University of Glasgow
School of Education

SES Society for Educational Studies

Reliability and Consistency in Judging New Teacher Practices – Why Does It Matter?

A Research Project Funded by the Society for Educational Studies National Award 2022

Authors: Professor Sarah K. Anderson (Principal Investigator), Professor James Conroy, Dr Sevda Ozsezer-Kurnuc, Dr Pinky Jain, and Professor Andrew Davies

With: Professor Rachel Lofthouse, Mr Daryl Phillips, and Ms Mary Lappin

Final Report to Funders
September 2024

Project Team

Name: Sarah K. Anderson, Principal Investigator
Title: Professor of Teacher Education
School: School of Education
Institution: University of Glasgow
Address: R664, Saint Andrew's Building, 11 Eldon St. Glasgow, G3 6NH
Direct Tel No: 07423231938
Email: sarah.anderson.3@glasgow.ac.uk

Name: James Conroy, Co-Investigator
Title: Professor of Philosophical and Religious Education, Vice Principal Emeritus
School: School of Education
Institution: University of Glasgow
Address: R375d, St. Andrews Building, 11 Eldon Street, Glasgow, G3 6NH
Direct Tel No: 01413307375
Email: James.Conroy@glasgow.ac.uk

Name: Sevda Ozsezer-Kurnuc
Title: Research Associate
School: School of Education
Institution: University of Glasgow and Turkish Ministry of National Education
Address: St Andrew's Building, Glasgow, G3 6NH
Direct Tel No: 07727070734
Email: sevda.ozsezerkurnuc@glasgow.ac.uk

Name: Mary Lappin
Title: Deputy Head of School & Senior Lecturer
School: School of Education
Institution: University of Glasgow
Address: R579 Level 5, St Andrew's Building, Glasgow, G3 6NH
Direct Tel No: 01413301886
Email: Mary.Lappin@glasgow.ac.uk

Contact Information for Partner Institutions

Name: Dr Pinky Jain
Title: Head of Teacher Education
School: Carnegie School of Education
Institution: Leeds Beckett University
Address: Carnegie Hall, 101, Headingley Campus
Direct Tel No: 01138120000
Email: Pinky.Jain@leedsbeckett.ac.uk

Name: Andrew James Davies
Title: Professor of Education/Director of the Centre for Research into Practice
School: Department of Education and Childhood Studies, School of Social Sciences, Swansea University
Address: Keir Hardie Building, Swansea University, Singleton Campus, Sketty, Swansea, Wales, SA2 8PP
Direct Tel No: N/A
Email: andrew.j.davies@swansea.ac.uk

Name: Rachel Lofthouse
Title: Professor of Teacher Education, Founder – CollectivED: The Centre for Mentoring, Coaching & Professional Learning
School: Carnegie School of Education
Institution: Leeds Beckett University
Address: Carnegie Hall, 121, Headingley Campus
Direct Tel No: 01138122326
Email: R.M.Lofthouse@leedsbeckett.ac.uk

Name: Daryl Phillips
Institution: Aberystwyth University
Direct Tel No: N/A
Email: dap105@aber.ac.uk

Contents

Cite this report as:

Anderson, S. K., Conroy, J., Ozsezer-Kurnuc, S., Jain, P., & Davies, A. (2024). *Reliability and consistency in judging new teacher practices – why does it matter? Project report to the Society for Educational Studies*.

List of Tables

List of Figures

Terminology

This section defines the notions used in this study, benefiting from the Glossary of the Council for the Accreditation of Educator Preparation (CAEP, n.d.) and the research methodology book of Cohen et al., (2018). Definitions of teacher and teaching effectiveness were followed from Darling-Hammond (2013, p. 12).

Candidate: any individual who engages in a teacher education provider's preparation process to complete the teacher education programme or to receive teaching licensure/certification. Candidates may also be known as pre-service teacher candidates, candidate teachers, student teachers, university students, or intern teachers.

Course: a specific set of lessons taken as a part of programme, consisting of coursework such as assignments and pen-and-paper exams to pass the course.

Dispositions: habits of professional action and moral commitments that underlie a (candidate) teacher's effectiveness.

Evaluatee: an individual being evaluated, assessed, rated, or judged. They are also known as ratees.

Evaluation**:** a process of assessment and making a judgement about merit of a programme, process, or individual (e.g., candidates, clinical faculty) based on available information.

Evaluator: any individual responsible for evaluating, assessing, rating, or judging the work of an evaluatee. They are also known as a rater and a scorer.

Field experience*:* a variety of candidate experiences in school settings, aimed at preparing them to teach students. A field experience may involve observation opportunities for candidates and/or practice opportunities to apply content and pedagogical knowledge in school settings. Hands-on practical experiences in general includes supervision from university-based teacher educators and/or mentoring from school-based teacher educators. *Field experience* may also be known as school-based experiences, field learning, onsite experience, sequence, placement, clinical experience, clinical practice, student teaching, internship, field work, and clinical internship.

In-service teacher: any schoolteacher employed to work in a school.

Inter-rater reliability: level of agreement or similarity among different observers who are evaluating the same thing (i.e., candidate, construct).

Judgement: *a* process of forming an opinion, making a decision, or placing value to a programme, process, or individual (e.g., candidates, clinical faculty).

New teacher**:** an in-service teacher who is recently employed, having a few years of experience in teaching.

Reliability: an umbrella term for dependability, consistency, and replicability over time, over instruments and over groups of respondents. Reliability is a precondition for validity.

Rubric: a type of scoring tool, communicates expected performance, organised in a table or matrix format, with criteria listed on the vertical axis and levels of performance on the horizontal axis.

School-based teacher educators: an individual involved in teacher preparation whose primary institutional home is a school and who takes on mentoring and partnership responsibilities in addition to their own school responsibilities. They may also be known as university liaisons, site facilitators, cooperating teachers, mentor teachers, classroom teachers, classroom mentor teachers, collaborating teachers, or school liaisons.

Standards: normative statements about teacher education providers and teacher candidate practices, performances, and outcomes that are the basis for an assessment. Standards can be created by various entities, such as TEPs ('institutional standards'), professional organisations ('professional standards'), and government bodies (national/state standards).

Student: a learner in a school setting or other structured learning environment but not a learner in a teacher education programme. Students can also be known as pupils.

Teacher educator: anyone who directly provides instruction or support services to the candidate in any type of teacher education provider settings. This includes university-based and school-based teacher educators.

Teacher education programme: a programme candidates enrol for their teacher preparedness in a specific field or level of school.

Teacher education provider*:* an entity responsible for the preparation of teachers at initial and advanced levels. It may be a public or private university, or an alternate body (i.e., Teach First). A teacher education provider could include more than one teacher education programme.

Teacher effectiveness: the personality traits, skills, and understandings an individual brings to teaching, including dispositions to act in certain ways, including strong content knowledge, knowledge of how to teach others in that area and skill in implementing productive teaching practices, understanding of learners and their development, abilities to organize and explain ideas as well as observe and think diagnostically, and adaptive expertise to make judgements about what is likely to work in a given context in response to pupils' needs (Darling-Hammond, 2013, p. 11)

Teaching efficiency: connotes the impeded route to achieve goals where effectiveness is the optimal to achieve such goals.

Teaching effectiveness: distinct from *teacher* effectiveness, teaching effectiveness refers to strong instruction that enables a wide range of pupils to learn. It is in part a function of teacher effectiveness (knowledge, skills, and dispositions) and is influenced by context of instruction (Darling-Hammond, 2013, p. 12).

Triangulation: utilisation of multiple techniques, perspectives and/or methods, to map out, or explain more fully, the richness and complexity of research findings. Triangulation is a way of demonstrating concurrent validity. Triangulation involves types such as time triangulation, space triangulation, theoretical triangulation, investigator triangulation, methodological triangulation.

University-based teacher educator: an individual involved in educator preparation whose primary institutional home is a college or university. University-based teacher educators are a specific type of boundary-spanning teacher educator who engage in evaluation, coaching, instruction, and partnership and assume expanded and multiple responsibilities within, and often across, each of these four domains. A university-based teacher educator may be otherwise known as a university supervisor, university liaison, clinical supervisor, or clinical faculty.

Validity: a term to describe the level of accuracy in measuring the theoretical constructs (i.e., proxies) of evaluation tool under investigation.

List of Abbreviations

AI	artificial intelligence
CLASS	Classroom Assessment Scoring System
Co-I	Co-Investigator
CPD	continual professional development
ECT	early career teacher
edTPA	Educative Teacher Performance Assessment
EPS	Expected Progress Statement
FAV	formative assessed visit
GTCE	General Teaching Council for England
GTCS	General Teaching Council for Scotland
HEI	higher education institution
I-LAST	Item-Level Assessment of Teaching Practice
InTASC	Interstate Teacher Assessment and Support Consortium
IRR	inter-rater reliability
ITE	initial teacher education
ITT	initial teacher training
LBU	Leeds Beckett University
NQT	newly qualified teacher
Ofsted	Office for Standards in Education
PACT	Performance Assessment for California Teachers
PDQ	Professional Development Qualities
PEI	Profile for Evaluation of Intern
PGCE	Postgraduate Certificate of Education
PI	Principal Investigator
PISA	Programme for International Student Assessment
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PSTL	Professional Standards for Teaching and Leadership
QAA	Quality Assurance Agency for Higher Education
QTS	qualified teacher status
RA	Research Associate
SCDE	Scottish Council of Deans of Education
SDG	Sustainable Development Goal

SJT	social judgement theory
SoE	School of Education
SPR	Standard for Provisional Registration
STP	Samples of Teaching Performance
TDA	Training and Development Agency for Schools
TEP	teacher education programme
TPP	teacher preparation programme
T-TESS	Texas Teacher Evaluation and Support System
TWS	Teacher Work Sample
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization

Executive Summary

In 2022 the Society for Educational Studies granted the National Award to Sarah Anderson (Principal Investigator), James Conroy (Co-Investigator), and Mary Lappin of the University of Glasgow School of Education. The generous award supported the 2-year project titled 'Reliability and consistency in judging new teacher practices – why does it matter?' The multi-phase project involved exploration of the nature of judgement-making processes regarding initial teacher education (ITE) students' practices in contexts of normed teaching standards. In addition to the awardees at the University of Glasgow, the project involved partnership with colleagues at Leeds Beckett University in England and Aberystwyth University in Wales. We considered this cross-national collaboration as crucial in light of the professional disassociation experienced by educators and researchers alike during and after the pandemic years; hence this project pursued an opportunity to come together and strengthen common understanding. The project sought to explore the possibilities for more accurate judgements to positively enhance teacher capacities, to reimagine the value and professional career trajectory of the 'teacher educator' as a reorientated role, and to investigate potential power dynamics among stakeholders that impact our collective understanding of professional competence.

The project is a multi-case analysis exploring the nature of judgements regarding ITE students' performance per normed teaching standards. It involves partnership with teachers, researchers, and university staff of three programmes in Scotland, Wales, and England. The project aims are:

- to better understand judgement processes in order to improve judgement-making on teaching effectiveness;
- to directly influence the practices of assessing and enhancing novice teachers' skills in clinical school placements, with the ultimate goal of enhancing pupil outcomes;
- to expand opportunities for dialogue across systems through a renewed sharing of practices, policies, and professional standards; and
- to meet the shared responsibility of training high-quality future educators in a sustainable model, foster networked improvement communities, and inform perspectives beyond Great Britain.

The following research questions are founded on these aims:

RQ1 What is the nature of shared judgement, consensus, and dissensus on observed teaching effectiveness among university-based teacher educators and school experience tutors/associate tutors and school-based mentor teachers?

RQ2 How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

RQ3 How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

Chapter 1: Introduction

This chapter provides a comprehensive overview of the research project focused on understanding how judgements are made about the effectiveness of ITE students' practices within the context of normed teaching standards. The project is a collaborative effort and reflects a cross-national approach to investigating teacher education practices in the UK. The chapter highlights the global challenges in teacher education, such as the shortage of qualified teachers, the increasing complexity of teaching, and the evolving accountability measures. It underscores the role of high-quality teacher preparation in addressing these challenges and the need to ensure that teaching standards and judgements about new teachers' readiness align with educational priorities and future uncertainties. Teacher quality, particularly in the context of the United Nations Sustainable Development Goal 4 (SDG 4), is emphasized as critical to achieving inclusive, quality education worldwide.

The report delves into the complexities of evaluating new teachers, including the inconsistencies and variability in judgements by teacher educators, mentor teachers, and university staff. These judgements, often influenced by subjective factors and varying standards across the UK, have significant implications for teacher development and educational outcomes. The chapter also explores the power dynamics between universities and schools in the judgement process and the need for collaboration to enhance reliability in assessments.

Chapter 2: Methodology

This chapter details the methodology employed in the research project aimed at understanding the decision-making processes used by university-based and school-based teacher educators to judge teacher candidates' readiness to teach. The study used a concurrent and convergent mixed methods design, organized into five phases.

The first phase involved a systematic literature review to assess existing methodologies and tools for evaluating teaching effectiveness. The second phase was a professional teaching standards policy review, comparing five sets of standards, including UNESCO and UK-based standards, to align evaluation criteria across participating institutions in Scotland, England, and Wales.

The third phase was a multi-case study across teacher education programmes (TEPs) in Scotland, England, and Wales, where empirical data was collected through a video observation task, a questionnaire, and focus groups and interviews with teacher educators. This phase was key to investigating the judgements made about teaching effectiveness.

The fourth phase used the Delphi panel technique to gather insights from national and international experts and build consensus on key findings. The final phase, a convergent cross-phase and cross-case meta-analysis, synthesized results across all phases to answer the three research questions.

Social judgement theory underpinned the research design, guiding the process of identifying and analysing the cues used in judgement-making. The methodology emphasized flexibility,

robustness, and ethical rigour, ensuring a comprehensive investigation of how judgements are made about teacher candidates' effectiveness. The chapter concludes with an outline of the project's ethical considerations, triangulation of methods, and data collection procedures.

Chapter 3: Systematic Literature Review

This chapter presents the findings of a systematic literature review aiming to expand knowledge on how judgements about teaching effectiveness are made. The review explored methodologies and data collection tools used in assessing teaching effectiveness, with a focus on validity, reliability, and judgement-making processes. The review was conducted using the Preferred Reporting Items for Systematic reviews and Meta-Analyses review process. It aimed to better understand: (a) methodologies and data collection tools used when making judgements about student teaching effectiveness; (b) ways in which validity and reliability are considered and conceptualized in judgement-making about new teacher effectiveness; (c) processes involved in assessing new teacher effectiveness within TEPs; and (d) how evaluation and results are used to improve judgement-making on new teacher effectiveness.

Key findings include the identification of three major areas of focus in the literature: validity (primarily concerning construct and face validity); reliability (mainly inter-rater consistency); and judgement-making (focusing on instrument development and use). The review analysed 45 peer-reviewed studies, revealing a strong emphasis on validity, followed by reliability, while relatively fewer studies directly addressed judgement-making instrument development and use. The review also highlighted a gap in research on judgement-making within the UK and the need for further exploration of how different rater groups make decisions. Most of the studies originated from the US, with no research specifically from the four UK nations. The findings underscored the complexity of ensuring valid and reliable assessments in teacher education, with varying tools and standards used globally.

These findings contribute to the field by mapping the existing methodologies used in evaluating teaching effectiveness and identifying areas for improvement, particularly in the reliability and transparency of judgement processes. It also emphasizes the need for more UK-based studies to align teacher evaluation practices across devolved educational contexts. The literature review indicates that improving the reliability of professional judgement can lead to better alignment of evaluations and more effective collaboration between schools and universities, while addressing power imbalances between university-based and school-based educators can foster more equitable judgement processes.

Chapter 4: Comparative Policy Analysis

This chapter presents a comparative analysis of professional teaching standards across England, Scotland, and Wales. The review highlights the distinct paths taken by each nation in defining and evaluating teacher competencies since devolution. Using UNESCO's *Global Framework for Professional Teaching Standards* and the US-based Interstate Teacher Assessment and Support Consortium standards as benchmarks, the chapter evaluates the alignment and variation in how newly qualified teachers are assessed across the three UK jurisdictions.

Key findings include the divergence in educational priorities and ideologies between the nations. In England, standards emphasize technical competencies and prescribed practices, focusing on curriculum knowledge and behaviour management. However, England lacks professional standards related to research engagement and continuous professional development, which are emphasized in both Scottish and Welsh frameworks. Scotland's standards prioritize reflective practice, social justice, and inquiry-based approaches, including a strong emphasis on Gaelic language provision and outdoor learning. Welsh standards are the most comprehensive, balancing pedagogy, professional development, and leadership while highlighting research-informed practice and the promotion of Welsh culture and language. The analysis also reveals differences in the conceptualization of teacher–student relationships. While Scotland and Wales emphasize holistic, context-sensitive approaches to student development, England's standards tend to position students as passive recipients of instruction. This has implications for how teacher preparation programmes are structured in each country, with Scotland and Wales fostering greater professional autonomy and reflective practice and England taking a more rigid, outcome-focused approach.

This research contributes to the field by providing a detailed crosswalk comparison of professional standards in devolved UK contexts, offering insights into how differing national policies shape teacher education and professional development. It also informs international research by aligning national standards with globally recognized frameworks, providing a basis for future comparative studies. The findings underscore the importance of integrating research, reflective practice, and professional learning into teaching standards to foster a more adaptive and innovative educational environment.

Chapter 5: Case Study 1

This chapter provides a detailed case study of judgement-making in ITE at the University of Glasgow, focusing on the practices of teacher educators, school experience tutors, and mentor teachers. Using a combination of video observation task, questionnaire, and focus groups and interviews, the case study explores how these different evaluators assess the effectiveness of student teachers during school placements. The study reveals significant variation in how different groups judged the same teaching practices. While all groups rated the 'learning environment' as the most effective aspect of teaching, there were differences in how they assessed areas like 'instructional strategies' and 'content' knowledge.

Evaluators used four main strategies to make their judgements:

- classroom cue utilization (focusing on observable actions of teachers and students);
- suggestions for lesson improvement (offering feedback on what could be improved);
- internal expectation criteria (relying on personal standards); and
- no identified strategy (instances where evaluators were unsure of their rationale).

Participants found it easiest to rate the 'learning environment' but struggled more with assessing 'instructional strategies' and 'assessment'. This reflects the difficulty in making subjective judgements about teaching effectiveness without direct access to lesson plans or context. There was broad consensus that consistent and reliable judgements are crucial for

fairness, maintaining standards in education, and ensuring that new teachers meet the expected level of competence. However, the findings suggest that evaluators approach judgements differently based on their roles and experiences, which could affect consistency. The study identified that personal biases, evaluator experience, and contextual factors like classroom complexity can influence judgements. Teacher educators tended to focus more on professional standards and the use of clear criteria, while mentor teachers emphasized the practical realities of classroom teaching.

This case study contributes to the understanding of how judgements about new teachers' effectiveness are made in a real-world context. It highlights the complexity of the evaluation process and the factors that influence judgements, including evaluator roles, biases, and the challenges of ensuring consistent standards. The research underscores the importance of developing more standardized approaches to judging teaching effectiveness while acknowledging the need for professional discretion. This study adds to the broader conversation on improving reliability and fairness in teacher education assessments across different educational contexts.

Chapter 6: Case Study 2

This chapter presents a case study on judgement-making practices within the ITE programme at Leeds Beckett University, focusing on teacher educators, tutors, and school-based mentor teachers. The case explores how these different groups assess teaching effectiveness, using a video observation task, questionnaire, and focus groups and interviews to evaluate teaching practices. Teacher educators, tutors, and mentor teachers exhibited different approaches when evaluating the same teaching practice. While all groups rated the 'learning environment' highly, 'instructional strategies' and 'assessment' were consistently rated lower by tutors and mentor teachers compared to university-based educators. In making their judgements, the evaluators employed the four key strategies listed above for Chapter 5.

All groups found it easier to judge the 'learning environment' but identified 'assessment' and 'research' as the most difficult dimensions to rate. Teacher educators highlighted challenges in assessing research-informed practices, while tutors and mentors noted the difficulties of observing effective assessment strategies in short teaching clips. The evaluators' judgements were influenced by their role (university-based versus school-based) and their prior experience. School-based mentors, for instance, placed greater emphasis on practical classroom management, while university-based educators focused more on adherence to teaching standards and research-based approaches. There was broad agreement across groups that ensuring accuracy, consistency, and fairness in judgements is crucial. The study highlighted the importance of multiple evaluators and using evidence-based judgements to minimize bias.

This case study contributes to understanding the complexities and inconsistencies in evaluating teacher candidates within ITE programmes, particularly the differences in how various stakeholders (university educators, tutors, and school mentors) assess teaching effectiveness. It underscores the challenges in achieving reliable and consistent judgements and suggests the need for more standardized evaluation processes that incorporate multiple

perspectives. This research also emphasizes the importance of clear judgement criteria and the integration of both practical and theoretical teaching dimensions in assessments. The case study reveals a complex interplay between consensus and dissensus in judgement-making, the potential for enhanced reliability to drive collaboration, and the significant influence of power dynamics on the roles of university-based and school-based teacher educators in ITE.

Chapter 7: Case Study 3

This chapter presents a case study of the ITE programme at Aberystwyth University in Wales, focusing on how judgements of teaching effectiveness are made. The case is one of three in this research project and highlights the unique context of ITE in Wales, particularly the processes for student teacher assessment and the challenges faced by the Aberystwyth ITE Partnership. The ITE programme at Aberystwyth University offers an innovative structure where student teachers experience both primary and secondary settings, regardless of their specialization. This approach is designed to enhance their understanding of teaching progression and increase their employability, especially within the growing 'all-through' school model in Wales. Student teachers are evaluated continuously against the Professional Standards for Teaching and Leadership set by the Welsh Government. These standards focus on five domains: pedagogy; collaboration; professional learning; innovation; and leadership. Assessment is carried out through a combination of mentor observations, regular reviews, and final holistic evaluations.

During the research period, the Aberystwyth ITE Partnership faced significant challenges. An Estyn inspection report in 2023 highlighted issues such as inconsistent mentoring quality, poor communication across the partnership, and a lack of coherence between university and school-based components of the programme. As a result, the partnership did not secure re-accreditation and will cease offering its Postgraduate Certificate of Education programme in 2024. Due to the context of the programme's re-accreditation challenges, the research team decided not to include the data collected. This decision was made to ensure integrity given the unique circumstances of the programme during the data collection period.

This case study provides valuable insights into the complexities of teacher education in Wales, particularly regarding the challenges of maintaining high standards and coherence in partnerships between universities and schools. The study contributes to the understanding of how power dynamics and collaborative processes between educational institutions can impact the effectiveness of teacher training programmes. Furthermore, it underscores the importance of consistent and high-quality mentorship in developing effective teachers. Despite the decision to omit data, this research highlights the need for a continued focus on the voices of mentor teachers and the evolving structures of teacher education in Wales.

Chapter 8: Delphi Panel

This chapter outlines the Delphi panel's role in consolidating the findings from earlier phases of the research project. The Delphi panel, which involved iterative discussions among nine international education experts, was used to build consensus on issues related to judgements

about teacher effectiveness. The process highlighted several key themes and produced recommendations for improving teacher education.

The panel reached consensus that current competency frameworks focus too much on observable behaviours, neglecting more complex and essential aspects of teaching, such as professional dispositions and the teacher's impact on student engagement and learning. Participants suggested shifting from assessing micro-competencies to evaluating a teacher's broad 'repertoire' of abilities over time. The panel emphasized that teaching effectiveness should not be evaluated solely based on rigid criteria, but through a more holistic, developmental lens. They advocated for ongoing collaboration between student teachers, mentor teachers, and teacher educators, stressing that judgement should be seen as part of a continuous learning process.

There was a strong call for better integration between university and school-based mentors, who play a crucial role in assessing new teachers. The experts highlighted the need for enhanced mentor development and consistent collaboration between universities and schools to create a shared language and framework for making professional judgements. While consistency in judgement-making is important, the panel recognized the risk of reducing teaching assessment to a 'laundry list' of competencies that oversimplifies the complexity of teaching. They advocated for a balance between standardization and allowing flexibility to account for individual teaching styles and contexts. The panel expressed scepticism about the use of artificial intelligence in assessing teacher effectiveness, emphasizing that judgement is inherently a human activity. They were concerned that over-reliance on data-driven approaches could overlook the nuances of teaching, such as relational and dispositional factors that are difficult to quantify. The experts discussed the intangible qualities that make a teacher effective – referred to as the 'it' factor. These include trustworthiness, relational skills, and the ability to foster student flourishing. Assessing these qualities requires a more relational and reflective approach, rather than one based solely on technical competencies.

This chapter contributes significantly to the understanding of how judgements about teacher effectiveness should be made. It critiques current competency frameworks for their focus on observable behaviours and proposes a shift towards a more comprehensive, collaborative, and reflective approach to teacher evaluation. By emphasizing the importance of relationality, mentorship, and context, the findings challenge the oversimplification of teacher assessments and advocate for more nuanced, human-centred evaluation methods. The Delphi panel's insights offer a framework for improving consistency and reliability in judgements while acknowledging the complexity of the teaching profession.

Chapter 9: Cross-Case and Cross-Phase Analysis

This chapter presents a convergent cross-case and cross-phase analysis to answer the three research questions of the study. It compares findings from case studies in Scotland and England, reviews relevant literature, analyses professional teaching standards, and incorporates insights from a Delphi panel of experts.

Shared Judgement in Teacher Education (RQ1)

- **Consensus:** Across the phases of the study, there was agreement on the importance of observable teaching competencies, such as classroom management and student engagement. Most participants valued professional judgement and emphasized growth and development over time rather than checklist-based evaluations. The 'learning environment' was generally considered the easiest aspect to judge.
- **Dissensus:** Differences emerged around instructional strategies and assessment, with university-based educators focusing more on reflective practices, while school-based mentors emphasized practical, immediate classroom performance. There was also debate on how much variability in judgement is acceptable, with some favouring flexibility while others feared that inconsistencies could undermine fairness.

Fostering Collaboration Between Schools and Universities (RQ2)

- Enhanced reliability of professional judgement, through standardized evaluation tools and shared assessment criteria was identified as key to fostering collaboration between schools and universities. Sustained residency models, co-designed frameworks, and continuous feedback loops were suggested as strategies to create more consistent evaluations and strengthen partnerships. Both case studies highlighted the importance of dialogue and shared decision-making to ensure that judgements reflect both theoretical and practical perspectives.

Power Dynamics in Judging Teacher Effectiveness (RQ3)

- Power dynamics shape the roles of university-based and school-based teacher educators. University educators typically hold more authority in summative evaluations, while school-based mentors provide practical insights. This creates a power imbalance that could be mitigated by giving mentors a more formal role in the assessment process and promoting co-construction of the teacher education experience. The study emphasized the need for mentors to be seen as equal partners in evaluation, fostering a more balanced and collaborative environment.

This research highlights the complexities involved in judging teaching effectiveness and offers several contributions:

- **Emphasizing a holistic approach:** The study advocates for a shift from checklist-based evaluations to a more holistic approach that values both observable competencies and less tangible dispositional traits.
- **Strengthening collaboration:** It provides practical recommendations for fostering stronger partnerships between schools and universities, including the use of co-designed assessment frameworks and shared decision-making processes.
- **Addressing power dynamics:** The research contributes to the understanding of how power dynamics influence teacher evaluations and offers strategies for creating more equitable partnerships in teacher education.

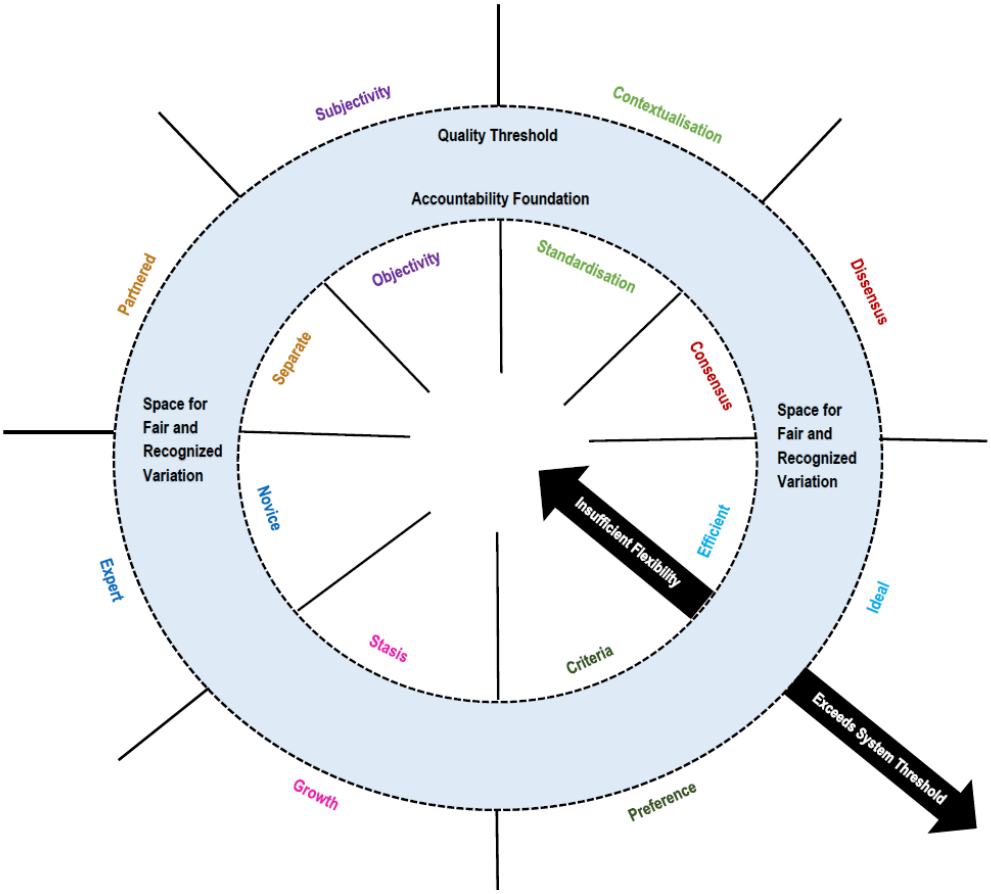
Overall, this chapter underscores the need for a more collaborative, reliable, and fair system for judging teacher effectiveness, balancing both theoretical knowledge and practical classroom experience.

Chapter 10: A Model of Dynamic, Adaptive Systems Thinking in Teacher Education

This chapter introduces the Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education (illustrated below), a conceptual framework developed through the research project to address the complexities of judging teaching effectiveness. The model is designed to help TEPs navigate the multifaceted challenges involved in evaluating new teachers' practices within an ever-evolving educational system.

The research reveals that linear approaches to judging teaching effectiveness overlook the inherent complexity of education systems. Judging teacher effectiveness involves balancing multiple, often contradictory, factors, such as standardization versus contextualization, efficiency versus ideality, and subjectivity versus objectivity. The Duplexity Model aims to reflect this complexity by promoting flexibility and fairness in judgement-making processes. The model is grounded in the notion of duplexity, which refers to the coexistence of two interrelated forces that are not oppositional but complementary. These include factors such as balancing fairness with complexity and ensuring that judgements are adaptable to the nuances of different teaching contexts.

Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education



The model applies complexity theory to teacher education, emphasizing the need for adaptive decision-making that can respond to the unpredictable and interconnected nature of teaching environments. The research highlights that teacher education is a dynamic system influenced by multiple variables, such as political, social, and contextual factors, requiring constant adjustment and resilience. The concept of 'teacher tensegrity' is introduced to describe the balance between tension and flexibility in teacher education. This principle, drawn from architecture and nature, suggests that teacher education systems must maintain structural integrity while being adaptable to various pressures and challenges, much like the resilience seen in natural systems. The model is applied in a hypothetical scenario to evaluate its practicality in an ITE programmes. The analysis shows how balancing factors such as standardization, fairness, and collaboration can provide more reliable and consistent judgements of teaching effectiveness. It highlights the need to adjust processes when overshooting or undershooting acceptable thresholds of quality.

This research contributes to the field of teacher education by offering a new way of thinking about judging teaching effectiveness, grounded in dynamic systems thinking and complexity theory. The Duplexity Model provides a flexible framework that acknowledges the inherent tensions in teacher evaluation processes, offering a balanced approach to decision-making. It encourages teacher educators and policymakers to consider both the immediate and long-term impacts of their judgements and to focus on sustainability in teacher preparation. By emphasizing adaptability, collaboration, and the integration of multiple perspectives, the model pushes forward the idea that fairness and reliability in judging new teachers require ongoing flexibility and reflection within complex educational systems. This framework represents a shift away from rigid, linear approaches, moving towards a more nuanced and adaptive understanding of teacher education.

Chapter 11: Conclusions and Recommendations

This chapter presents the conclusions and recommendations of the research project, which investigated the reliability and consistency in judging new teacher practices within ITE programmes. The chapter also acknowledges limitations. The study utilized social judgement theory to explore judgement-making processes and examined how shared judgements, collaboration, and power dynamics influence teacher evaluations.

Key Findings

- **Importance of reliable and consistent judgements:** The study highlighted that consistent and reliable judgements of teaching effectiveness are essential for fairness, ensuring that all teacher candidates are evaluated equitably. Inconsistent judgements can lead to disadvantages for some candidates, affecting career progression, support, and development opportunities. Reliable judgements also help in identifying future teachers who can positively impact pupil learning.
- **Impact on teacher development:** The experiences and feedback that student teachers receive during their mentorship period significantly influence their professional development and commitment to the teaching profession. Ensuring quality mentorship

and fair assessments during this formative period is crucial for the long-term success of teacher candidates and the broader teaching profession.

- **Collaboration and power dynamics:** The research identified the need for greater collaboration between schools and universities, as well as the importance of addressing power imbalances between university-based teacher educators and school-based mentors. Enhancing collaboration through joint decision-making, feedback loops, and shared responsibilities can lead to more equitable and effective judgement processes.
- **Complexity in judging teaching effectiveness:** The study underscored the complexity involved in evaluating new teachers, noting that current systems often oversimplify the process. A more adaptive and flexible approach is necessary to account for the various contextual factors that influence teaching performance.

The research contributes to a deeper understanding of the challenges and complexities involved in judging teaching effectiveness. It calls for more adaptable, collaborative, and fair evaluation systems in teacher education. Future research should focus on exploring non-university-based teacher preparation programmes, expanding comparative analyses of professional teaching standards, and refining the conceptual Duplexity Model introduced in the study. In conclusion, the project emphasizes the need for systemic changes in how teacher effectiveness is judged, advocating for a more flexible, collaborative, and fair approach that aligns with the complex realities of teaching.

Recommendations

Taken together, findings from this project support several recommendations for improving the judgement-making process of teaching effectiveness for TEPs, school partners, and policymakers. These recommendations are predicated on the principles of SDG 4 that high-quality teaching, for all students, in all circumstances is a right.

Teacher Education Programmes

1. Examine entrance requirements and evaluation processes to ensure they do not narrow the talent pool.
2. Eliminate ineffective programmatic requirements in ITE that do not demonstrate predictive validity of a positive impact on pupil learning and development.
3. Revise ITE structure and curriculum with a focus on creating opportunities and learning experiences in which future teachers develop skills needed to deal with complexity and uncertainty and to translate theory into their practice.
4. Prepare student teachers for systems thinking through using systems thinking.
5. Explore opportunities to expand the amount of time prospective educators spend in clinical experiences in which future teachers can sustain relationships needed to develop the sophisticated skill set required for effective teaching in increasingly complex classrooms (e.g., multi-year residencies, 1-year mentored residencies).

6. Conduct a collaborative research study to examine the effectiveness of the 1-year Postgraduate Diploma in Education/Postgraduate Certificate in Education model of teacher preparation.
7. Standards should be calibrated to better reflect different levels of experience rather than a one-size-fits-all approach across the continuum of the professional teaching career.
8. Develop judgement processes with explicit performance expectations (e.g., look fors).
9. Create opportunities for developing clinical judgement skills.
10. Adopt, revise, or create evaluation measures of teaching effectiveness that better address the complexities of teaching and allow for a fair degree of dissensus.
11. Expand professional development opportunities for teacher educators and mentor teachers.
12. Create a diploma, certificate, or endorsement for teacher educators and for mentor teachers.
13. Emphasize the role of mentor teachers in their subject area expertise, as historians, artists, mathematicians, etc.
14. Create a specialized TEP advisory board focused on clinical partnerships and practice.
15. Form Research Practice Partnerships where researchers and practitioners work together to address educational challenges and improve student outcomes.
16. Jointly make placement decisions.
17. Only place teacher candidates with mentor teachers who demonstrate high-quality instruction.
18. Partner with schools to identify preparation gaps and opportunities.
19. Develop a comprehensive probation process for new teachers that involves teacher educator programmes.
20. Reduce bureaucratic workload for all involved in partnership to prepare teachers.
21. Explore team teaching/co-teaching models involving school-based mentor teachers and university-based teacher educators.
22. Gather actionable feedback from ITE graduates and mentor teachers to inform programme improvements.
23. Use systems thinking to incorporate feedback loops into teacher education.
24. Collaborate with other TEPs nationally and internationally to inform continuous improvement efforts.

School Partners

1. Place future teachers with school-based mentor teachers who have demonstrated exceptional teaching practices and are committed to working with teacher candidates.

2. Ensure schools where teacher candidates are placed provide high-quality, research-based instruction, effective social emotional learning, and evidence-based interventions to address the needs of all pupils, including those at risk.
3. Employ a university-based teacher in residence to facilitate collaborative approaches and discussions, potentially considering residency models.
4. Pair newer school-based mentor teachers with more experienced colleagues and offer differentiated programming and support for both groups.
5. Ensure the school's overall strategic plan includes a personnel strategy and specific goals for talent management and the development of teacher candidates, aligned with the school's mission, vision, and strategy.
6. Explore flexible or non-traditional work arrangements for school-based and university-based teacher educators to enhance collaboration and efficiency.
7. Identify and celebrate highly effective student teachers and teachers, and provide them with opportunities to serve as educational ambassadors for the profession.

Policymakers

1. Ensure professional standards for educators, including those related to teaching and headship, are clear, accessible, and applicable to diverse teaching contexts. Include specific responsibilities for mentoring and educating future teachers.
2. Adjust funding models to provide fair compensation to TEPs for the actual costs associated with strong clinical experiences.
3. Promote innovative teacher preparation programmes that address barriers and improve educator outcomes through increased funding, research, and recognition of promising initiatives.
4. Invest in research on assessment practices, the validity and reliability of accountability measures, and the data points used to determine the quality of teacher preparation.
5. Provide fair compensation to teachers who serve as mentor teachers during ITE preparation experiences.
6. Offer comprehensive technical assistance to TEPs and schools to support the development, implementation, improvement, and expansion of teacher educator preparation programmes nationwide.
7. Establish a system-wide academy that provides professional development, networking, and mentorship opportunities for new mentor teachers to strengthen their skills.
8. Include the voices of student teachers on government committees involved in teacher education and professional development.
9. Examine and revise policies that may hinder or discourage experienced teachers with extensive tacit knowledge from entering teacher education roles.

10. Ensure that compensation for university-based teacher educators is competitive with the broader education system and consider relevant work experience when determining starting salaries.
11. Convene groups of teacher educators and connect them with their Members of Parliament to advocate for policies that support teacher education and professional development.

1 Introduction

In 2022 the Society for Educational Studies granted the National Award to Sarah Anderson (Principal Investigator), James Conroy (Co-Investigator), and Mary Lappin of the University of Glasgow School of Education. The generous award supported the 2-year project titled 'Reliability and consistency in judging new teacher practice – why does it matter?' The multi-phase project explored the nature of judgement-making processes regarding initial teacher education (ITE) students' practices in contexts of normed teaching standards. In addition to the awardees at the University of Glasgow in Scotland, project partners were colleagues at Leeds Beckett University in England and Aberystwyth University in Wales. This cross-national collaboration was considered crucial in light of the professional disassociation experienced by educators and researchers alike during and after the Covid-19 pandemic years, and the project provided an opportunity to come together and strengthen common understanding.

The project sought to explore the possibilities for more accurate judgement-making to positively enhance teacher capacities, to reimagine the value and professional career trajectory of the 'teacher educator' as a reorientated role, and to investigate potential power dynamics among stakeholders that impact our collective understanding of professional competence. This research acknowledges the petition for stakeholders to participate fully in the 'development, implementation, monitoring, and evaluation of education policy' (Education International & UNESCO [United Nations Educational, Scientific and Cultural Organization], 2019, p. 4), a core responsibility of those who prepare future teachers.

In this chapter, we outline the inception and purpose of the research and provide a contextual backdrop for the project, including teacher education as situated in the global sphere. We also position the transformative work of teaching with a focus on the preparation of future teachers taken up collaboratively by teacher education programmes (TEPs) and schools, and we provide a theoretical framework for the project. We bring these forward acknowledging the importance of practising teachers to the imperative of providing high-quality, school-based experiences during teacher preparation and the teachers' role in shared decision-making and assessing candidates' readiness for their classroom responsibilities. This is offered in recognition of the call in the UNESCO report *Reimagining Our Futures Together* (2021) for teacher education 'to be rethought to align with educational priorities and orient better towards future challenges and prospects' (p. 85). The Introduction concludes with an overview of the chapters in this project report.

1.1 A High-Quality Teaching Profession

As teacher educators, we are motivated to contribute to more clearly defining teaching as a profession and making collective work and high-quality research our norm. We are focused on learning from the practices of our peers and anchoring our work in our common goal, the United Nations (UN) Sustainable Development Goal 4 (SDG 4), to deliver an 'inclusive and quality education for all' (UN, 2022). Our aim is to prepare high-quality teachers who can deliver this goal. Globally, we are not on a trajectory to meet the targets of SDG 4 by 2030. A

recent UNESCO report on the profession put forward the urgent need for 44 million teachers worldwide to attain the goal; this includes replacing over half of the existing teachers who are leaving the profession (UNESCO, 2024). This crisis could impact dramatically on the learning and development of young people, and on top of that SDG 4 is a key enabler of most of the other 16 SDGs. Alongside the alarming statistics on the global shortage of teachers (UNESCO, 2024), many other challenges threaten educational access and quality. These include large class sizes, Covid-19-related learning loss, overburdened educators, increased classroom complexity, educational disparities, and financial strain on educational systems (UN, n.d.). In this context, teacher preparation is key, and it has been noted as playing an important role in the systematic reform of schools (Bransford et al., 2005). Teacher quality remains one of the most influential in-school resources for improving pupil learning (National Academy of Education, 2024).

Within the UK, teacher education has changed considerably since devolution of education to the home nations, with each having separate systems under separate governments. The ongoing and serious shortfall of teacher supply, semi-permanent in England (MacLean et al., 2024) and sporadic elsewhere in the UK and beyond, has resulted in the global rise of 'quick fixes' in the form of fast-track, for-profit, and non-university-based teacher preparation programmes. Some education systems, such as that in Scotland, have resisted the superficial allure of alternative programmes (Anderson & McMahon, 2024) and maintain a commitment to clinical experience. University-based teacher education continues to emphasize the role of education as a public good and recognizes the complexity of teaching which goes beyond technical dimensions (Darling-Hammond & Lieberman, 2012). Intellectual robustness (Donaldson, 2011) and sophisticated practical capacities remain at the heart of this approach to teacher preparation (Dickson, 2011), along with development of an enquiring stance (Cochran-Smith, 2009).

While we recognize that the quality of teacher education is not the only, or indeed possibly not the most important, contributor to student outcomes (Tan et al., 2023), it is nonetheless the case that teachers' beliefs and their dependant practices contribute substantially to such outcomes (Silverman et al., 2023). Thus, the importance of preparation programmes for the practices of teaching should not be underestimated. The different professional standards across England, Scotland, and Wales can have notable implications for teacher development and educational outcomes, especially considering the scope of impact of teacher preparation, which is much broader than simply providing a source of new teachers (Ell et al., 2019). Among complicating factors shared across the UK and beyond are the increasing reliance on adjunct and school-based supervisors and a perceived and perseverating disconnect between theory and practice. Moreover, Ziechner and Bier (2015) recognized the often 'marginalized' status of those who supervise clinical experiences and the under-resourcing of this work, which can undermine student support. Zeichner and Bier further noted the lack of value of faculty involvement in strong, school-based clinical experiences in the university promotion and reward system (p. 25). Interestingly, Clapham et al. (2023) identified that the Research Excellence Framework, the UK's system to assess the quality of research in higher education

institutions, has also reflected marginalization of research on ITE, a signal for how higher education values teacher education.

1.2 Background and Concerns

There is a sustained and deepening interest locally, nationally, and globally in the purposes of teacher qualifications and the related professional teaching standards and policies. Evaluation of student teachers' readiness to teach is a central component of high-quality teacher preparation, and this takes place most often during practice placements in schools (National Academy of Education, 2024). Higher amounts of practice teaching during preparation has indeed been found to be a benefit to new teachers' success when they enter the classroom (Boyd et al., 2008). The Clinical Practice Commission of the American Association of Colleges for Teacher Education (AACTE) established clinical practice as the basis for high-quality educator preparation, a statement strongly supported by the research base with evidence of improved preparation for future teachers and improved learning for pupils (AACTE, 2018). AACTE noted: 'Because the actual process of learning to teach requires sustained and ongoing opportunities to engage in authentic performance in diverse learning environments, clinical practice is a valuable, necessary, and fundamentally non-negotiable component of high-quality teacher preparation' (p. 14).

An inspection approach continues to take shape and dominate discourse in teacher education, in what, as discussed earlier, has been referred to as a crisis in teacher education provision (Mutton & Burns, 2024). This has sparked discord among teachers and teacher educators alike (Cochran-Smith, 2021) concerning perceived disproportionate levels of accountability in the form of high-stakes observations and evaluations and performative measures adopted in the course of teacher preparation, including those used in deciding entry into the profession. Darling-Hammond (2017) argued that there is a strong relationship between high-performing school education systems, the quality of student teachers, and robust intellectual and professional barriers to admission into the profession. Such 'appropriate' barriers include effective assessment of pre-service and early career teachers. Hattie (2023) observed that while TEPs claim to determine competence based on a comprehensive set of core attributes, this core remains different across providers and systems in a manner that Levine (2006) labelled 'unruly' and 'disordered' (p. 109).

Interestingly, a review by Klassen and Kim (2019) of 32 studies revealed only small correlations between academic and non-academic criteria during preparation and later teacher effectiveness. Research by Sandholtz and Shea (2011) contested the accuracy of supervisors' judgements of student teacher performance, questioning the reliability of determinations of readiness to teach. Donaldson (2010) reflected that the efficacy of judgements as to excellence in early career teachers according to professional standards is little better than random. Research by Raths and Lyman (2003) suggested that because of failures of professional agreement as to what constitutes a judgement of competence, many students in teacher education manage to pass into the profession despite significant incompetence. And Haigh and Ell (2014) found that mentor teachers take an 'idiosyncratic approach' to reaching decisions about teaching and that even where judges have a shared vision of quality teaching,

significantly different findings often emerge (p. 19). In addition, student teachers themselves have indicated that inconsistent assessment of their practice is a stressor and even reduces their classroom effectiveness (Murray-Harvey et al., 2000). Such failures of agreement are not uncommon and can be attributed to a range of factors (such as contextual differences, time constraints, and asymmetric attributions of importance). There are implications from this variability and dissensus in judgement-making to be explored.

The accurate and consistent judgement of teaching competence during clinical experiences continues to be an area of increasing interest and concern (Asher, 2018; Haigh et al., 2013; Schmoker, 2023; Seidenberg, 2017), particularly in an era of high accountability and increased scrutiny of teacher preparation. Efforts dedicated to defining what constitutes teachers' knowledge, skills, and competencies have been ongoing for decades, particularly since the mid-1980s (Tigelaar & van Tartwijk, 2010). These efforts have translated into standards and criteria in the pursuit of teacher effectiveness, serving as foundation and contributing guidelines for teacher education curriculums, assessment, and quality assurance (Yinger & Daniel, 2010). One of the earliest examples is the 1987 introduction in the US of standards by the Interstate New Teacher Assessment and Support Consortium (InTASC) (Papanastasiou et al., 2012), which set out to define effecting teaching for all learners and establish a progression towards sophisticated teaching practices (Council of Chief State School Officers [CCSSO], 2013). More recently, Darling-Hammond et al. (2023) put forward 'the "what" of teacher education' (p. 4), a synthesis of research recognizing the social and cultural context in which teaching occurs and how effective preparation programmes can support future teachers in developing characteristics of high-quality teaching.

While professional standards vary greatly in detail and encompass a wide range of dimensions, they can be broadly categorized into three fundamental areas of focus: essential subject matter knowledge; pedagogical content knowledge; and professional values and dispositions. Effective teaching emerges from the synergy of these dimensions, as it hinges on imparting specific content (subject knowledge) through proficient instructional techniques (pedagogical knowledge), implemented through and underpinned by an overarching set of professional skills and attributes (values and dispositions). Wyatt-Smith and Looney (2016) recognized professional standards as the 'codified representations of teachers' work' (p. 805). However, prior research pointed out that accreditation bodies and many professional standards are government-centric and, at times, leave out professional judgement and overall teacher professionalism as a concept (Papanastasiou et al., 2012; Yinger & Daniel, 2010).

There are clearly many challenges associated with making judgements about teacher candidates' practice and the quality of teaching. According to McLean Davies et al. (2015), the standard practice among those observing lessons of student teachers is to provide a critique with feedback, which quite often reveals more about the evaluators' preferences than pupils' learning in the classroom. Haigh et al. (2013) have given extensive consideration to these challenges and came up with a list that, while by no means exhaustive, elucidates many of the challenges we might consider: different purposes of evaluations; impact of context on practice; who defines good practice; equity in assessment; expectations of evaluators; normative standards; accountability; transparency; procedures adopted; constraints of money

and time; and issues of reliability (p. 2). The integration of theory and application during practical experiences in schools remains an abiding concern for systems globally, with judgements as to effective practice and their concomitant criteria at its centre (Conroy et al., 2013).

It is within this context and with these complexities that we consider the nature of the judgements, and the rationales for these, made by individuals who judge teacher candidates' readiness to teach. In doing so, we consider it essential to first clearly define what we mean by judgement. The Oxford Learner's Dictionaries (n.d.) defines judgement as 'the ability to make considered decisions or come to sensible conclusions'. It is the process of forming an opinion, making a decision, and assigning value to a programme, process, or individual; the judgement is the result of our deliberations. Professional judgements, and related actions, are guided by the evidence and agreed standards of the profession and occur within the close working relationships of that professional community (Organisation for Economic Co-operation and Development [OECD], 2018). In this, we wish to distinguish judgement from other epistemic considerations such as knowledge; while judgements most often entail knowing certain kinds of things, these are not synonymous, but require other, often synthetic capacities of discrimination, homologization, integration, and synthesis as well as psychological distancing. Thus, in order to understand professional judgement, it is necessary to differentiate between *professional* judgement and *personal* judgement, the essence of the former lying in the context and criteria used for decision-making and the consequences of these judgements.

Human judgement, according to Wyatt-Smith et al. (2024), is a core element in many professions, underpinning the translation of evidence into action. Professional judgements typically occur in a more formal context and are guided by expertise and cohere to professional standards of the field; these judgements are expected to align with the goals and principles of the profession (Porter et al., 2001). Professional judgement is informed by specialized knowledge, education, and expertise. Decisions made in this capacity can have broad consequences, affecting organizations, stakeholders, and the pupils themselves, and professionals are held accountable for the outcomes of their decisions. Those making professional judgements are accountable to their employers, clients, regulatory bodies, and stakeholders, and they are often held to higher standards, thus often facing consequences for any perceived failure. Such professional judgements are not made by reproducible formulas; they are generated through application of wisdom and authority with discretion (Wyatt-Smith et al., 2024), especially in matters affecting consequential decisions such as entry into the teaching profession. Such wisdom is, itself, the product of the kinds of synthetic abilities we have adumbrated above. By the same token, inadequate judgements or failures of judgement, as well as the inability to distinguish the personal from the professional, can lead to undesired variability resulting in manifest injustice and the diminishment of trust. Professional judgement is therefore distinct from personal judgement, which is predicated on individual opinions, personal experiences, preferences, values, and even emotions, and may not necessarily follow a standardized process, meaning decisions can vary widely (Kahneman et al., 2021). Here, individuals are primarily accountable to themselves and possibly those

directly affected by their decisions. Ultimately, while personal judgement is more subjective and individual-centred, professional judgement is expected to be objective, informed by expertise, and guided by established standards within a profession. Thus, just as teachers work to create conditions that allow for pupils to learn without fear of failure and to be vulnerable as they grapple with new ideas and challenging knowledge, teacher education must attend to co-creating experiences in schools that are free from fear of unfair judgement so that future teachers can engage with the complexities of learning their craft.

1.3 Transformative Work

The changing shape of teacher education requires a richer understanding of the nature of judging effectiveness of new teachers in cooperation with school partners. As the UNESCO (2021) report noted, universities are uniquely positioned to shape the next generation of educators (p. 88), and yet the commitment to teacher education jointly implemented with schools must be elevated, in particular if what Goodwin (2023) has called 'dead ideas' are to change. As Beck (2023) has illustrated, there are substantial and powerful recuperative forces at play in education whereby initiatives that appear to carry the seeds of constructive progress and the possibilities of change deliver much less than is promised. In this sense, change is often incremental and constrained by a range of material and political constraints. As Beck further demonstrated, those who create the agenda (often civil servants) get to shape the desired outcomes!

The transformative aspects of our research relate to the degree to which established norms are challenged in three key aspects: the ways in which classroom-based mentor teachers and university-based teacher educators judge ITE students' performance; who institutions rely on to judge teaching effectiveness (i.e., school-based mentor teachers, associate tutors, and university staff); and how ITEs use concomitant judgements of teaching effectiveness, particularly in a context of evolving power dynamics.

This project aims to be transformative for participants involved in the project through sharing best practices, enabling peer observation, developing new knowledge, building networked relationships, deepening commitments to reliability in judgements (Education International & UNESCO, 2019), reconciling tensions and dilemmas given the multiple positions on problems both old and new, and taking responsibility for decisions where the resulting actions impact the broader benefit of others. The project also seeks to alter the role of those who supervise and assess student teachers during their school-based experiences. This essential role of those with clinical practice orientations has largely been diminished in academia. We hope to re-emphasize the value of the 'teacher academic' as an indispensable partner in teacher education, the connector of university to school, who is instrumental in the skills of translating theory into practice. At the forefront is the willingness of TEPs to reconsider their role as intermediary organizations and focus on coordinating relationships among stakeholders and lending social capital to teacher educators positioned within the central space of schools; this is an important opportunity for development. The project is an occasion for TEPs to demonstrate adaptability and change approaches to shared judgement-making based on emerging insights.

These transformational aspects aim to adjust the way TEPs engage in the joint process of teacher education with classroom teachers and stakeholders in a spirit of co-agency. Teacher participation in the project is an opportunity to empower them to assume a more active, responsible, and effective role in the ITE process. As la Velle (2020) stated, 'learning to teach is transformative, complex and life-long' (p. 141), and mentoring teachers are at the centre of that process for novice teachers, re-emphasizing the unique and valued role that teachers hold in the social contract of education (UNESCO, 2021). We recognize in this endeavour that change is rarely linear or smooth, and learning to teach and teaching others to teach are both transformational processes. Embracing the curve of uncertainty may help us protect and transform teacher education and, in doing so, recover a bit of the great humanistic mission of teaching without denying the variability of that mission.

1.4 Aims and Research Questions

This project is focused on developing a common understanding of judgement-making on effective teaching in ITE, particularly in school-based experiences. The multi-phase project explores the nature of judgements regarding ITE students' performance per normed teaching standards. It involves partnership with teachers, researchers, and university staff in three programmes in Scotland, Wales, and England. The aims are:

- to better understand judgement processes in order to improve judgement-making on teaching effectiveness;
- to directly influence the practices of assessing and enhancing novice teachers' skills in clinical school placements, with the ultimate goal of enhancing pupil outcomes;
- to expand opportunities for dialogue across systems through a renewed sharing of practices, policies, and professional standards; and
- to meet the shared responsibility of training high-quality future educators in a sustainable model, foster networked improvement communities, and inform perspectives beyond Great Britain.

Based on these aims, the project seeks to answer three overarching research questions (RQs):

RQ1 What is the nature of shared judgement, consensus, and dissensus on observed teaching effectiveness among university-based teacher educators and school experience tutors/associate tutors and school-based mentor teachers?

RQ2 How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

RQ3 How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

This multi-phase research is designed to provide a rich picture of:

- the importance of the judgements made about teaching effectiveness (RQ1);
- the process of judging teaching effectiveness used by school-based teacher educators (i.e., mentor teachers), university-based school experience tutors/associate tutors, and university-based teacher educators (RQ1, RQ2);

- the ways in which professional teaching standards and professional judgement are perceived and used in judgement-making (RQ1, RQ2);
- the expectations for professional practice and teaching effectiveness (RQ2);
- the nature of power dynamics in the process of judging new teachers' readiness to teach (RQ2, RQ3);
- the justification as to how judgements are made (RQ1, RQ2, RQ3); and
- the influences on judgement-making (RQ1, RQ2, RQ3).

This research is underpinned by our experiences as teacher educators and the goal of ensuring children and young people are taught by a steady succession of highly qualified, competent teachers. At the heart of the research is a desire to recognize the important work that teachers do and to find opportunities, through solidarity, to enable success amid tensions in dynamic education systems.

1.5 Conceptual Framework

We view the nature of judgements regarding ITE students' performance in consideration of normed teaching standards as socially constructed and fundamentally situated; therefore, judgements must be understood in context. Social judgement theory (SJT), which emphasizes careful identification and analysis of the context of judgement, aptly supports and informs the project design (Cooksey, 1988, 1996; Hammond et al.,1977; Hovland & Sherif, 1980). SJT highlights the indicators and guidelines used by judges, making it a fitting framework from which to investigate the decisions associate tutors, university staff, and mentor teachers make in multifaceted and dynamic learning situations in each of the three ITE contexts. The theory recognizes that professional judgement is a distinctly cognitive act as well as a socially positioned practice (Allal, 2013). Judgement of new teacher performance will depend on what evaluators think constitutes effective teaching and the level of performance of the knowledge, skills, and dispositions required by normed teaching standards they find acceptable. Additionally, teaching standards themselves are socially constructed within a larger social, economic, and political narrative of teacher education (Cochran-Smith, 2003), and attempts to understand how ITE students are judged requires consideration of underlying constructs. We therefore view SJT as a tool to work with rather than a position we take (Biesta, 2020, p. 9); as a framework, it served to steer decisions and allow us to understand the differences between principles for how things are done and what happens in practice.

Examination of the nature of judgements regarding ITE students' performance in this project was therefore guided through the stages suggested by SJT (Cooksey, 1996):

1. conceptualizing the judgement problem
2. understanding the context
3. identifying the cues and dimensions for judgement-making
4. determining a sample of cue profiles
5. sampling participating judges
6. obtaining judgements
7. capturing individuals' judgement rationales
8. comparing these rationales

When those conducting evaluations make judgements about the effectiveness of a student teacher's teaching, they are synthesizing and interpreting evidence and making decisions that can have significant implications (e.g., to do with licensure or career progression). Issues around the validity and reliability of judgements arise, as teacher educators make these assessments as 'professionals working in complex, dynamic, and partially unique educational environments' (Moss et al., 2006, p. 109). In this project, we examine the contexts that socially situate evaluation of new teachers' practices at three university-based TEPs and interrogate and compare judgement policies used to assess readiness to teach.

While the suitability of SJT for this project is clear, it is essential to critique the theory so that we can address potential weaknesses in project design and methodological choices. There is a concern that SJT is too simplistic to take account of the myriad effects of variables on judgement-making, which include interpretation of evidence, the quality of an argument, the individual's position on/involvement in a particular issue, and credibility of sources (O'Keefe, 2015). Additionally, variability in human nature plays a part, with those involved in teacher education displaying individual differences and biases. To address these concerns, data collection during Phase 3 of the research included open-ended questions to explore justifications for judgements made, and the responses were further examined and validated through focus groups, which considered individuals' positions and reasoning. To describe the participants' positionality, data was also collected on qualifications, years of teaching experience and educational roles held. Furthermore, the Delphi panel process (Green, 2014), used to generate reciprocal attempts at understanding, allowed for distanciation from respective roles possibly shaped by power dynamics, thus providing opportunity for deeper insights to emerge.

SJT draws attention to the variations in evaluator involvement and the possibility that judges with similar positions may evaluate differently. It also draws attention to the value of dissensus and the potential benefits of a pluralistic approach (Moss & Schutz, 2001); with power dynamics influencing the process of judgement-making, there is potential for deeper understanding when learning comes from people's differences. Indeed, as illustrated in Chapter 8, senior professionals are wary of the deployment of technological surrogates, such as artificial intelligence, in judgement-making precisely because pathway variation is a fundamentally human characteristic of the art and science of teaching. Alert to all these subtleties, we aim, through the lens of SJT, to better understand judgement processes so that the reliability of judgement-making can be enhanced.

1.6 Structure of the Report

In Chapter 2, we present the overall project methodology, a convergent mixed methods design (Creswell & Creswell, 2023) with five phases of data collection and analysis. In this design, data from multiple phases were collected at the same time and analysed in parallel, and results were merged in the final phase to draw conclusions (Creswell & Zhang, 2009). The methodologies adopted in Phases 1–5 are briefly outlined in Chapter 2 and then fully explicated in each respective chapter.

Chapter 3 presents the findings of a systematic literature review which identified methodological trends and key areas of focus within existing literature on judgement-making on teacher effectiveness, thereby highlighting current gaps in knowledge and offering guidance on future research directions and teacher education practices. The review elucidates the processes involved in judgement-making within TEPs beyond the UK, achieved by closely examining the evaluation tools used to assess teaching effectiveness. The review offers a narrative summary of recent evidence and, as part of the larger project, serves both to broaden knowledge regarding judgements we make on teaching effectiveness and to inform the convergent research design.

In Chapter 4, we present a comparative crosswalk analysis of the professional standards for newly qualified teachers in England, Scotland, and Wales. This comparison allowed us to develop an understanding of both the universal and the particular aspects of standards informing judgement-making in the partnering institutions, on which the evaluation tools employed during school-based experiences are based. This helped the team, working across national boundaries, to understand the extent to which the different standards refer to the same thing. The comparison of current standards in the three jurisdictions is anchored alongside UNESCO's global professional teaching standards (Education International & UNESCO, 2019) and InTASC's standards (CCSSO, 2013). The resulting analysis provides novel insights into the educational standards used to judge student teachers' performance and identifies common dimensions used in judgement-making across the three jurisdictions.

Case studies of judgement-making in three TEPs, one each in Scotland, England, and Wales, are presented in Chapters 5–7. These descriptive cases are based on empirical data gathered through a video observation task, a questionnaire, and focus groups or interviews with school-based mentor teachers, university school experience tutors/associate tutors, and university teacher educators. These case studies also involved review of contextual information about teacher education provision at the three participating institutions.

Chapter 8 provides the results on expert consensus building using the Delphi panel technique. Following the Delphi method (Green, 2014), we brought together nine national and international experts in education to take up preliminary findings from Phases 1–3 of the research (presented in Chapters 3–7) in a full day of discussion and consensus building; the goal was to generate a reliable consensus opinion on the topic of judgement from a group of experts through an iterative process of questions interspersed with controlled feedback. Agreement among the experts is identified and themes related to the purpose, consistency, and nature of judging new teacher practices are presented.

In Chapter 9, we answer the RQs set out above and address project aims in response to the cross-phase and case meta-analysis from Phases 1–4. The convergent analysis sought to attend as fully as possible to available evidence, considering alternate interpretations, bringing forward the most significant aspects of the study, and situating findings in prevailing thinking and discourse (Yin, 2018). Findings relate to reliability and consistency in judging new teacher practices and why this is important.

Chapter 10 presents our argument that linear ways of considering judgement-making have overlooked important facets for effective preparation of teachers in an incredibly complex and ever-changing education system. We present a conceptual model based on convergent findings in an attempt to tease out some of these complexities and inform judgement-making practice. The model has implications for TEPs, partnering local authorities, and policymakers, which are presented as recommendations in Chapter 11. Areas for further research are also set out in the final chapter.

1.7 Conclusion

In this chapter, we introduced the multi-phase mixed methods research project undertaken to explore the nature of judgement-making processes regarding new teachers' teaching effectiveness, including how TEPs define, evaluate, and use concomitant judgements of teaching effectiveness amid contextual power dynamics. The chapter introduced the challenges and the purpose of the study. It also delineated the broader environment of teacher education that impacts on how judgements on new teachers' practices are conceptualized, conducted, and established. In the following chapter, we provide an overview of the research design and approach to data collection, offering a rationale for our choices and highlighting ways our approach was adaptable and flexible in this study of a thoroughly complex topic.

2 Methodology

Through this 2-year project, we sought to uncover the decision-making processes of those who judge teacher candidates' readiness to teach (i.e., university-based teacher educators and school-based teacher educators). We carried out a detailed investigation of what judges look for specifically when making decisions about readiness to teach. This chapter provides an overview of the methodological approach, offering a rationale for choices and highlighting that the approach was adaptable and flexible in response to a thoroughly complex topic. The project was guided by three overarching research questions (RQs):

RQ1 What is the nature of shared judgement, consensus, and dissensus on observed teaching effectiveness among university-based teacher educators and school experience tutors/associate tutors and school-based mentor teachers?

RQ2 How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

RQ3 How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in initial teacher education (ITE) shaped by power dynamics?

2.1 Research Phases

One method alone was not sufficient to fill existing gaps in knowledge about judgement-making in teacher education, particularly in the UK's devolved education context. The methods needed to capture the complex tasks involved in gaining a more complete understanding of reliability, consistency, and efficacy in new teacher practices.

This chapter describes the overall approach, a *convergent* mixed methods design (Creswell & Creswell, 2023) involving five phases of data collection. Thus, data for multiple phases were collected and analysed in parallel, and the results were merged in the final phase to draw conclusions (Creswell & Zhang, 2009). A combination of quantitative and qualitative methods of data collection and analysis was considered necessary for a robust investigation of the often tacit, and always multivariate, topic of study. The methodology adopted in each of the phases is briefly outlined in this chapter and then fully explicated in the chapters presenting findings.

Phase 1: Systematic literature review (Chapter 3): This was based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) review process (Bryman, 2016; Page et al., 2021), a thorough and transparent review process that shows the strengths and weaknesses of the review, thus giving confidence in results and allowing for potential replication. It was used to better understand: (a) methodologies and data collection tools used when making judgements about student teaching effectiveness; (b) ways in which validity and reliability are considered and conceptualized in judgement-making about new teacher effectiveness; (c) processes involved in assessing new teacher effectiveness within teacher education programmes (TEPs); and (d) how evaluation and results are used to improve judgement-making on new teacher effectiveness.

Phase 2: Professional teaching standards policy review (Chapter 4): This involved an iterative analysis of policy documents, including an assessment of the alignment of professional teaching standards in terms of content and themes (Merriam & Tisdell, 2016). This covered five sets of professional teaching standards:

- UNESCO (United Nations Educational, Scientific and Cultural Organization): *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019);
- Scotland: *The Standard for Provisional Registration* (General Teaching Council for Scotland [GTCS], 2021b);
- England: *Teachers' Standards* (Department for Education, 2021);
- Wales: *Professional Standards for Teaching and Leadership* (Welsh Government, 2009); and
- US: *InTASC Model Core Teaching Standards* (Council of Chief State School Officers [CCSSO], 2013).

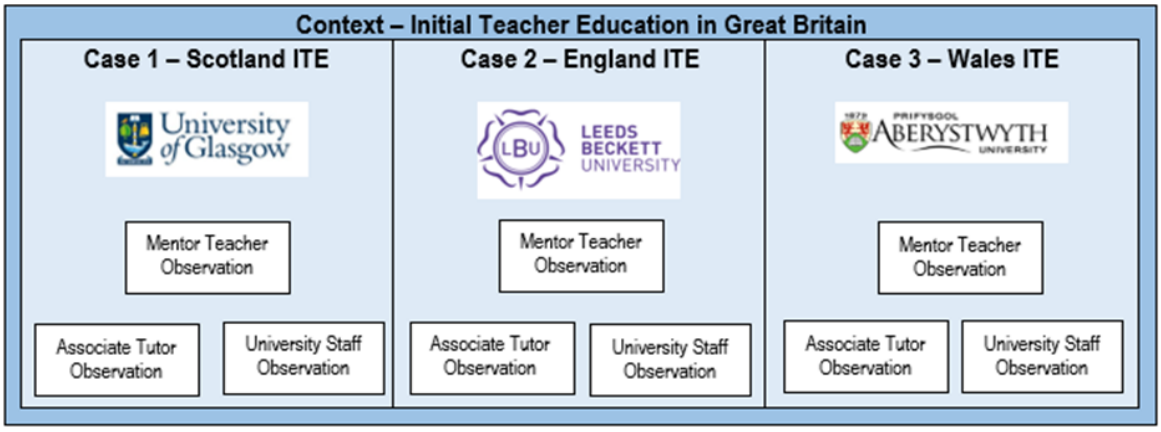
This assessment of alignment was necessary to develop an understanding of the universal and particular aspects of standards informing judgement-making within the context of the partnering institutions and the evaluation tools they employ during school-based experiences, where observations of teaching occur. Methodologically, we realize these standards are dialogical and predicated on having a grasp of how standards have been configured over time in the context of devolution in the UK.

Phase 3: Case studies of TEPs in Scotland, England, and Wales (Chapters 5–7): A comparative, embedded, and descriptive multiple case study approach (Yin, 2014) was employed, which included collection of empirical data via a video observation task, a questionnaire, and focus groups with university-based teacher educators and school-based mentor teachers (see Figure 2.1). In cases where focus groups were not preferred (i.e., confidentiality issues or scheduling conflicts arose), individual interviews were conducted. This phase also involved gathering contextual information about teacher education provision at the three participating institutions.

Aspects of Phase 3 methods common to all three cases (e.g., instrumentation and focus group questions) are described in Section 2.7. Inevitably, there was variation across the three settings, reflecting the practical realities of conducting research in different contexts. Thus, in respective chapters, nuances related to, for example, context, recruitment, and participants are explained in order to avoid a reductionist approach.

Figure 2.1

Embedded Multi-Case Research Design



Phase 4: Delphi panel technique (Chapter 8): Using the Delphi method (Green, 2014), nine national and international education experts were invited to take up preliminary findings from Phases 1–3 in a full day of discussion and consensus building; the goal of the expert panel was to generate, through an iterative process of questioning interspersed with controlled feedback, a reliable consensus opinion on the topic of judgement-making.

Phase 5: Convergent cross-phase and cross-case meta-analysis (Chapter 9): The findings from Phases 1–4 were synthesized to achieve a richer understanding and to answer the three main RQs. This involved examining the significant patterns and relationships emerging from the different lines of enquiry and 'thinking upward' conceptually to draw meaningful conclusions (Yin, 2018, p. 197).

This provided a rich picture of:

- the importance of the judgements made about teaching effectiveness (RQ1);
- the process of judging teaching effectiveness used by school-based teacher educators (i.e., mentor teachers), university-based school experience tutors, and university-based teacher educators (RQ1; RQ2)
- the ways in which professional teaching standards and professional judgement are perceived and used in judgement-making (RQ1, RQ2);
- the expectations for professional practice and teaching effectiveness (RQ2);
- the nature of power dynamics in judging new teachers' readiness to teach (RQ2, RQ3);
- the justification for how judgements are made (RQ1, RQ2, RQ3); and
- the influences on judgements (RQ1, RQ2, RQ3).

2.2 Research Settings

In this project, we considered the contexts that socially situate judgements of new teachers' practices in three university-based TEPs, and we interrogated and compared the judgement

policies used to decide on readiness to teach. The research took place at three universities that provide ITE – one in Scotland, one in England, and one in Wales (Table 2.1). An aim of the project was to stimulate collaborative, cross-institutional research to further a common understanding of effective teaching in ITE. As such, it was desirable to bring together a team from across the UK.

Table 2.1

Overview of Participating Institutions

Location	Glasgow, Scotland	Leeds, England	Aberystwyth, Wales
Institution	University of Glasgow	LBU Leeds Beckett University	1872 Prifysgol Aberystwyth University
Population size	635,000	792,500	14,650
Size of ITE programme (Undergraduate and Postgraduate Certificate/Diploma in Education)	1,300	600	50
Governing authority	General Teaching Council for Scotland	Teaching Regulation Agency	Education Workforce Council
Approximate student placements annually	2,500	750	150

Some members of the research team had prior connections through professional networks, research, and external examining. The co-investigators at the University of Glasgow initially suggested potential partners engaged in teacher education, and the Principal Investigator (PI) then reached out with an initial enquiry as to their interest in participating. The PI held virtual calls to discuss the project and potential collaboration, in which the partners confirmed their participation. A collaborative approach, while in many ways more difficult to accomplish, was desired to expand opportunities for dialogue across systems through a renewed sharing of practices, policies, and professional standards amid a substantial reform landscape (Anderson, 2023) and significant political change.

2.3 Application of Social Judgement Theory to the Study Design

Social judgement theory (SJT; Cooksey, 1988, 1996) informed design, implementation, and analysis in this project. SJT emphasizes careful identification and examination of the context of judgements and the cues and rationales (i.e., policies) used by judges, and it suggests specific stages to consider within any investigation of human judgement. This guided the

research design, with the nature of judgements regarding students' performance in ITE investigated according to the stages suggested by SJT (Cooksey, 1996). While SJT informed all phases of the study, it was particularly informative for the Phase 3 case studies. The stages considered in this research, as suggested by SJT, were:

1. **Conceptualizing the judgement problem:** This stage related to the decisions that teacher educators (based in universities and schools) make about student teachers' teaching effectiveness. It has been noted that these decisions are highly variable and idiosyncratic and that this has significant consequences for the individual student teacher and the profession. To understand this conceptualization, both decision-making as it unfolds as a performed task and the values or perspectives that influence future judgements were considered. This included the 'entangled probabilistic relationships with which a decision maker must cope' (Cooksey, 1996, p. 142) and zones of uncertainty around cause and effect related to cue information. The systematic literature review and the policy review of professional teaching standards contributed to conceptualizing the judgement problem (see Chapters 3 and 4).
2. **Understanding context:** This stage involved understanding the environment in which judgements are made. This included understanding the conditions and circumstances, including criterion measures and power dynamics, and identifying the kinds of cue information found useful by experienced judges (e.g., visual and auditory cues in an observation). This facilitated comparisons designed to highlight judgement activities. The research considered evaluation processes at each participating institution (see Chapters 5–7) and professional teaching standards of each home nation (see Chapter 4).
3. **Identifying dimensions of judgement-making:** This stage required focusing down to establish the smaller set of cues that are potentially most relevant when making judgements. It included examination of what occurs during a taught lesson and what could reasonably be included in a video observation task and accompanying questionnaire. The research design also included a consideration of the common dimensions of teaching standards across the three nations (see Chapter 4).
4. **Determining a sample of indicator profiles:** This stage involved identifying a teaching sample to simulate the natural process of teacher education. A video was selected of teaching that was representative of a classroom-based observation, similar to what would be carried out during ITE, which would be used to elicit judgements from participants. The video lasted 15 minutes and an accompanying contextual vignette was prepared.
5. **Sampling participating judges:** In this stage, a sample of judges was selected, reflecting the various roles of individuals involved in the judgement problem, which in this study meant those who conduct observations of teaching effectiveness (i.e., university-based school experience/link tutors, school-based mentor teachers and university-based teacher educators). Cooksey (1996) noted that in this stage it is important to be aware of experiential background. In this case, experiential background might impact on participants' capacity to cope with the observation task requirements. Therefore, information related to experience was collected in the case studies.
6. **Obtaining judgements:** In this stage, judgements were captured using a 15-minute video observation task and an accompanying questionnaire (see Appendix A2.1). Participants

were shown an example of new teacher practice and asked to rate this on various dimensions. This approach was taken because it was not feasible to have multiple judges observing in a real classroom, or to observe through livestreaming of a current student teacher's practice or through a video recording of a current student teacher's classroom, due to the extensive permissions required, ethical concerns, and General Data Protection Regulation (GDPR) issues.

7. **Capturing individuals' judgement policies:** This stage involved capturing the judgement strategies, or policies, participants used in the video observation task. To examine variation across the different participant groups, descriptive statistics were produced, and the rationales for judgements and cues captured through open-ended questions were explored using qualitative thematic analysis (Braun & Clarke, 2006).
8. **Comparing policies:** In this stage, comparative analyses were carried out to determine patterns of consensus and dissensus among the judges. This covered systemic influences on judgement, levels of agreement, different weighting of cues, potential predictability in the way judgements are made, and any other emergent insights from the judges. These patterns and converging data from the cross-case analysis (Morse, 1994) were then brought forward to the Delphi panel of experts for further examination and discussion (see Chapter 8).

This staged SJT framework thus shaped the multi-phase design. It enabled the capturing and comparison of judgement decisions and strategies used by participants as they determined teaching effectiveness, and it framed the wider conversation about the shared responsibility of determining readiness to teach.

2.4 Convergent Cross-Phase Analysis

A synthesis of findings from Phases 1–4 was then carried out (findings are presented in Chapters 9 and 10). The convergent analysis utilized Morse's (1994) four-stage framework: comprehending; synthesizing; theorizing; and recontextualizing. This framework was integrated with Miles & Huberman's (1994) analysis strategies: broad coding; pattern coding; memoing; distilling and ordering; testing executive summary statements; and developing propositions to achieve a richer understanding of findings. The analysis sought to attend as fully as possible to available evidence related to the RQs, consider alternate interpretations, bring forward the most significant aspects of the study, and situate outcomes within prevailing thinking and discourse (Yin, 2018).

2.5 Trustworthiness

To establish trustworthiness and ensure high-quality research, a number of strategies were used in the project design, implementation, and analysis stages. Strategies used across the entire research project are explained in this section. Considerations of validity, reliability, and dependability are more fully explicated in the chapters on findings.

Because of the continuous interaction between theoretical concerns and data collection in this complex study, several considerations, including positionality, were addressed in advance of the data collection to ensure the research design could be implemented well. All the members

of the research team were former schoolteachers who, at the time of the project, were working full time in teacher education at their respective institutions (which were the institutions participating in the case studies). The Research Associate (RA), also a former teacher, was hired specifically to work on this project and had not previously been employed at any of the participating institutions. The PI had previously worked in teacher education in the US, and this positioned the project for future collaboration and comparative research in that setting. To avoid substantiating any preconceptions the researchers had, their potential for bias was addressed through bracketing (Creswell, 2007). At research team meetings, members of the team discussed values, biases, and experiences related to the topic that could potentially influence data collection, analysis, and reporting (Merriam & Tisdell, 2016). Thus researchers held each other accountable for potential bias in the decisions, analyses, and auditing of preliminary findings.

The PI maintained the research database in the University of Glasgow protected Microsoft Teams system; access to raw empirical data was only provided to the PI and the RA. To maintain anonymity, the RA coded all potentially identifiable data before analysis was carried out by team members. Focus groups and individual semi-structured interviews were conducted by the RA, who had interview experience and, as noted, had not previously been employed at any of the participating institutions.

The triangulation of methods allowed researchers to study different but related aspects; data from multiple sources and using multiple measures served to capture the different dimensions of judgement-making and contribute to explanation building (Yin, 2018). To enhance dependability of the project, the researchers aimed for transparency of methods to show how conclusions were arrived at.

Use of the same video observation task for all participants and the use of the same questions in focus groups and interviews allowed for replication logic and checking across each of the case study locations. Triangulation in each of the three TEPs made it possible to determine with more confidence whether findings in one context were applicable across other contexts, and the Delphi panel of experts' assessment of and comments on the research findings also increased confidence in findings. In addition, as Merriam & Tisdell (2016, p. 237) noted, conducting research in an ethical manner ensures validity and reliability in qualitative research; thus the project was reviewed and approved by ethics committees in each partner institution.

2.6 Ethics

Ethical approval was gained from the University of Glasgow on 21 November 2022 (see Appendix A2.2). This approval was subsequently accepted by ethics committees at the other partner institutions. Data sharing and collaboration agreements were finalized by legal teams at the three institutions on 21 April 2023; some delays were encountered regarding translation of legalese across the devolved home nations. When one member of the project team changed employment, moving to a new institution, in March 2024, the agreements were extended to the new institution to ensure protection of data, since the team member continued to contribute to project. No data collection was conducted at this university.

All those involved in the video observation task, focus groups, individual interviews, and Delphi panel gave written informed consent to participate. The data were de-identified, with codes and pseudonyms used in place of participant names. Participant demographic data were coded and stored separately, using a naming convention system. All data were stored in a password-protected digital filing system on the University of Glasgow OneDrive, on the researchers' desktop computers, which were protected by the University of Glasgow (SSD), and in a joint Microsoft Teams folder that was not accessible by anyone outside the research team. All research team members had completed GDPR and information security training at their respective institutions.

2.7 Methods for Case Study Data Collection and Analysis (Phase 3)

Phase 3 employed a comparative, embedded, descriptive multiple case study design involving a mixed methods approach to data collection and a cross-case synthesis (Yin, 2018). The case study methods described here were applied in all three TEP settings; nuances in context, recruitment, and participant demographics are explained in the chapters reporting findings for each TEP.

The case study phase involved collection of data from several sources to provide a multidimensional examination of the three settings. A combination of the approaches of Yin (2018), who focuses on design rigour, and Merriam & Tisdell (2016), who take a constructivist-education epistemological approach, guided the case study design; these approaches complement each other in a way that met the need of this research on judgement-making (Yazan, 2015). Following Yin's (2014) explicit set of procedures, a case study protocol was developed (see Appendix A2.3), which provided the research team with general procedures and plans to be followed at each site and assisted them in anticipating problems as well as staying focused on the topic of enquiry. The protocol included five main sections: an overview; methods; procedures; data collection questions; and a guide for the final report (Yin, 2014, pp. 84–94). As the case studies were descriptive, no attempt was made to control variables, infer direct causality, or imply generalization to all judgements of teaching effectiveness.

A sample of judges who conduct observations of teaching effectiveness in the three case studies was obtained. A purposive sample was selected from each institution, including university-based teacher educators and school experience/link tutors and school-based mentor teachers from each institution as they demonstrated the perspective within each defined context and could provide enough information for in-depth exploration (Merriam, 1998).

2.7.1 Case Study Data Collection

Data collection in the case studies focused on judgements about teaching effectiveness, and the RQs were explored through two sources of data: a video observation task designed for this study with an accompanying questionnaire; and a series of focus groups and interviews with judges in each TEP (see Appendix A2.1).

2.7.1.1 Video Observation Task and Questionnaire

The third stage of the procedural methodology of SJT (Cooksey, 1996) is identifying relevant dimensions of judgement-making. This involved focusing down to establish the dimensions that were potentially most relevant so that these could be incorporated into a simulated judgement task. Dimensions of judgement-making, therefore, arose from common understandings of teaching effectiveness, from which the judge can make a decision. To establish a common understanding in this study and to help develop the observation questionnaire, we aligned UNESCO's *Global Framework of Professional Teaching Standards* with the professional standards frameworks of England, Scotland, and Wales as well as the CCSSO's *InTASC Model Core Teaching Standards* from the US ('InTASC' stands for Interstate Teacher Assessment and Support Consortium). The detailed results and findings of this analysis are presented in Chapter 3.

Next, the domains reflective of what could reasonably be observed through perceptual information, or 'cues', in a teaching video were selected. Observable domains included teaching knowledge, understanding, and practice; these are indicated in italics in Table 2.2. The other domains were related to abstract conceptions, such as values, and non-classroom-based skills, such as engaging families; as these were not observable in the video observation task, they were not included in the task.

Table 2.2

Domains Included in the Video Observation Task

Standard	UNESCO	Scotland	England	Wales	US
Domains	*I. Teaching knowledge & understanding* *II. Teaching practice* III. Teaching relations	I. Being a teacher in Scotland *II. Professional knowledge & understanding* *III. Professional skills and abilities*	*I. Teaching* II. Personal & professional conduct	*I. Pedagogy* *II. Professional learning* *III. Collaboration* IV. Innovation V. Leadership	*I. The learner & learning* *II. Content knowledge* *III. Instructional practices* IV. Professional responsibilities

Note. The standards are from: CCSSO (2013); Department for Education (2021); GTCS (2021b); Education International & UNESCO (2019); Welsh Government (2009).

Observable domains are italicized; the remainder were not observable in the video observation task.

The dimensions that could be examined within these observable domains were then specified (Table 2.3) and included in an observation questionnaire with descriptors relating to judgement-making (see Appendix A2.1). Together, the domains and dimensions formed an

agreed reference for describing the core work of teachers in the three institutions in an effort to generate knowledge to improve understanding of judgement-making.

Table 2.3

How Dimensions of Teaching Are Demonstrated in the Video Observation Task

Dimension	Practice
Learners	Shows understanding of learning and development and individual variations within and across the cognitive, linguistic, social, emotional, and physical areas; regards the needs of all individuals and the class as a whole; learning experiences are developmentally appropriate and intellectually challenging
Content	Demonstrates core knowledge and skills of the content area are being taught; learning experiences make the subject matter accessible and meaningful to learners to ensure mastery of the content
Research	Reflects core research and analytical methods that apply in teaching, including with regard to effective assessment of learners
Planning & preparation	Demonstrates planning and preparation which supports learners in reaching identified learning objectives
Instructional strategies	Includes an appropriate range of teaching activities which reflect and align with both the nature of the subject content being taught and the learning, support, and development needs of the learners; instruction facilitates engagement and integration of digital technologies
Learning environment	Demonstrates organization and facilitation of learners' activities so that they can participate constructively in a safe and secure environment and in a cooperative manner; the learning environment encourages positive social interaction, active engagement in learning, and self-motivation
Assessment	Demonstrates consistent, fair, valid, and reliable assessment of student learning using an appropriate range of methods to evaluate attainment of learning objectives

Note. These broad descriptions of practice were synthesized from: CCSSO (2013); Danielson (2007); Marzano et al. (2011); Education International & UNESCO (2019).

To further study judges' value systems and determine the level of consistency among them, these dimensions were also used as the basis for capturing judges' ratings of how satisfactory the teaching practice was. Using a 5-point nominal scale, from high (5) to low (1), judges could indicate the quality of the teaching they observed. High-quality teaching reflected the degree of sophistication in the new teachers' application of knowledge and skills in the respective dimensions. It is critical to acknowledge that this assesses levels of *teaching performance*, not the *teacher* (Danielson, 2007, p. 39). Teaching effectiveness is distinct from

teacher effectiveness, with teaching effectiveness referring to strong instruction that enables a wide range of pupils to learn (Darling-Hammond, 2013, p. 12). It is in part a function of teacher effectiveness (knowledge, skills, and dispositions), which can be observed and is influenced by the context in which instruction occurs.

The video selected for the task was created by the department of education in a US state for the explicit purpose of promoting a shared understanding of instructional quality to increase reliability in assessing classroom instruction; it was freely available online (https://youtu.be/Jyh3M8SCB3M). This context, outside the jurisdiction of the three UK nations, was chosen for the video observation task as it helped disentangle some context-specific variables, which meant that participants could avoid subjective judgements stemming from their familiarity with the Scottish, English, or Welsh contexts. A high school English lesson was selected from among the available content areas, as literacy across learning is considered the responsibility of all teachers and the literacy skills of reading, writing, listening, and speaking would be commonly observed across disciplines.

The video observation task was deployed through the University of Glasgow secure Qualtrics system. Participants (i.e., judges) observed the student teacher in the video and discerned their teaching effectiveness based on the seven dimensions in Table 2.3.

The video simulated the natural process of observation in teacher education. Participants were asked provide judgements in each of the seven areas and an overall judgement of the teaching effectiveness, and to indicate which areas were most and least difficult to judge. They were also asked in open-ended responses to explain *how* and *why* they made judgement decisions in order to capture the cues utilized, their judgement policies, and the factors that potentially influence this. While the seven areas were provided before watching the video, the descriptors were presented along with the questionnaire after the video was viewed.

The second part of the task involved completion of the questionnaire on aspects of judgement-making and potential influencing factors, which had been derived from prior research (Biesta, 2020; Cameron-Jones & O'Hara, 1994; Habermas, 1996; Haigh & Ell, 2014; Haigh et al., 2013; Hand & Rong, 2014; Hattie & Clinton, 2001; Hegender, 2010; Hoy, 1994; Johnson, 2013; Menter, 2016; Moss et al., 2006; Murray-Harvey et al., 2000; Raths & Lyman, 2003; Schmoker, 2006; Wyatt-Smith & Klenowski, 2013). These influencing factors were rated on a 7-point scale from strongly agree (7) to strongly disagree (1), with a neutral option (4). The questionnaire is included in Appendix A2.1.

The video observation task and accompanying questionnaire was piloted with two university TEs and four teachers not currently in the role of mentor teacher to ensure the dimensions were perceived as genuine and the instructions for the task contained enough information. It was also evaluated for participant time burden, clarity of instructions, errors, and any difficulties or skip patterns. The task was then adjusted based on feedback, and the final version was developed. The questionnaire was reviewed by two research officers from the Robert Owen Centre for Educational Change at the University of Glasgow and deemed 'fit for purpose'. For the Welsh context, the evaluation task description and the questionnaire were translated into Welsh by project partners at Aberystwyth and circulated in English and

Welsh. The task was then distributed, based on purposive sampling at each TEP, via the University of Glasgow secure Qualtrics account. Recruitment strategies specific to each TEP are explained in the respective chapters on case study findings.

2.7.1.2 Focus Groups and Individual Interviews

Focus groups were organized in each TEP (see Figure 2.1) to facilitate a discussion about the results of the video observation task and to corroborate initial analysis identifying judgement policies. Participants were self-selected via the video observation task: the final query of the task provided information about the focus groups being arranged and asked 'Are you willing to participate in a focus group to discuss your response and initial results of the study?' Those who selected 'yes' were then prompted to provide contact information so they could be invited to a focus group. The focus groups for all TEPs were facilitated by the RA, who had not been involved in any ITE processes in the three TEPs and thus was a neutral member of the team and suited to this role. There were some instances where individuals could not contribute in a focus group setting due to confidentiality or scheduling conflicts, so instead they took part in individual interviews.

A set of semi-structured questions was used to facilitate the discussion. The individual interviews followed the same interview protocol as the focus groups. To prevent any confusion or misremembering of the dimensions that participants had been asked about when assessing the candidate featured in the video, the definitions of each of the dimensions were made available on a PowerPoint slide or the interviewee could ask the RA to read them out at any point.

The protocol was adjusted based on the results of the evaluation task for each group of teacher educators (e.g., mentor teachers) and each institution. For instance, if there was consistency in the evaluations, questions focused on exploring the reasons for this consistency. Conversely, when inconsistencies were revealed, questions investigated the reasons behind the variation. The questions included:

- What could be the reasons for consistencies (or inconsistencies) between raters?
    - *Follow-up:* Is it okay that there are consistencies (or inconsistencies)?
- What would make judgements of evaluators more consistent?
    - *Follow-up:* I wonder, what should we be after if not consistency?
- We found that, among the seven evaluation areas rated, the most explicit inconsistency (consistency) in judgements was found for 'learners' (or 'content' or 'research', or 'planning & preparation' or 'instructional strategies' or 'learning environment' or 'assessment'). Interestingly, this area was also selected as the easiest (or the hardest) element to evaluate in the video. What are your thoughts on the finding that 'learners' (or one of the other dimensions) had the most inconsistent (consistent) rating but was also considered easiest (or hardest) to rate?

  Alternative question: We found a high (or satisfactory) degree of variability (or consistency) in relation to the area selected as the easiest and most difficult element to

evaluate in the video. So, what are your thoughts on the finding of these inconsistencies (or consistencies) among raters?

- What are your views on using professional judgement to assess teaching effectiveness? After that, could you please also tell me about your views on using professional standards to judge teaching effectiveness?
    - *Prompt:* How do you find that standards affect what is defined as effective teaching, and how does that influence people's judgement?
- How might schools and universities work together to gain greater reliability in evaluation of teaching effectiveness?
- Is there any barrier or asset you would like to bring attention to that would impact schools and universities working together?
- Is there anything you would like to add about reliability and consistency or inconsistency in judging teaching effectiveness, from your perspective?

The focus groups and individual interviews were conducted and recorded via the University of Glasgow secure Zoom account and auto-transcribed by Zoom; the facilitator rewatched the recording and made any necessary corrections to the transcripts. When participants had difficulty accessing Zoom at the school they were based in, Microsoft Teams, accessed via the University of Glasgow, was used as an alternative platform to conduct the interviews. Transcripts of the focus group discussions, which lasted around 45 minutes, contributed to the qualitative data set for this study. Data were analysed with the aim of better understanding how the participants arrived at judgements. With the data from both individual and focus group interviews, the analysis focused on the consensus and dissensus that emerged among the judges in relation to the video observation vignettes. The analysis first examined the decisions made by each participant, looking at individual differences that might impact decision-making and then focused on the response to the video vignette.

2.7.2 Single-Case Analysis

Quantitative data included ratings of teaching effectiveness and Likert-type responses to the questionnaire. Descriptive statistics (e.g., frequency, mean, and standard deviation) were produced to examine variance across the participant groups and TEPs (Pyrczak & Oh, 2018). A trend analysis was carried out to examine the level of consistency among raters. Comparative analysis was used to examine patterns of consensus and dissensus.

Qualitative data from the open-ended questions in the questionnaire and the focus group and interview transcripts were analysed using the constant comparative method (Glaser & Strauss, 1967) to construct inductive codes, categories, subcategories, or themes. This iterative process was carried out across the three participant roles at the three locations. Data were explored through the six steps of qualitative thematic analysis: familiarization; initial coding; generating themes; validating themes; defining themes; and interpreting and reporting (Braun & Clarke, 2006). To ensure reliability, guidelines on thematic analysis (Morse, 1994) were adhered to.

The analysis process involved two stages, each comprising multiple steps. In the first stage, a template was created for each open-ended question on the questionnaire and each focus group question; these were used to populate participant responses (from 10 questions in the questionnaire and from 7 questions in the focus groups and individual interviews). For each set of responses, a member of the research team identified the core ideas underlying each participant's assessment and then looked for emerging patterns across all participant responses. Lastly, an independent audit was conducted by another member of the research team to determine a consensus of findings. In the second stage of analysis, data were cross-analysed for core ideas within each set of responses, resulting in themes and answers to the RQs. Consensus was achieved between the initial researcher and the auditor, and the results are presented in Chapters 5–7.

Researcher memoing was used during analysis to foster reflexivity and assist the researchers in their ongoing analytical thinking about concepts and themes. An example memo is provided in Table 2.4. These memos provided an important step in which concepts derived from the qualitative data were used as building blocks for synthesizing findings and constructing an argument.

Table 2.4

Example of a Researcher Data Analysis Memo for a Focus Group Question

Focus group question: What could be reasons for inconsistencies between raters?
Data: The first one that springs to mind is from my memory, the lesson seemed to be an end lesson, where lots of underpinning had been done in previous lessons, or in part of the lesson that we weren't able to see. So, the learners seemed to already know what was expected of them. And therefore, on the day, the person who was teaching didn't have so much instruction to do, because she was reminding them what they had already done. So, perhaps the perception of some of the mentors was that they didn't see that teaching taking place. I personally felt it was okay, because I could see from the learners responses and how they were interacting with the teacher that that learning had taken place, and that it had been reasonably effective and successful. And so that was my view. So, I could have been wrong because I was making assumptions based on how I saw the learners reacting. And maybe somebody else with a different kind of mindset would not want to make that kind of presumption. They may even, it may even have been a different teacher who did that learning and not the one that we were watching. That's something else that that might have made them think, well, I didn't see that happening. And we kind of saw the end result. And then there are maybe just personal preferences as to how to teach. The teacher was, to me, very animated and taught probably quite differently than we do in British schools. You know, quite animated, but it kind of suited her subject, I felt. I felt it suited her subject. It suited her personality. The students definitely responded to it. And so, I was quite happy to kind of go with it.

Researcher memo:

- Reality of the way the school functions and interdisciplinarity means pupil learning cannot necessarily be attributed to an individual teacher.
- Also speaks to how judgements are made: learners responses, interactions with the teacher, if learning had taken place (something new – reached the objectives).
- Would we call making a judgement about teaching based on the learners response an assumption? Or is this use of observation and professional judgement?
- If one lesson isn't enough, could it truly be summative – exhibiting the necessary skills at some time during multiple observations or a day of teaching? This requires a portfolio and the mentor teacher observing over a more extended period of time and using different sources of evidence to show competence. The observation itself could be used, but only for that which might provide evidence of certain requirements.

2.7.3 Cross-Case Analysis

A cross-case analysis was conducted to build a general explanation that fit the three cases, giving consideration to the details specific to each case (e.g., differences related to the devolved educational settings; Merriam & Tisdell, 2016). The analysis was carried out using Morse's (1994) four-stage framework: comprehending; synthesizing; theorizing; and recontextualizing. The analysis sought to identify relationships, contradictions, and consistencies, and it highlighted key findings in order to situate the results alongside prior research on judging teaching effectiveness.

2.8 Conclusion

This chapter details the methodological decisions made in the five phases of the project, which incorporated quantitative and qualitative methods. The complexity of the RQs and the focus on the nuances of judgement-making warranted a mixed methods approach. The chapter outlines how the methods, grounded on SJT (Cooksey, 1996), stayed true to this complexity. The chapters that follow share findings from each phase of the research. Chapter 3 covers the systematic literature review on judgement-making processes. Chapter 4 presents the analysis of professional teaching standards. Chapters 5, 6, and 7 cover the case studies of judgement-making tasks in the TEPs in Scotland, England, and Wales, respectively. Chapter 8 examines the consensus outcomes of the Delphi panel. In Chapter 9, the convergent findings across phases and across cases are discussed. Chapter 10 presents an emerging model based on the findings, which aims to inform judgement-making. Finally, Chapter 11 contains conclusions and recommendations.

3 Systematic Literature Review: Judgement-Making on Teaching Effectiveness

3.1 Introduction

In Phase 1 of this project, a systematic review was carried out to gather together knowledge regarding judgement-making on teaching effectiveness. The aim was to better understand: (a) methodologies and data collection tools used when making judgements about student teaching effectiveness; (b) ways in which validity and reliability are considered and conceptualized in judgement-making about new teacher effectiveness; (c) processes involved in assessing new teacher effectiveness within teacher education programmes (TEPs); and (d) how evaluation results are used to improve judgement-making on new teacher effectiveness.

Using the theoretical framework of social judgement theory (SJT; Cooksey, 1996), the systematic literature review helps conceptualize the judgement problem under investigation. The review identifies methodological trends and key research themes within the existing literature on judgement-making on teacher effectiveness, highlighting current gaps in knowledge and guiding future research and practice in teacher education. As well as broadening knowledge regarding judgement-making on teaching effectiveness, the findings informed the convergent research design of the larger project.

Section 3.2 outlines the background to the systematic review and the methods used. The findings of the review are presented in Sections 3.3 to 3.5. Section 3.3 elucidates the processes involved in judgement-making in teacher preparation programmes (TPPs), achieved by close examination of the evaluation tools used in assessing teaching effectiveness. Section 3.4 offers a narrative summary of the research evidence concerning the judgement of teaching effectiveness in the studies. Section 3.5 concludes the chapter.

3.2 Systematic Review Methods

This section introduces the research aims and questions that guided our systematic literature review. It then details our search strategy and screening procedure for identifying relevant studies as well as our approach to extracting data from the studies that were selected for inclusion. We used the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) framework (Page et al., 2021) while also following established systematic review methods outlined by Bryman (2016). In addition, the study summaries underwent thematic analysis following Braun and Clarke's (2006) six-step framework. Recurring themes from the findings of the review were identified inductively, iteratively, and collaboratively by the PI and RA.

3.2.1 Research Aims and Questions

This phase of the study aimed to:

1. explore the most recent evidence related to judging teacher effectiveness in the UK and other countries
2. document subject-specific and methodological approaches in the most recent literature

3. map and analyse the processes involved in judgement-making of teaching effectiveness

The systematic review was guided by three key questions, which influenced the choice of search terms and criteria for selecting studies.

1. What types of evidence are used to make judgements of teaching effectiveness?
2. How are reliability and validity of evaluations/tools used in judging student teachers assessed?
3. How are the findings relating to judgements of teaching effectiveness used?

3.2.2 Search Strategy

The search strategy involved conducting electronic searches across established and credible scientific databases. A total of 19 databases, all centred around or relevant to education, were searched, and these are listed in Table 3.1.

Table 3.1

Databases Searched

Atla Religion Database (via EBSCOhost)
British Education Index (via EBSCOhost)
Child Development & Adolescent Studies (via EBSCOhost)
Education Abstracts (via EBSCOhost)
Educational Administration Abstracts (via EBSCOhost)
ERIC (via EBSCOhost)
Professional Development Collection (via EBSCOhost)
PsycBOOKS (via APA PsycNet)
PsycINFO (via EBSCOhost)
Teacher Reference Center (via EBSCOhost)
The Philosopher's Index with Full Text (via EBSCOhost)
ProQuest Academic
Australian Education Index (via Proquest)
ERIC (via Proquest)
Social Science Premium Collection (via Proquest)
Oxford Research Encyclopedia of Education
Nexis
Emerging Trends in the Social and Behavioral Sciences
International Encyclopedia of Education

Search terms were chosen based on the research questions, and the search strings used are provided in Table 3.2.

Table 3.3 provides the criteria used to identify studies for potential inclusion in the review. Searches were limited to peer-reviewed articles written in English or Welsh, and to obtain the most recent evidence, only articles published from 2010 were included (this meant that the timeframe for published studies was 2010–2023). Searches were not limited by research methodology, so a range of empirical and non-empirical studies could be included.

Table 3.2

Search Strings

‘judgement’ OR ‘evaluation’ OR ‘assessment’ OR ‘professional judgement’ OR ‘shared judgement’
AND
‘reliability’ OR ‘consistency’ OR ‘trustworthiness’ OR ‘validity’
AND
‘teaching effectiveness’ OR ‘teacher effectiveness’ OR ‘teacher quality’ OR ‘teacher preparedness’ OR ‘teacher preparation’
AND
‘initial teacher education’ OR ‘beginning teacher*’ OR ‘teacher candidate*’ OR ‘novice teacher*’ OR ‘new teacher*’ OR ‘teacher education’ OR ‘teacher training’ OR ‘educator preparation’

Table 3.3

Inclusion Criteria

- Peer-reviewed article
- Published in English or Welsh
- Published during 2010–2023
- From any country
- Study includes one or more of the following aspects of teacher evaluation:
  - how (pre-service and in-service) teachers’ classroom practice is evaluated
  - criteria used to judge teaching effectiveness
  - validity and trustworthiness of teaching evaluation instruments
  - relationships between raters and candidates or new teachers (i.e., power relationships)
  - consistency of judgements of teaching practice

3.2.3 Screening of Articles

After the database searches, following the recommendations of the PRISMA guidelines for systematic reviews, studies were screened in a multi-step process to identify relevant studies for inclusion in the review (Figure 3.1).

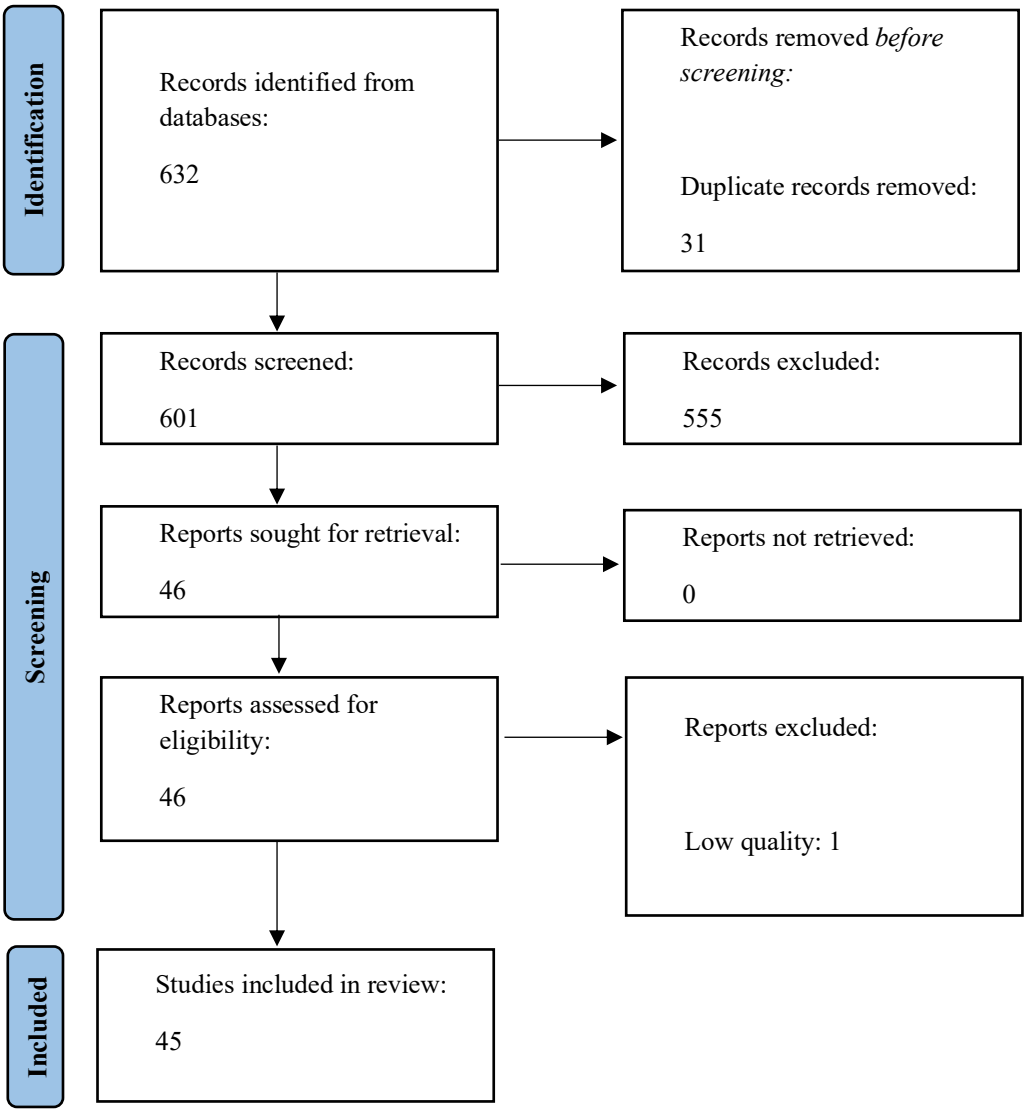
The Research Associate (RA) conducted the database searches and retrieved studies, and the Principal Investigator (PI) confirmed the results. All identified studies were exported to Rayyan to facilitate the screening process (Ouzzani et al., 2016).

A total of 632 peer-reviewed studies were retrieved following the database searches, and 31 duplicates were removed, leaving 601. The abstracts and titles of these studies were screened by the RA, using the criteria outlined in Table 3.3. Initial disagreement about whether to include 7 of these studies was successfully resolved through discussion, and these, along with 548 other studies, were excluded, leaving 46 studies. There was a high percentage of agreement (98.8%) between the RA and the PI; Cohen’s kappa coefficient was 0.91, indicating ‘almost perfect agreement’.

The Research Associate (RA) conducted the database searches and retrieved studies, and the Principal Investigator (PI) confirmed the results. All identified studies were exported to Rayyan to facilitate the screening process (Ouzzani et al., 2016).

Figure 3.1

The Study Selection Process



Note. Adapted from Page et al. (2021)

A total of 632 peer-reviewed studies were retrieved following the database searches, and 31 duplicates were removed, leaving 601. The abstracts and titles of these studies were screened by the RA, using the criteria outlined in Table 3.3. Initial disagreement about whether to include 7 of these studies was successfully resolved through discussion, and these, along with 548 other studies, were excluded, leaving 46 studies. There was a high percentage of agreement (98.8%) between the RA and the PI; Cohen's kappa coefficient was 0.91, indicating 'almost perfect agreement'.

Full-text screening of the remaining 46 studies was guided by *relevance* (fit with the inclusion criteria) and *quality*. No studies were excluded because of lack of relevance, but one was excluded because of poor quality (it lacked research aims, research questions, and

clear methodology). Therefore, the screening process resulted in selection of 45 studies. These studies are referred to in this report according to assigned numbers. Table 3.6 provides the study number and citation for each study, and the full references are provided in Appendix A3.1.

3.2.4 Summarizing Studies for Data Extraction and Analysis

During this stage, the 45 studies selected for inclusion in the review were read thoroughly by the RA and summarized in a Word document following the framework outlined in Table 3.4. This allowed the RA to identify initial themes. The PI examined the initial themes and cross-checked them against each study to validate the decisions made by the RA.

Table 3.4

Framework for Summarizing Study Characteristics

Study number	A number was assigned to each study.
Citation	The citation was recorded for each study.
Study aim(s)	The original statement of the overarching aim or objective of the study was copied directly. Where no aim or objective was included in the study, research questions were reworded as the study aim.
Research question(s)	The original research questions were copied directly. Where no research question was included in the study, the study aim or objective was reworded as the research question. Where no research question, aim, or objective was included, the research question was inferred based on the analysis presented in the study.
Research focus and context	The research focus, research aims, and contextual information about the study were noted. Where the focus of the study was broader than judging teaching effectiveness (i.e., reviewing effectiveness of a teacher preparation programme) only the aspect related to judgement-making was summarized.
Evaluation instrument and context	Information about the use of an evaluation tool was summarized, including the specific evaluation tool used and the setting (e.g., candidate evaluation, new teacher evaluation). When the tool was not named, a name was assigned by the RA based on the main focus and the intended purpose of the evaluation.
Methodology	The data collection methods, data collection tools, and sample size were summarized. Where the data collection method was not explicitly stated, this was inferred. Mixed methods studies which collected mainly quantitative data were marked with an asterisk to indicate the predominance of quantitative data.
Findings	Key findings pertinent to our study were summarized. To enhance readability and bolster the reliability and transparency of data analysis, we explicitly indicated findings related to respective themes of our study. Findings on

	development and design of teaching evaluation tools, implementation of teaching evaluation tools, and rater consistency were summarized.

Summaries also captured information relevant to the inductively identified themes and sub-themes shown in Table 3.5. Our descriptions of the sub-themes are included in the table, though we are aware that there is some blurring of the definitions of validity and reliability in the literature (Cohen et al., 2018). All summaries were audited by the PI three times, and once by a Cooperating Investigator, to ensure accuracy of results.

Table 3.5

Identified Themes and Sub-Themes

Main theme	Sub-theme	Description
Instrument development and implementation	Instrument development	Who developed the tool Basis for development of the tool – whether it was grounded on theoretical work, empirical findings, or standards
	Instrument structure	Format and components of the tool (e.g., the evaluation domains, the number of items it includes) Psychometric properties – details of the reliability, validity, and other psychometric characteristics of the tool
	Instrument implementation and result use	How the tool is implemented (e.g., the number of raters, the number of evaluations) Training for raters – whether training is available for raters and details about the training that exists Results use – how the results obtained from the tool are intended to be used
Reliability	Consistency and accuracy	Consistency – the degree of agreement in ratings by different observers (i.e., inter-rater reliability), by the same observer (i.e., intra-rater reliability), and over time (i.e., test–retest reliability) Accuracy – how close a measure or observation is to the 'true' construct being measured
	Internal consistency reliability	Extent to which the items or components within a tool consistently measure the same underlying construct (i.e., teaching quality)
	Influences on rater reliability and how to improve it	Causes of bias, inconsistency, and inaccuracy in judgement-making and suggestions for how to improve reliability
Validity	Face validity	Whether the content of the test is suitable for the aims (e.g., user perspective, experience, satisfaction)
	Content validity	Degree to which a test comprehensively represents the entire scope of the construct it aims to measure, with the test items adequately covering all relevant aspects of the defined content area

Construct validity	Extent to which a construct is accurately defined and operationalized to measure the intended construct without interference from other constructs
Predictive validity	How well an evaluation predicts future outcomes
Consequential validity	Broader consequences, including the impact of assessment on ratees
	Extent to which actions based on the evaluation are both legitimate and fulfilled

Note. Definitions are adapted from Cohen et al. (2018).

3.3 Findings From the Systematic Review

This section presents key findings from our review of the 45 included studies, organized within four distinct themes to shed light on trends in the literature and identify gaps that demand further investigation. The section starts with an overall summary of the studies in Table 3.6, including: the study number and citation; a study description; the main themes and sub-themes; contextual information about the research; the research methods used and type of evidence gathered; and the data collection tools and participants. This serves as a navigational guide for the rest of the section.

Section 3.3.1 covers the research themes and research context of the 45 studies. Section 3.3.2 provides an examination of TEPs and the evaluation tools identified in the studies. Section 3.3.3 presents evidence regarding the reliability of these tools, and Section 3.3.4 presents evidence on the validity of the tools.

Table 3.6

Summary of Studies Included in the Review

	Publication	Study description	Theme and sub-theme	Study context	Methods and evidence type	Data collection tools and participants
1	Hylton et al. (2022)	Validation of a candidate evaluation instrument that was developed and used in a TEP	V: construct validity*	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on mid-term and final rating of candidates (n = 1,486), including self-rating and rating by school-based and university-based teacher educators
2	Dewaele et al. (2021)	Factors influencing pre-service teacher evaluations of an instructor	R: influences on rater reliability and how to improve it	Schoolteacher evaluation, 2+ TEPs across universities, Germany and Australia	Empirical, mixed methods,** primary	Evaluation instrument, including a comment section for explanations, used by candidates (n = 266) to rate a teacher
3	Tobón et al. (2021)	Validation of an emerging evaluation tool, SOCME-10, for new teacher populations	V: construct validity*	Schoolteacher evaluation, not in the context of a TEP, Mexico	Empirical, mixed methods,** primary	Questionnaires, with a comment section, to experienced schoolteachers (experts; *n* = 21) and new teachers (*n* = 25) Evaluation instrument used by new teachers (n = 557) to self-rate
4	Tanguay (2020)	Perspectives of university-based teacher educators on edTPA (Educative Teacher Performance Assessment) as a standardized state-mandated tool	V: face validity*	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, qualitative, primary and secondary	Interviews with university-based teacher educators (*n* = 8) Documents related to candidates and the TEP – artifacts, programme workshop materials
5	Sandoval et al. (2020)	Alignment of a TEP with edTPA outcomes related to equity	V: consequential validity	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, qualitative, secondary	Documents related to candidates – course essays (*n* = 53) portfolios (*n* = 9)

6	Roloff et al. (2020)	Predictive validity of entry characteristics and teacher education grades on teachers' future instructional quality	V: predictive validity	Evaluation of early career teachers in school context, 2+ TEPs across universities, Germany	Empirical, quantitative, primary and secondary	Evaluation instrument used by classroom students (n = 3,768) to rate schoolteachers (n = 113)Questionnaires to teachers (*n* = 113) Administrative records related to teachers during teacher education
7	Shahzad & Mehmood (2019)	Development and validation of an emerging evaluation tool for higher education teaching (Teaching Effectiveness Scale) to be used by university students	V: construct validity*	Higher education lecturer evaluation, not in the context of a TEP, Pakistan	Empirical, mixed methods, primary	Interviews with higher education lecturers (experts; *n* = 10) Focus groups with graduates (*n* = 3 groups) Literature search Subject matter experts questionnaires (*n* = 16) Evaluation instrument used by higher education students (*n* = 698) to rate higher education lecturers
8	Yahiji et al. (2019)	Examination of an assessment model used in field experience (validity, reliability, objectivity, practicality)	V: face validity	Candidate assessment, single TEP in a university, Indonesia	Empirical, mixed methods, primary and secondary	Questionnaires and focus groups with experts – university-based teacher educators (*n* = 14) and school-based teacher educators (*n* = 14) Documents related to candidates – assignments
9	Mkhasibe et al. (2018)	Comparing teacher mentors' and university supervisors' perceptions of student teachers' readiness to teach	R: consistency and accuracy*	Candidate assessment, single TEP in a university, South Africa	Empirical, qualitative, primary and secondary	Focus group with school-based teacher educators (*n* = 12 participants) Pre-existing observation reports prepared by university-based teacher educators (n = 3) as part of their evaluation of candidates
10	Basit & Khurshid (2018)	Satisfaction of teacher educators and candidates with candidate assessment techniques	V: face validity	Candidate assessment, 2+ TEPs across universities, Pakistan	Empirical, quantitative, primary	Questionnaire to university-based teacher educators (*n* = 300) and candidates (*n* = 890)
11	Ata & Kozan (2018)	Construct validity and reliability of Intern Keys Teacher Candidate Assessment, based on interpretable factor structure	V: construct validity*	Candidate evaluation with an authentic tool, 2+ TEPs across universities, US	Empirical, quantitative, primary	Evaluation instrument used by university-based teacher educators (n = 116) to evaluate candidates

12	Goldhaber et al. (2017)	Predictive value of edTPA scores on workforce entry and teaching quality	V: predictive validity*	Early career teachers in school context, 2+ TEPs across universities, US	Empirical, quantitative, secondary	Administrative records related to candidates (n = 2,362) Student achievement data for employed teachers (n = 277)
13	Kennedy & Lees (2016)	Candidates' growth through Classroom Assessment Scoring System (CLASS) scores supported feedback and tiered support	V: consequential validity	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, mixed methods, primary and secondary	Focus group with candidates (n = 19) and pre-existing evaluation results of same candidates (n = 19)
14	Masuwai & Saad (2016)	Face and content validity (representativeness, relevance) of an evaluation instrument (Teaching and Learning Guiding Principles Instrument)	V: content validity*	Teacher educator assessment, 2+ TEPs across universities, Malaysia	Empirical, mixed methods,** primary	Questionnaires, with a comment section, to university-based teacher educators (expert judgers; n = 9)
15	Brown et al. (2015)	Documenting candidates' professional growth through Profile for Evaluation of Intern (PEI)	V: consequential validity*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on rating of candidates (n = 97) including candidate self-evaluation and school-based and university-based teacher educators evaluations
16	Maharaj (2014)	School administrators' view of the Teacher Performance Appraisal for new and experienced teachers	V: face validity*	Schoolteacher evaluation, not in the context of a TEP, Canada	Empirical, mixed methods,** primary	Questionnaires, with a comment section, to school principals and vice-principals (n = 166)
17	Kingsley & Romine (2014)	Construct validity, dimensionality, and reliability of Item-Level Assessment of Teaching Practice (I-LAST), a learning-oriented evaluation tool	V: construct validity*	Candidate assessment, single TEP in a university, US	Empirical, quantitative, primary	Evaluation instrument used by school-based teacher educators (n = 46) to rate candidates Questionnaires to university-based teacher educators (n = 3) and school-based teacher educators (n = 3)

18	Hamid et al. (2012)	Predictive value of teachers' cognitive ability and personality for performance	V: construct validity*	Schoolteacher evaluation, not in the context of a TEP, Malaysia	Empirical, quantitative, primary	Evaluation instrument for self-rating by schoolteachers (n = 1,366)
19	Smalley & Retallick (2012)	Evaluation practices in agricultural TPPs	IDI: instrument implementation and result use*	Candidate assessment, 2+ TEPs across universities, US	Empirical, quantitative, primary	Questionnaires to coordinators of agricultural education TPPs (n = 66)
20	Ritzhaupt et al. (2010)	Candidates' perspectives of e-portfolios	V: face validity	Candidate assessment, 2+ TEPs in a university, US	Empirical, quantitative, primary	Questionnaires to candidates (n = 224)
21	Beare et al. (2014)	Examining bias in employment supervisors' evaluations of new teachers based on their socioeconomic status and ethnicity, using data from the Systemwide Evaluation of Professional Teacher Preparation Programs	R: influences on rater reliability and how to improve it*	Early career teachers in school context, 2+ TEPs in a university, US	Empirical, quantitative, secondary	Pre-existing evaluation results from 22 institutions over 5 years, based on employment supervisors' assessment of new teachers' preparedness to teach, including year-long ratings
22	Behizadeh & Neely (2018)	Consequential validity of edTPA in a social justice-oriented TEP	V: consequential validity*	Candidate assessment, single TEP in a university, Georgia, US	Empirical, qualitative, primary	Reflective commentary by candidates (n = 16)
23	Bell et al. (2018)	Administrators' judgement accuracy, based on assessment of their thinking and reasoning strategies	R: consistency and accuracy*	Schoolteacher evaluation, not in the context of a TEP, US	Empirical, mixed methods, primary	Evaluation instrument for school administrators (n = 35) to rate teachers Think-aloud exercises during rating
24	Chaplin et al. (2014)	Correlation between and among teacher effectiveness measures: Research-based Inclusive System of Evaluation; 7Cs; and students' value-added achievements	R: consistency and accuracy*	Schoolteacher evaluation, not in the context of a TEP, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on school principals' and classroom students' rating of teachers (n = 329) Administrative records related to student achievement for these teachers

25	Conderman & Walker (2015)	Examining similarities between candidates' and instructors' concerns about candidate dispositions	R: consistency and accuracy*	Candidate evaluation with an authentic tool, 2+ TEPs in a university, US	Empirical, quantitative, primary	Evaluation instrument based on candidates' (n = 248) and university-based teacher educators' (n = 80) assessment of candidate disposition
26	Choi et al. (2016)	Reliability and validity of the Teacher education dispositions rating form, developed and used in a TEP	R: internal consistency reliability*	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, quantitative, primary	Evaluation instrument (mid-term and final ratings) used by university-based and school-based teacher educators to rate candidates (n = 147) Evaluation instrument to rate candidates' engagement with students
27	Johnston et al. (2018)	Advancing psychometric assessment of nine previously validated dispositional indicators (EDA tool)	V: construct validity*	Candidate assessment, 2+ TEPs across universities, US	Empirical, quantitative, primary	Interviews with stakeholders including university-based and school-based teacher educators and candidates (n = 22) Alignment of indicators assessed using Q-sort procedure with stakeholders (n = 16)
28	Lazarev et al. (2017)	Ability of Texas Teacher Evaluation and Support System (T-TESS) rubric to distinguish teaching quality; internal consistency of T-TESS	R: internal consistency reliability*	Schoolteacher evaluation, not in the context of a TEP, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on rating by qualified raters (n = 8,250 records) Administrative records held by schools (n = 51)
29	Lyness et al. (2021)	IRR of portfolios scored by Performance Assessment for California Teachers (PACT) evaluators, comparing findings across statistical methods, challenges	R: consistency and accuracy*	Candidate evaluation with an authentic tool, single TEP in a university, US	Empirical, mixed methods, primary and secondary	Evaluation instrument used by local raters (n = 2), based on assessment of portfolios (n = 19) Interviews with raters (n = 10) Pre-existing evaluation results based on double-scored portfolios as 'true scores'

30	Montecinos et al. (2010)	Consequential validity of a candidate evaluation tool, Samples of Teaching Performance (STP)	V: consequential validity*	Candidate assessment, 2+ TEPs across universities, Chile	Empirical, mixed methods, primary	Evaluation instrument used by 2 school-based educators to rate candidates (n = 24 reports) Questionnaires, with a comment space, to candidates (n = 62) and school-based teacher educators (n = 40) Focus groups with candidates (n = 47 participants) and school-based teacher educators (n = 40 participants)
31	Murley et al. (2014)	Inter-rater reliability in university course instructors' and trained project participants' use of the Teacher Work Sample (TWS) Scoring Rubric; perspectives on scoring prompts	R: consistency and accuracy*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, mixed methods, primary and secondary	Evaluation instrument and feedback form (n = 100 teacher work samples) completed by university-based and school-based teacher educators Pre-existing evaluation results of work samples as 'true scores'
32	Papanastasiou et al. (2012)	Examining the coherence between programme and state standards in a TEP	IDI: instrument development	Candidate assessment, single TEP in a university, US	Empirical, qualitative, secondary	Documents related to programmes – portfolio creation guidelines, standards, lesson plans
33	Parkes & Powell (2015)	Commentary on problems with and alternatives to edTPA	V: predictive valdity*	Candidate assessment, not in the context of a TEP, US	Non-empirical	No data collection
34	Pufpaff et al. (2015)	Rater agreement before and after digital training	R: consistency and accuracy*	Candidate assessment, single TEP in a university, US	Empirical, mixed methods, primary and secondary	Evaluation instrument based on university-based teacher educators' assessment of candidate assignments Questionnaire to university-based teacher educators (n = 10) Pre-existing evaluation results of course instructors as 'true scores'
35	Saltis et al. (2020)	Alignment of mentor teacher and candidate's rating on candidate's professional dispositions (PDQ)	R: consistency and accuracy*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on mid-term and end-of-year rating of candidates over 3 years, including candidate self-rating (n =

						1,220), ratings of mentor teachers (n = 2,094) and university-based teacher educators (n = 1,367)
36	Tait-McCutcheon & Knewstubb (2018)	Alignment between self, peer and lecturer-assessment of candidates; possible reasons of divergence	R: consistency and accuracy*	Candidate assessment, Single TEP in a university, New Zealand	Empirical, mixed methods, primary	Evaluation instrument for rating candidates (n = 34), based on self-rating and rating by candidate peer group and university-based teacher educators Interviews with candidates (n = 14)
37	Tracz et al. (2017)	Predictive relationship between selectivity standards and principal supervisors' ratings of teachers via the Systemwide Evaluation of Professional Teacher Preparation Programs	V: predictive validity*	Early career teachers in school context, 2+ TEPs across universities, US	Empirical, quantitative, secondary	Pre-existing evaluation results based on employer principals' rating of graduates (n = 11,723) Administrative records of graduated teachers – SAT (n = 289) and GPA (n = 3,420) results
38	Voss et al. (2011)	Developing and validating an instrument for assessing teachers' general pedagogical and psychological knowledge (the tool is called Pedagogical and Psychological Knowledge)	V: construct validity*	Candidate assessment, 2+ TEPs across universities (and across federal states), Germany	Empirical, mixed methods, primary and secondary	Questionnaires to expert judgers – university-based teacher educators and schoolteachers (n = 20) Evaluation instrument based on rating of candidates (n = 27) by schoolteachers (n = 71), school students (n = 620), and candidates (self-rating; n = 845) Literature review
39	Tillema (2010)	Commentary on using formative assessment for teacher professional development	IDI: instrument implementation and result use	Schoolteacher evaluation, context-free	Non-empirical	No data collection
40	Yinger & Daniel (2010)	Commentary on standards and accreditation processes in teacher education	IDI: instrument development	Candidate assessment, context-free	Non-empirical	No data collection
41	Tigelaar & van Tartwijk (2010)	Commentary on prospective teacher evaluation methods such as portfolios, self-assessment	IDI: instrument implementation and result use*	Candidate assessment, context-free	Non-empirical	No data collection

42	Rafiq et al. (2022)	Examining public and private university lecturers' evaluation proformas	IDI: instrument structure*	Higher education lecturer evaluation, not in the context of a TEP, Pakistan	Empirical, qualitative, secondary	Documentary analysis of evaluation proformas (n = 8) for higher education lecturers
43	Rafiq & Qaisar (2021)	University lecturers' views about their evaluation in a private university	V: face validity*	Higher education lecturer evaluation, not in the context of a TEP, Pakistan	Empirical, quantitative, primary	Questionnaire to higher education lecturers (n = 150)
44	Khan et al. (2017)	Review of teacher evaluation methods, student achievement-based assessment	IDI: instrument implementation and result use*	Schoolteacher evaluation, not in the context of a TEP, Pakistan	Non-empirical	No data collection
45	Rizwan & Masrur (2018)	Schoolteachers' content knowledge of a standard ('instructional planning and strategy') and their attitudes towards it	V: consequential validity	Schoolteacher evaluation, not in the context of a TEP, Pakistan	Empirical, quantitative, primary	Evaluation instrument based on schoolteachers' self-rating (n = 345)

Note. IRR: inter-rater reliability; IDI: Instrument development and implementation; R: reliability; TEP: teacher education programme; V: validity.

* More than one theme appears in the study; ** quantitative data collection via questionnaire with some qualitative data provided via an open-ended comment section.

3.3.1 Research Themes and Study Context

This section describes the research themes in the studies included in the review and the context of the studies.

3.3.1.1 Research Themes

Our review of 45 studies revealed three key areas of interest in teacher evaluation: validity; reliability; and instrument development and implementation (Table 3.7). Interestingly, most studies tackled multiple themes. In terms of the primary themes, validity emerged as the top area of focus, with over half the studies (n = 25) delving into this area. Reliability followed, with 13 studies, and judgement-making was the least common theme, receiving direct attention in 7 studies.

Table 3.7

Research Themes

Main theme	Sub-theme	Study	*n*
Validity (n = 25)	Construct validity	1, 3, 7, 11, 17, 18, 27, 38	8
	Face validity	4, 8, 10, 16, 20, 43	6
	Consequential validity	5, 13, 15, 22, 30, 45	6
	Predictive validity	6, 12, 33, 37	4
	Content validity	14	1
Reliability (n = 13)	Consistency and accuracy	9, 23, 24, 25, 31, 35, 36	7
	Influences on rater reliability and how to improve it	2, 21, 29, 34	4
	Internal consistency reliability	26, 28	2
Instrument development and implementation (n = 7)	Instrument implementation and result use	19, 39, 41, 44	4
	Instrument development	32, 40	2
	Instrument structure	42	1

Studies related to validity of judgement and tools most commonly aimed to address construct (n = 8), face (n = 6) and consequential validity (n = 6), followed by predictive (n = 4) and content (n = 1) validity. Notably, while this analysis only provides insight into the primary sub-themes addressed in each of the studies, many studies incorporated multiple aspects of validity. For instance, although frequency analysis shows that only one study was primarily focused on content validity, multiple studies in fact incorporated content validity as part of their broader examination of construct validity. In these cases, we recorded construct validity over content validity because construct validity was seen as 'subsum[ing] other types of validity' (Cohen et al., 2018, p. 256).

Studies that focused on reliability of judgement and tools most commonly aimed to identify consistency and accuracy (n = 7). This was followed by studies addressing rater reliability and how to improve it (n = 4) and studies addressing internal consistency reliability (n = 2). Importantly, none of the studies directly examined the rationale behind rater decisions. However, three studies did identify and explore rationale, with the aim of elucidating

accuracy and consistency in judgement-making. Studies 29 and 36, as part of examinations of inter-rater reliability (IRR), included interviews with raters to explore their reasoning, and Study 23 carried out statistical analysis based on independent-sample *t* tests to examine the relationship between accuracy of administrator scoring and reasoning strategies.

The remaining studies, focusing on instrument development and implementation, examined instrument implementation and use of results (*n* = 4), instrument development (*n* = 2), and instrument structure (*n* = 1).

3.3.1.2 Country Context

As shown in Table 3.8, the studies were predominately set in the US (*n* = 25), followed by Pakistan (*n* = 6), and Germany and Malaysia (both *n* = 2). Countries with one study were Canada, Chile, Indonesia, Mexico, New Zealand, and South Africa. One study involved student teachers in two countries: Germany and Austria. Notably, the aim of this study was not to examine teacher education in these countries; rather it was to examine whether pre-service teachers' ratings of an instructor showed bias. Three studies did not indicate a specific country context. No research was identified from the four UK home nations.

Table 3.8

Country Context

Country	Study	*n*
US	1, 4, 5, 11, 12, 13, 15, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 37	25
Pakistan	7, 10, 42, 43, 44, 45	6
Germany	6, 38	2
Malaysia	14, 18	2
Canada	16	1
Chile	30	1
Indonesia	8	1
Mexico	3	1
New Zealand	36	1
South Africa	9	1
Germany and Austria	2	1
N/A	30, 39, 41	3

3.3.1.3 Scope of Teacher Education Programmes

When the study aim was to examine teacher education, this took place within the context of university-based TEPs (Table 3.9). The majority of these studies involved a single TEP (*n* = 13) or multiple TEPs within a single university (*n* = 6). Some studies involved multiple TEPs across universities in the same country (*n* = 10). Only one study involved multiple TEPs in different countries.

Notably, 15 studies were not relevant to a contextual enquiry to for two reasons: they either did not collect or use data from TEPs or focused on other evaluation contexts of teaching effectiveness, such as university instructor evaluation or practising teacher evaluation.

Table 3.9

Scope of Teacher Education Programmes

Scope	Study	*n*
Single TEP within one university	1, 4, 5, 8, 9, 13, 17, 22, 26, 29, 32, 34, 36	13
Multiple TEPs within the same university	15, 20, 21, 25, 31, 35	6
Multiple TEPs across universities in the same country	6, 10, 11, 12, 14, 19, 27, 30, 37, 38	10
Multiple TEPs in different countries	2	1
TEP not mentioned	3, 7, 16, 18, 23, 24, 28, 33, 39, 40, 41, 42, 43, 44, 45	15

3.3.1.4 Methods and Participants

As shown in Table 3.10, most studies were empirical, or data driven: 40 compared to 5 non-empirical studies. Of the empirical studies, most were based on quantitative methods (*n* = 20) and mixed methods (*n* = 14), followed by studies using qualitative methods (*n* = 6). It should be noted that quantitative studies using a questionnaire that included a section for open-ended commentary were categorized as 'mixed methods' research (these are recorded with a double asterisk in Table 3.6).

Among the empirically driven studies, a notable portion (*n* = 19) relied solely on primary data collection, while others used secondary data (i.e., pre-existing data; *n* = 11) or combination of primary and secondary data (*n* = 10).

Table 3.10

Research Type, Research Methods, and Evidence Type

Research type	Study	*n*
Empirical	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 37, 38, 42, 43, 45	40
Non-empirical	33, 39, 40, 41, 44	5

Methods	Study	*n*
Quantitative	1, 6, 10, 11, 12, 15, 17, 18, 19, 20, 21, 24, 25, 26, 27, 28, 35, 37, 43, 45	20

Mixed methods	2, 3, 7, 8, 13, 14, 16, 23, 29, 30, 31, 34, 36, 38	14
Qualitative	4, 5, 9, 22, 32, 42	6
N/A (non-empirical)	33, 39, 40, 41, 44	5

Evidence type	Study	*n*
Primary	2, 3, 10, 11, 14, 16, 17, 18, 19, 20, 22, 23, 25, 26, 27, 30, 36, 43, 45	19
Secondary	1, 5, 12, 15, 21, 24, 28, 32, 35, 37, 42	11
Primary and secondary	4, 6, 7, 8, 9, 13, 29, 31, 34, 38	10
N/A (non-empirical)	33, 39, 40, 41, 44	5

As shown in Table 3.11, studies that collected primary data employed a variety of data collection techniques. The two most common were evaluation instruments (n = 19) and questionnaires (n = 14). Evaluation instruments were used to elicit judgements and ratings on candidates and teachers, and these encompassed fabricated (i.e., created for research purposes; n = 7), authentic (i.e., actively in use in TEPs, n = 7), and emerging (i.e., in the development stage, n = 5) instruments. Questionnaires were used to gather data on participants' experiences with and opinions of evaluation tools or models, and data on programme and participant characteristics. In addition, primary data collection involved obtaining direct verbal data and feedback through interviews and focus groups (both n = 5), and through a think-aloud exercise (n = 1). Written views and feedback were collected via open-ended commentary sections in questionnaires (n = 5) as well as through a feedback form and reflective commentary (n = 1).

Studies drawing on secondary data used various pre-existing data sources. The most common source was evaluation outputs, encompassing evaluation results, such as ratings assigned to a teacher candidate (n = 11) and an evaluation report (n = 1). This was followed by documents, encompassing programme documentation, such as reviews and administrative records (n = 3) and samples of coursework and fieldwork (n = 3); though six instances of use of these types of document were recorded, the overall number of studies collecting documents was five, as one of the studies gathered both programme and candidate documents. The next most common type of secondary data was administrative records, encompassing data on candidates (i.e., ethnicity, cognitive skills, SAT scores, GPA; (n = 3), data on classroom students (i.e., achievement in mathematics; (n = 2), and data on school characteristics (n = 1). A literature review was carried out in two studies to inform construction of evaluation instruments.

Table 3.11

Primary and Secondary Data Collection

Primary data collection		Study	*n*
Evaluation instrument (*n* = 19)	Fabricated evaluation instruments	2, 6, 18, 26, 36, 38, 45	7
	Authentic evaluation instrument	11, 23, 25, 26, 29, 31, 34	7
	Emerging evaluation instrument	3, 7, 17, 30, 38	5
Questionnaire (*n* = 14)		3, 6, 7, 8, 10, 14, 16, 19, 20, 27, 30, 34, 38, 43	14
Verbal data and feedback (*n* = 11)	Interview	4, 7, 27, 29, 36	5
	Focus group	7, 8, 9, 13, 30	5
	Think-aloud exercise	23	1
Written views and feedback (*n* =7)	Space for open-ended comments (as part of a questionnaire)	2, 3, 14, 16, 30	5
	Reflective commentary	22	1
	Feedback form	31	1

Secondary data collection		Study	*n*
Evaluation outputs (*n* = 12)	Evaluation results	1, 12, 13, 15, 21, 24, 28, 29, 31, 35, 37	11
	Evaluation report	9	1
Documents (*n* = 6)	Programme document	4, 32, 42	3
	Coursework and fieldwork sample	4, 5, 8	3
Administrative records (*n* = 6)	Candidates' ethnicity, cognitive skills, SAT scores, GPA	6, 12, 37	3
	Classroom students' achievement	12, 24	2
	School characteristics	28	1
Literature (*n* = 2)	Information on item development	7, 38	2

Table 3.12 shows the distribution of research participants in the empirical studies (n = 40). Our examination revealed that in almost half of the empirical studies (n = 17), participants were university-based teacher educators. This was followed by teacher candidates (n = 13) and school-based teacher educators (n = 10). Others included classroom teachers (n = 5), classroom students, employment supervisors, and qualified raters (all n = 3), higher education lecturers and school principals (both n = 2), candidate peers, higher education students, and TEP coordinators (all n = 1). Data collected from these participants typically took the form of participant views, experiences, assessment of candidates and teachers, and assessment of teacher evaluation tools.

Table 3.12

Research Participants

Participants	Study	*n*
University-based teacher educators	1, 4, 8, 9, 10, 11, 13, 14, 15, 25, 26, 27, 31, 34, 35, 36, 38	17
Candidates	1, 2, 10, 13, 15, 20, 22, 25, 27, 30, 35, 36, 38	13
School-based teacher educators	1, 8, 9, 15, 17, 26, 27, 30, 31, 35	10
Classroom teachers	3, 6, 18, 38, 45	5
Classroom students	6, 24, 38	3
Employment supervisors	21, 23, 37	3
Qualified raters	12, 28, 29	3
Higher education lecturers	7, 43	2
School principals	16, 24	2
Candidate peers	36	1
Higher education students	7	1
Teacher education programme coordinators	19	1

3.3.1.5 Evaluation Context

Among the 45 studies in our review, five types of evaluation context emerged: candidate assessment in a university-based teacher education context (n = 27); evaluation of early career teachers in a school context (n = 4); evaluation of practising teachers in a school context (n = 10); evaluation of teacher educators in universities (n = 1); and evaluation of lecturers in universities (n = 3; Table 3.13). This reflects a wider scope than might be expected based on the search strings we used, which focused on ‘candidates’. We report this finding for transparency, to ensure that readers have a clear understanding of the scope and

limitations of our review and to allow for interpretation of the findings within the appropriate context. We include other contexts to provide a broader picture of judgement-making in teacher effectiveness in some domains (e.g., the rare focus on factors influencing judgement in candidate evaluation settings). Our purpose in reporting evaluation context is not to generalize.

In studies that examined evaluation of candidates during teacher education (n = 27), the context was always a university-based TEP; no other context was found. In other words, when the study was about teacher education, this took place within the context of university-based TEPs. This is a notable finding, indicating the need for further investigation of the work of alternate education provisions.

Evaluation of early career teachers in a school context (n = 4) reflects an intention to follow up or predict candidates' success as practising schoolteachers (i.e., predicative validity). Three studies compared data on practising schoolteachers – both new and experienced teachers – with data for the same individuals from before or during teacher education (Studies 6, 12, 37). Specifically, Study 6 collected ratings from classroom students on their schoolteachers' instructional quality and compared this with results on the teachers' cognitive abilities, personality data, and course grades received during teacher education. Study 12 used edTPA (Educative Teacher Performance Assessment) portfolio ratings to explore if these predict employment status and student attainment in mathematics and reading. Study 37 examined the relationship between teacher education selectivity standards (i.e., based on SAT and GPA results) and principal supervisors' rating of teachers. The other study (Study 21) involved evaluation of teacher educators in university-based teacher education provisions by examining ratings assigned by employer supervisors to investigate whether they showed any bias based on new teachers' characteristics (i.e., socioeconomic status and ethnicity). One study looked at evaluation of teacher educators in universities.

Among the studies that examined evaluation of practising teachers in a school context (n = 10), two (Studies 3, 16) included evaluation of both new and experienced teachers. The studies on evaluation of higher education lecturers (n = 3) did not necessarily look at practices related to teacher education.

Table 3.13

Evaluation Context

Evaluation context	Study	*n*
Evaluation of candidates in university-based teacher education provisions	1, 4, 5, 8, 9, 10, 11, 13, 15, 17, 19, 20, 22, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 38, 40, 41	27
Evaluation of practising teachers in school context	2, 3,* 16,* 18, 23, 24, 28, 39, 44, 45	10
Candidate assessment in alternate teacher education provisions	None	0
Evaluation of early career teachers in school context	6, 12, 21, 37	4
Evaluation of higher education lecturers	7, 42, 43	3

Evaluation of teacher educators in university-based teacher education provisions	14	1
Evaluation of candidates in alternate teacher education provisions	NA	0

Note. * Study includes both new and experienced teachers.

3.3.2 Approaches to Teacher Education and Development and Use of Evaluation Instruments

In this section, the distinct approaches to teacher education in the studies are described, and the evaluation instruments in the studies are examined to reveal, where available, information on development, design and implementation of instruments.

3.3.2.1 Teacher Education

Our review of the studies revealed that research in the field has been conducted exclusively within university-based TEPs, with no studies identified in alternative teacher education provisions. University programmes often have a high degree of autonomy over their curriculum design, including decisions on programme content, duration, and structure (Study 30). However, this autonomy does not always extend to assessment practices. In some cases, TEPs are mandated by external authorities to adopt assessment tools (e.g., in Study 4, the edTPA is mandated) or independent researchers have developed the tools used to meet externally imposed national or state standards (e.g., see Samples of Teaching Performance [STP] in Study 30). In practice, these have caused challenges for TEPs, which have to adjust their ways of working and curriculums rather than enjoying autonomy.

Examination of the studies also revealed that in general TEPs begin with foundational coursework in education theory and pedagogy, gradually incorporating field experiences. However, various studies underscored the gap in successfully translating theoretical knowledge into practical application (Studies 5, 6, 17). To deal with this issue, several TEPs that distinguish themselves from traditional university-based programmes have emerged and are worthy of note:

1. **Extended Internship Program (Study 17):** This programme integrates theory with classroom practice through year-long internships in partnered district schools. Interns start when the school year begins and fully participate in all school activities, meetings, and parent–teacher conferences, receiving collaborative support from district and university faculty.

2. **Teaching, Learning, and Leading with Schools and Communities Program (Study 13):** This field-based learning model (also referred to as guided field-based apprenticeship) was designed as a collaborative effort between a university, community, and schools. Student teachers engage in supervised field experiences from their freshman year through eight semesters. Coursework and clinicals have been nearly completely replaced with supervised field-based student teacher learning experiences. University coursework is designed to complement field learning rather than being separate from it. Faculty members accompany candidates daily and

provide direct supervision and feedback grounded in observational evidence, using an evaluation tool called the Classroom Assessment Scoring System (CLASS).

Study 18 focused on one out of eight semesters of the programme, in which students in their second year are required to participate in a birth-to-three experience in Early Head Start classrooms (infant/toddler settings). At the same time, students enrol in two courses (on early childhood special education and language/literacy development), designed to support the field component. The birth-to-three experience, lasting 12 weeks, is divided into three modules, and classroom teachers and early childhood faculty provide constant direct supervision and field-based instruction. Module 1 lasts 3 weeks and involves university-based seminars and community-based experiences in the form of visits to diverse infant/toddler and preschool programmes. The aim is for candidates to learn about child development and the role of teachers in providing developmentally appropriate learning experiences. Module 2 lasts 6 weeks (this is the period during which data was collected in Study 18). Pairs of candidates are placed in classrooms three mornings per week for 6 weeks, with a seminar at the beginning and end of each day. Candidates develop individual and collaborative activity plans under the direct supervision of classroom teachers. Classroom teachers provide immediate and consistent feedback (feedback type 1). Faculty provide daily supervision and individual feedback (during and after classroom visits). Faculty also rate all candidates weekly using CLASS and provide formal narrative and quantitative feedback over the 6 weeks of the module (feedback type 2). CLASS scores are also used to determine candidates' progress and development, and that informs the level of tiered support (i.e., universal, targeted, intensified support) the candidate receives from faculty.

3. **Early Field Experience (Study 19):** This programme, set in agricultural TPPs, serves two purposes: it provides opportunities for aspiring teachers to explore teaching and it assists TPP students to transition to teaching positions. Whether or not Early Field Experience is compulsory depends on the specific TPP and the requirements set out by the educational institution. It could run prior to student teaching; also the experiences could be offered within or outside of the agricultural education curriculum. The purpose of the programme is the application of pre-service teacher knowledge and skills in various settings, which could include teaching lessons, tutoring students, or observing in the classroom. Programmes require a minimum number of contact hours and a minimum number of lessons taught while in the field. The programme often involves interaction with peers, a cooperating teacher, and a university supervisor.

4. **Teacher education in Germany (Study 38):** In Germany, there are two phases of teacher education. The first phase, lasting 4–5 years, takes place in university. Here, candidate teachers attend general courses, in subjects like psychology, pedagogy, and sociology, as well as two subject-based courses. The second phase, called the Referendariat, involves practical training. Candidates are placed in schools and begin by observing experienced teachers (learning by observation). After roughly 6 months, they take on teaching responsibility, moving gradually from guided to independent

teaching, for around 10 hours per week. While they gain hands-on experience, trainees also continue their theoretical learning for 6–8 hours per week; this covers both general teaching methods and subject-specific strategies for their chosen areas. Additionally, the Referendariat provides a valuable support system. Candidates are supported through mentoring, peer interaction, and both instructional and psychological guidance. The organization of the Referendariat (e.g., the amount of teaching experience, the length of the observation phase) varies to some extent across the federal states.

3.3.2.2 Identified Evaluation Instruments

In the 45 studies included in the review, a total of 37 evaluation instruments were identified. These were categorized into three types:

- **authentic (*n* = 28):** actively used and/or integral to assessing teaching effectiveness in real-world practice. Of the authentic instruments identified, 17 were exclusively for evaluation of candidates, 7 were for evaluation of in-service teachers, 3 were for evaluation of higher education instructors, and 1 was for evaluation of teacher educators.
- **emerging (*n* = 5):** still in the developmental stage and yet to be used in real-world practice. Among the instruments in this category, the groups to be evaluated were candidates, new teachers, and in-service teachers.
- **fabricated (*n* = 7):** developed specifically for use in research and not intended for real-world application. These tools were not examined in depth in this study, as they do not reflect real-world practice.

The emerging evaluation tools were:

- SOCME-10 (Study 3), validated for use in evaluation of new and experienced practising teachers;
- Item-Level Assessment of Teaching Practice (I-LAST; Study 17), validated for use in candidate teacher evaluation;
- EDA (Study 27), validated for use in candidate teacher evaluation; and
- Pedagogical Psychological Knowledge (PPK; Study 38), validated for use in candidate teacher evaluation

Authentic in-service teacher evaluation tools included:

- Teacher Performance Appraisal (Study 16), used in Canada for assessment of in-service teachers (new and experienced) by school principals – new teachers were assessed on 8 competencies and experienced teachers were assessed on 16 competencies;
- Systemwide Evaluation of Professional Teacher Preparation Programs (Studies 21, 37), used in California, US, to follow up on candidates in their first year of employment, based on assessment by principals (‘employment supervisors’);
- Teaching and Learning Framework (Study 23), used in Los Angeles, US, for assessment of in-service teachers by school administrators;
- Research-based Inclusive System of Evaluation (assessment of in-service teachers by their principals), 7Cs (assessment of in-service teachers by their students), and value-added achievements (Study 24) were used in combination in Pittsburgh, US; and

- Texas Teacher Evaluation and Support System (T-TESS; Study 28), used in Texas, US, for assessment of in-service teachers by qualified raters

We aimed to examine the authentic and emerging instruments (n = 32) in relation to three strands: development information (i.e., developer, grounding); design information (i.e., assessment dimensions, rating scales); and implementation information (i.e., raters, evaluation sites, evaluation approaches, use of findings). However, for the emerging instruments (n = 4) and the in-service teacher evaluation instruments (n = 7), we could only examine design information. There were two reasons for this: First, in relation to the in-service teacher evaluation tools, we wanted to ensure fair comparison between different types of tool, and as our search was focused on 'candidate'-related terms, the identified in-service teacher evaluation tools were not representative of tools in this category. Second, there were obvious barriers to direct comparison of development information and implementation information for candidate evaluation tools with tools designed for in-service teachers, higher education instructors, and teacher educator evaluation (i.e., there are more officially designed tools for in-service teachers, and comparing this with candidates would skew our findings). However, examining design information for these tools was considered sensible in order to glean insights into new developments and whether the criteria used to evaluate candidates were relevant to the aims of in-service teacher evaluation.

For the *authentic* evaluation instruments that were for evaluation of candidates exclusively (n = 17), we carried out analysis in relation to all three strands: development; design; and implementation information. However, six of these instruments could not be included due to insufficient information (Studies 8, 9, 10, 19, 20, 34). The findings for the other 11 instruments are outlined in Table 3.14.

Table 3.14

Overview of Authentic Evaluation Instruments for Assessment of Candidates

Tool and study	Development	Design	Implementation	Psychometric properties
Competence Assessment* Study 1	**Context:** developed and used by a TEP **Developer:** single university-based teacher education provider ('a committee of faculty experts'), Virginia, US, 2001–2002; has undergone various improvements since its creation **Grounding:** • professional association standards – InTASC, National Council for the Accreditation of Teacher Education, Council for the Accreditation of Educator Preparation, National Council of Teachers of English, National Council of Teachers of Mathematics, National Council of Teachers of Science, Council for Exceptional Children • prior academic research – qualities of effective teachers • in-house data – views and assessments of faculty experts and school-based mentors • theoretical/conceptual framework	**Focus:** evaluation of teaching effectiveness, based on competencies (knowledge, skills, dispositions) **Rating system:** Each performance indicator (item) is rated individually, and an overall judgement is assigned, but it was not made clear how. **Rating scale:** 4-point scale – below expectations (1), developing (2), meets expectations (3), exceeds expectations (4) **Structure:** 30 competencies ('performance indicators') under four domains – planning, onstage teaching, assessment, professionalism	**Summative decision:** results are used to determine pass/fail for the competency of teaching during teaching experience **One-time formative feedback (for all) to guide candidate improvement:** Results are used to identify areas for growth, the intention being to guide learning and improvement and facilitate coaching discussions. There is no specific support other than this. **Evaluation site and frequency:** two time points – middle and end of field experience **Method:** observation and self-evaluation **Evaluation approach:** formative and summative **Raters:** candidates, mentor teachers, university supervisors **Rating process and QA:** A trio provides ratings independently, the intention being to provide a judgement about student teachers' performance based on multiple sources of information, including observation and coaching conversations. No information was provided on how agreement is reached or what happens in case of no agreement. **Training for raters:** training is provided for candidates only	Face validity has been established due to 15 years of use and various refinements. Study 1 concluded that the tool provides partial, reasonably valid, and reliable evidence of student teachers' competencies.

Educative Teacher Performance Assessment (edTPA) Studies 4, 5, 12, 22, 33	**Context:** adopted and used by various TEPs **Developer:** education research centre – Stanford Centre for Assessment, Learning, and Equity (SCALE) in collaboration with American Association of Colleges for Teacher Education, US, 2011–2012; underwent various field tests and was officially launched in 2013–2014 **Grounding:** • pre-existing tool – multiple candidate evaluation instruments: InTASC standard portfolio, Performance Assessment for California Teachers (PACT) National Board Portfolio; shares similarities with in-service teacher evaluation tool: National Board for Professional Teaching Standards assessment • professional association standards – InTASC, National Council for Teachers of Mathematics, InTASC Model Core Teaching Standards and Learning Progressions	**Focus:** evaluation of teaching effectiveness, based on performance **Rating system:** Each rubric is rated and summative scores ('holistic' scores) are provided, ranging from 15 to 75 assuming no missing rubric scores. **Rating scale:** 5-point scale **Structure:** The tool encompasses three tasks – planning, instruction, assessment – each comprising five rubrics, totalling 15 rubrics overall. Depending on state requirements, additional tasks and rubrics may be incorporated – e.g., in Washington state, teacher candidates are evaluated on three additional student voice rubrics; however, these additional rubrics do not contribute to the candidates' summative scores (Study 12).	**Summative decision:** Results are used to evaluate the candidates' teaching readiness for certification. No written or verbal feedback is provided for the decision, and no formative feedback is offered to candidates, including failed candidates **Programme improvement:** results are used for programme-level improvement and accreditation purposes **Evaluation site and frequency:** once, on submission of a portfolio in the final stage of the TPP **Method:** portfolio assessment **Evaluation approach:** summative **Raters:** single outsourced evaluator **Rating process and QA:** recommends an additional rater if the score is 'at or near', but no data was shared by SCALE to show whether this was done **Training for raters:** yes	It has been argued that due to years of use, the tool has established reliability and validity (Study 4). However, some other studies (e.g., Study 33) have criticized SCALE's validation study and the claim about content validity. Study 29 criticized the lack of availability of data on double scoring, which would allow researchers to test the IRR of edTPA scoring; the authors called for SCALE to make this data available (Study 29).
Intern Keys Teacher Candidate Assessment Study 11	**Context:** adopted and used by various TEPs **Developer:** state education department, Georgia, US	**Focus:** evaluation of teaching effectiveness **Rating system:** Each performance indicator (item) is rated individually.	**Result use:** no information **Evaluation site and frequency:** field experience; no information on frequency **Method:** observation	An initial validation study conducted by Georgia State Department, US, was criticized (Study 11).

	Grounding: • pre-existing tool – an in-service teacher assessment tool, Teacher Assessment on Performance Standards, which has strong similarities with the widely used Stronge Teacher Effectiveness Performance Evaluation System	No information was provided on how a holistic judgement is assigned. **Rating scale:** 4-point scale – from ineffective (1) to exemplary (4) **Structure:** single factor structure, based on 'teaching effectiveness'; 72 criteria under 10 items (standards) – instructional planning, assessment uses, positive learning environment, assessment strategies, academically challenging atmosphere, communication, professional knowledge, professionalism, instructional strategies, differentiated instruction	**Evaluation approach:** no information **Raters:** no information **Rating process and QA:** no information **Training for raters:** no information	Study 11 showed robust construct validity, exhibiting a singular factor structure in line with theoretical assumptions. It also found high internal reliability (Cronbach's alpha = .944). The tool was recommended for use in pre-service and early in-service teacher training programmes, as it is effective in assessing instructional performance of teacher candidates.
Classroom Assessment Scoring System (CLASS): Toddler version Study 13	**Context:** adopted and used by various TEPs **Developer:** independent researchers (US **Grounding:** • pre-existing tool – a candidate assessment tool: CLASS • professional association literature – recommended practices of professional organizations regarding adult–child interaction and support for social and emotional competence	**Focus:** evaluation of teaching effectiveness **Rating system:** Each dimension is rated individually. Information was not provided on whether, or how, a holistic judgement is assigned. **Rating scale:** 7-point scale – from low (1) to high (7) **Structure:** organized around eight dimensions – positive climate, negative climate, teacher sensitivity, regard for child perspectives, behaviour guidance, quality of feedback, facilitation of learning/development, language modelling/support	**Progress-oriented formative feedback and monitoring:** Weekly ratings and narrative feedback – based on observations by university supervisors, using monitoring graphs, and by peer groups, in recorded video clips – are used to monitor progress and tailor support. The identification of support needs (Levels 1–3) is based on feedback provided primarily by university supervisors. Additionally, daily narrative feedback collected from university supervisors and school mentors is used to inform candidates about their engagement with students. **Evaluation site and frequency:** six time points (weekly) during field experience **Method:** observation and peer feedback	No information was provided for face, content and construct validity and internal reliability. In study 13, university supervisors, before starting the evaluation process, secured a high level of IRR (.90 and above) across all dimensions of CLASS.

Profile for Evaluation of Intern (PEI) Study 15	**Context:** developed and used by a TEP **Developer:** single university-based teacher education provider ('teacher educators'), Mid-Atlantic, US **Grounding:** • state and/or national standards • professional association standards – Association for Childhood Education International standards • prior academic research – literature search for item development • in-house data – feedback on the tool items, domains, and scoring from faculty across the	**Focus:** evaluation of teaching effectiveness – measures teaching performance of intern teachers **Rating system:** Each domain is rated individually. An overall judgement is assigned, but it was not made clear how. **Rating scale:** 5-point scale – from performance needs significant improvement (1) to performance is of notable excellence (5) **Structure:** 30 criteria under four domains – preparation and planning, instruction and classroom management, assessment, professional development	**Evaluation approach:** progress-oriented evaluation (monitoring to support progress) **Raters:** university supervisors, candidate peer groups **Rating process and QA:** Weekly ratings are provided independently. University teacher educators provide holistic and targeted feedback during each observation and progress-monitoring graphs throughout the module with narratives indicating strengths and suggestions for further development. Candidates do not self-rate but at the end of field experience, they reflect on their experiences of teaching and their mentors' assessments are gathered via interviews. **Training for raters:** no information **Summative decision:** Results are used to assess candidate teaching readiness for licensure. A minimum rating of 3 out of 5 in both placements is required to obtain licensure. **One-time formative feedback (conditional) to guide candidate improvement:** candidates receive feedback and support if any ratings are 1 or 2 **Programme improvement:** Mid-term evaluation results are used to identify programme weaknesses – i.e., mean scores were low, so an action research component was incorporated by the second placement to address the issue; this led to overall improvement in mean scores in the final evaluation.	PEI is still under development, but face validity and content validity have been established through grounding of the tool, based on standards, research evidence, and stakeholder feedback. However, quantitative ecological validity and reliability tests have not been carried out . Study 15 IRR, with 80% agreement.

	programme and key school-based stakeholders (e.g., principals, teacher leaders)		**Evaluation site and frequency:** two time points – the middle of the first field experience and the end of the second one **Method:** observation and self-evaluation **Evaluation approach:** formative and summative **Raters:** candidates, mentor teachers, university supervisors **Rating process and QA:** A trio provides ratings independently and comes together to triangulate the individual and final scores for each of the teacher candidates. When trio scores are exact, this is the final score. When trio scores are exact-adjacent or adjacent, the final score is aggregated from individual scores. When scores are at least not adjacent, additional documentation is reviewed by the trio and discussion continues until agreement is reached. **Training for raters:** training for mentor teachers only; no information provided on training content	
Disposition Assessment* Study 25	**Context:** developed and used by a TEP **Developer:** single university-based teacher education provider ('faculty members'), Midwestern US **Grounding:** • in-house data – internal department meetings, focus groups, and subcommittees • institutional conceptual framework	**Focus:** assessment of disposition **Rating system:** Each indicator (item) is rated individually. No holistic judgement is made. **Rating scale:** dichotomous scale – faculty version: 'acceptable' or 'alert/unacceptable'; candidate version: 'possess' or 'not possess' **Structure:** Two versions exist, one for use by candidates and one for use by teacher educators, though these are similar. There are 17 sub-indicators in	**Progress-oriented formative feedback and monitoring:** Results are used to identify candidates' disposition status (levels 1 to 4) throughout each semester during their course so that necessary measures to support their appropriate development ('dispositional growth') can be taken – i.e., asking candidates to prepare a development plan and, if the problem is persistent, referring them to the designated committee for further investigation and creation of a remedy plan. The candidate's disposition status is based on: the total	Internal reliability of the self-assessment data was found to be high (Kuder-Richardson 20 = .88). Chi-square tests were statistically significant for each sub-area of the self-assessment ($p < .05$). No information was provided for face, content, and construct

	the candidate version and 23 sub-indicators in the faculty version, under 6 disposition indicators – caring, lifelong learning, scholarship, creative and critical thinking, collaboration, diversity (it was claimed that the lifelong learning and scholarship indicators make this tool unique)	weaknesses identified by the same or different raters, the same weakness being identified by different raters, whether the weakness is persistent over semesters, whether the development plan prepared by the candidate is satisfactory, and the number of times a Teacher Education Deficiency Report is filed each semester. Notably, candidate self-assessment is used to self-identify two disposition goals to work towards. **Evaluation site and frequency:** multiple times ('each semester') for each course throughout the programme **Method:** observation and self-evaluation **Evaluation approach:** formative assessments **Raters:** candidates and university supervisors **Rating process and QA:** Each instructor rates candidates independently. For candidates who are referred to the committee for further investigation (Level 4), the committee examines whether the identified issue is localized to a specific course or extends across multiple courses. As a result of the committee's investigation, candidate self-ratings are used as reference point in discussions and a focused remediation plan is set out, tailored to the candidate's needs. **Training for raters:** training on the instrument for candidates only – compulsory, to take place two times, before or during the first week of each semester of	validity of the instrument.

			first clinical, and it is led by candidates' peers	
Teacher education disposition rating form Study 26	**Context:** developed and used by a TEP **Developer:** single university-based teacher education provider, US **Grounding:** • professional association standards – professional dispositions • prior academic research – literature on teacher dispositions • in-house data – programme faculty input	**Focus:** assessment of disposition **Rating system:** Each item is rated individually. No information was provided on whether, or how, holistic judgement is assigned. **Rating scale:** 3-point scale – below expectation (1), meets expectation (2), exceeds expectation (3) **Structure:** 19 items under six main dispositions – responsibility, respect, integrity, caring/humanity, fairness, belief that all students can learn	**Result use:** no information **Evaluation site and frequency:** no information **Method:** observation **Evaluation approach:** no information **Raters:** no information **Rating process and QA:** no information **Training for raters:** no training	Findings show that, by and large, the tool exhibited a singular factor structure in line with theoretical assumptions. A high internal consistency was found for the 19 items. Test–retest reliability and IRR estimates showed consistent ratings for only responsibility and the composite score over time. Despite statistically significant correlations, all IRR estimates remained small, indicating poor correspondence of ratings by field-based supervisors and university supervisors.
Performance Assessment for California Teachers (PACT) Study 29	**Context:** adopted and used by various TEPs **Developer:** education research centre – Stanford Centre for Assessment, Learning, and Equity, US **Grounding:** no information	**Focus:** evaluation of teaching effectiveness, based on readiness to teach **Rating system:** Each rubric (under five tasks) is scored individually. An overall judgement is assigned, but it was not made clear how. **Rating scale:** 4-point scale – fail (1), basic or pass (2), proficient (3), advanced (4)	**Summative decision:** results are used to evaluate candidate teaching readiness and inform the decision of pass/fail for a teaching credential **Formative feedback to inform summative evaluation (conditional):** feedback provided only if a portfolio fails evaluation by both the original and second evaluators; in such instances, portfolios undergo remediation with review and	Study 29 found poor IRR among qualified raters. No information was provided on face, content, and construct validity and internal reliability of the instrument.

Structure: 12 rubrics (items) under five tasks – planning, instruction, assessment, reflecting, (encouraging student use of) academic language

feedback from PACT coordinators or faculty

Evaluation site and frequency: once, on submission of a portfolio (including evidence of artifacts and commentaries)

Method: portfolio assessment

Evaluation approach: summative

Raters: internal evaluators – part-time faculty, school administrators, and classroom teachers

Rating process and QA: Each portfolio is rated initially by a single scorer. Failed portfolios undergo a second evaluation by a different evaluator; agreement of both evaluators on failure results in portfolio remediation with support and feedback from PACT coordinators or faculty. Disagreement leads to the involvement of a third evaluator, often a lead faculty member, to break the deadlock. Additionally, 15% of passed portfolios are randomly double-scored to adhere to California state guidelines.

Training for raters: All evaluators undergo 2-day training on the evaluation instrument, including demonstrations, practice exercises, and feedback sessions on evaluation strategies and assessment sequencing. Before assuming their roles, evaluators are required to achieve IRR scores based on example assessment scores, demonstrating perfect agreement in 6 out of the 12 rubrics with no discrepancies exceeding one rubric level. Furthermore, evaluators are mandated to attend annual

			recalibration events after initial training to refresh their understanding and maintain consistency.	
Samples of Teaching Performance (STP) Study 30	**Context:** developed and used by various TEPs **Developer:** consortia of university-based teacher education providers ('17 universities'), Chile; independent researchers (Montecinos et al., 2005), Chile **Grounding:** • state and/or national standards – Performance Standards for the Initial Preparation of Teachers • theoretical/conceptual framework –model for developing high-quality assessments for the teaching profession; Teacher Work Sample (TWS) methodology as a protocol to collect evidence from candidates	**Focus:** evaluation of teaching effectiveness, based on capacity to plan, deliver, and evaluate a unit of instruction **Rating system:** a rubric with detailed narrative description **Rating scale:** 3-point scale – unsatisfactory (1 or 1.5), basic (2 or 2.5), competent (3.0) **Structure:** six standards – contextualizing teaching, setting learning goals, lesson plans, assessment plan, decision-making and analysis of results, reflective self-assessment	**Result use:** no information **Programme accreditation:** results are used for programme improvement and accreditation purposes **Evaluation site and frequency:** on submission of a sample report **Method:** report assessment **Evaluation approach:** no information **Raters:** no information **Rating process and QA:** no information **Training for raters:** no information	Content validity of the instrument was established previously IRR was calculated as rWG values for each of the six standards – values ranged between 0.64 and 0.77
Teacher Work Sample (TWS) Scoring Rubric: revised version Study 31	**Context:** modified and used by TEPs **Developer:** single university-based teacher education provider – the original TWS was developed by the Renaissance Partnership involving 11 US universities; this was modified by the university in 2010, Kentucky, US **Grounding:** • pre-existing tool – built on a candidate assessment tool,	**Focus:** teaching effectiveness, based on candidates' preparation and performance in teaching and their capacity/ability to positively impact student learning **Rating system:** Each rubric indicator is rated and a holistic judgement is assigned. **Rating scale:** 4-point scale – beginning (1), developing (2), proficient (3), exemplary (4)	**Summative decision:** Results lead to a pass or fail for the seminar course in the TEP. **Evaluation site and frequency:** once, on submission of a work sample in the seminar course taken in the last semester of the TPP **Method:** work sample assessment (i.e., whether this demonstrates evidence of candidates' capacity to positively impact student learning) **Evaluation approach:** summative	No information was provided on face, content, and construct validity and internal reliability, but the intention behind revising the TWS was to increase reliability by making it more relevant to programme needs.

	TWS, which was grounded in both state and national teaching standards • state and/or national standards – the modified version specifically adhered to state standards • in-house data – feedback from candidates, faculty members, and school mentors	**Structure:** 24 rubric indicators under five key teaching and learning processes – contextual factors, learning goal pre/post assessment, design for instruction, analysis of student learning, reflection of teaching	**Raters:** single course instructor of Student Teaching Seminar course **Rating process and QA:** no information **Training for raters:** yes, training on the instrument, including example evaluation practices	
Professional Development Qualities (PDQ) Study 35	**Context:** developed and used by a TEP **Developer:** single university-based teacher education provider, 2015, Western US **Grounding:** • professional association standards • prior academic research – work on disposition item development • in-house data – stakeholder need analysis survey with special education teachers and university professors	**Focus:** dispositional evaluation **Rating system:** Each item is scored and a (holistic) mean score is calculated. A brief definition of the item and rating is included to help with reliability of ratings. **Rating scale:** 3-point scale – unacceptable (1), approaching (2), target (3) **Structure:** 12 disposition items – professional appearance, attendance, professional responsibility, ethical behaviour, response to feedback, reflexive practice, collaboration, professional initiative, respect for diversity, student engagement, communication skills, portrays competence and confidence	**Progress-oriented formative feedback and monitoring:** Results are used to track candidates' professional disposition over time, the intention being to encourage discussion among candidates and those who support them in school settings. This includes interventions (Study 26). **Evaluation site and frequency:** three times – at the beginning of the programme and during first ('middle') and second ('final') field experience **Method:** observation and self-evaluation **Evaluation approach:** progress-oriented evaluation **Raters:** candidates, mentor teachers, and university supervisors **Rating process and QA:** A trio provides ratings independently. There is no intention to align ratings, as the evaluation is learning focused. **Training for raters:** no information	Tool reliability and validity was confirmed (Brewer et al., n.d.; Brewer et al., 2011). Study 35 found a high level of internal consistency of scores (Cronbach's alpha = .85).

Note. * A name assigned by the RA; this was done where tools did not have a specific name.

InTASC: Interstate Teacher Assessment and Support Consortium; IRR: inter-rater reliability; TEP: teacher education programme; TPP: teacher preparation programme.

3.3.2.3 Development of Evaluation Instruments

This section focuses on the development of the 11 authentic tools (i.e., those used in real teacher education settings) for judging candidates' teaching effectiveness, including the groundings for each tool. The tools are as follows, with an asterisk indicating a title assigned by the RA:

- Competence Assessment* (Study 1)
- edTPA (Studies 4, 5, 12, 22, 33)
- Intern Keys Teacher Candidate Assessment (Study 11)
- CLASS: Toddler version (Study 13)
- Profile for Evaluation of Intern (PEI; Study 15)
- Disposition Assessment* (Study 25)
- Teacher education dispositions rating form (Study 26)
- Performance Assessment for California Teachers (PACT; Study 29)
- Samples of Teaching Performance (STP; Study 30)
- Teacher Work Sample (TWS) Scoring Rubric: revised version (Study 31)
- Professional Development Qualities (PDQ; Study 35)

Our analysis revealed that all of the instruments but one (STP, created in Chile) were created in the US (Table 3.15). Among these, PACT stands as an early example in the field, feeding into the development of edTPA in later years (Study 29). edTPA was the first nationally available candidate evaluation instrument in the US. Following its official launch in 2013–2014 after several field tests, it was adopted by over 600 TPPs in 40 states within 3 years (Study 12). A recent study stated that it is the most widely used tool (Study 4). Notably, among the tools examined here, edTPA and PACT are similar in essence – they have similar content and design, and both were developed by the Stanford Centre for Assessment, Learning, and Equity; however, they differ in terms of use of outsourced (edTPA) or local (PACT) evaluators (Study 29). Not all states in the US use edTPA as an assessment tool. Indeed, some instruments, such as PEI, were developed by single university-based teacher education providers with the intention of addressing the weaknesses of edTPA, such as reliance on outsourced raters and adaptability issues (Study 15).

As shown in Table 3.15, the tools were developed initially by various institutions and researchers. Five were developed by university-based teacher education providers (Competence Assessment,* Disposition Assessment*, PDQ, PEI, Teacher education dispositions rating form). One instrument (TWS Scoring Rubric: revised version) was initially developed by a consortium of universities and later modified by a teacher education provider to meet their programme needs, such as clarity, reliability, and alignment to state standards rather than national standards. Two tools were developed by education research centres (edTPA, PACT), two by independent researchers (CLASS: Toddler version, STP), and one by a state education department (Intern Keys Teacher Candidate Assessment).

Notably, although researchers (i.e., in Studies 2, 33) have criticized the high rate of use of adopted (and mandated) tools and policymakers have had low trust in university-based teacher education providers and did not give providers the freedom to create their own instruments, in our examination the majority of the examined tools (n = 6) were originally developed by teacher education providers themselves. Interestingly, a Pakistani study (Study 42) found that public universities used national tools, while private ones created their own

teacher evaluation tools. While in public sector universities in Pakistan, teacher evaluation tends to be more of a formality, teacher evaluation in private sector universities feeds into decisions related to salary, promotion, and even demotion (Study 43).

Table 3.15

Background Information on the Development of Tools

Country where tool was created	Tool	*n*
US	Competence Assessment* (Study 1), PEI (Study 15), Disposition Assessment* (Study 25), Teacher education dispositions rating form (Study 26), PDQ (Study 35), edTPA (Studies 4, 5, 12, 22, 33), Intern Keys Teacher Candidate Assessment (Study 11), CLASS: Toddler version (Study 13), TWS Scoring Rubric: revised version (Study 31), PACT (Study 29)	10
Chile	STP (Study 30)	1

Origin	Tool	*n*
Instrument developed by a university-based TEP	Competence Assessment* (Study 1), PEI (Study 15), Disposition Assessment* (Study 25), Teacher education dispositions rating form (Study 26), PDQ (Study 35)	5
Existing instrument adopted by a TEP	Originally developed by independent researchers: CLASS: Toddler version (Study 13), STP (Study 30) Originally developed by an education research centre: edTPA (Studies 4, 5, 12, 22, 33), PACT (Study 29) Originally developed by a state education department: Intern Keys Teacher Candidate Assessment (Study 11)	5
Existing instrument modified by a TEP	TWS Scoring Rubric: revised version (Study 31)	1

Note. * A name assigned by the RA; this was done where tools did not have a specific name.

TEP: teacher education programme

CLASS: Classroom Assessment Scoring System; edTPA: Educative Teacher Performance Assessment; PACT: Performance Assessment for California Teachers; PDQ: Professional Development Qualities; PEI: Profile for Evaluation of Intern; STP: Samples of Teaching Performance; TWS: Teacher Work Sample.

In our examination of the foundations of the tools, we found that this was not clear for the PACT tool, but for the other 10 tools, we identified four main sources of information: evidence (n = 8); standards (n = 7); pre-existing evaluation tools (n = 4); and institutional conceptual framework (n = 1). Some tools drew on a combination of sources, as outlined in Table 3.16.

Table 3.16

Sources of Information Used in Tool Development

Number of sources used	Type of source	Tool
Three	Evidence, standards, and pre-existing instrument (*n* = 1)	TWS Scoring Rubric: revised version (Study 31)
	Evidence and standards (*n* = 5)	Competence Assessment* (Study 1), PDQ (Study 35), PEI (Study 15), STP (Study 30), Teacher education dispositions rating form (Study 26)
Two	Evidence and pre-existing instrument (*n* = 1)	CLASS: Toddler version (Study 13)
	Standards and pre-existing instrument (*n* = 1)	edTPA (Studies 4, 5, 12, 22, 33)
	Evidence and institutional conceptual framework (*n* = 1)	Disposition Assessment* (Study 25)
One	Pre-existing instrument (*n* = 1)	Intern Keys Teacher Candidate Assessment (Study 11)

Note. * A name assigned by the RA; this was done where tools did not have a specific name.

While some studies indicated that standards were used in the development of evaluation tools to define assessment criteria, thus clarifying expectations for both candidates and teacher educators and ultimately contributing to establishing content and construct validity (Studies 32, 40, 41, 42), the exact nature of these standards was not specified in these studies. Therefore, we sought to categorize tools according to the type of standards they drew on – i.e., whether these were state/county standards, institutional standards, or professional standards – to understand the extent to which authentic tools were created with content and construct validity in mind . Similarly, we broke down the evidence used in tool development by in-house data (such as data collected via need analysis and data on the views of university-based and school-based teacher educators), prior literature (such as academic research – i.e., evidence of effective teaching), theoretical frameworks (i.e., Danielson's framework), and professional association literature (i.e., recommended practices). These findings are presented in Table 3.17 along with the tools that drew on pre-existing tools and institutional conceptual frameworks.

In terms of evidence, this most commonly involved a combination of evidence from literature and from the programme (*n* = 4), followed by literature alone and in-house data alone (both *n* = 2). Looking at the use of standards, professional standards (*n* = 4) were the most common, followed by state/national standards (*n* = 2) and a combination of both types of standard (*n* = 1, PEI). No studies used institutional standards in tool development, but there was one instance where the tool (Disposition Assessment*) was created based on an institutional conceptual framework.

Additionally, four evaluation instruments drew on existing instruments or shared similarities with existing ones. Specifically, two were developed from candidate evaluation tools (TWS Scoring Rubric: revised version, CLASS: Toddler version), one from an in-service teacher

evaluation tool (Intern Keys Teacher Candidate Assessment) and one from both candidate and in-service teacher evaluation tools (edTPA).

Table 3.17

Developmental Grounding and Sources Used

Grounding	Type	Tool	*n*
Evidence (*n* = 8)	Evidence in the literature – academic research, theoretical frameworks, professional association literature	CLASS: Toddler version (Study 13), STP (Study 30)	2
	Evidence from the programme	Disposition Assessment* (Study 25), TWS Scoring Rubric: revised version (Study 31)	2
	Both literature and programme evidence	Competence Assessment* (Study 1), PEI (Study 15), Teacher education dispositions rating form (Study 26), PDQ (Study 35)	4
Standard (*n* = 7)	Professional standard	Competence Assessment* (Study 1), Teacher education dispositions rating form (Study 26), PDQ (Study 35), edTPA (Studies 4, 5, 12, 22, 33)	4
	State/national standard	STP (Study 30), TWS Scoring Rubric: revised version (Study 31)	2
	Both professional and state/national standard	PEI (Study 15)	1
Pre-existing tool (*n* = 4)	Candidate evaluation tool	CLASS: Toddler version: (Study 13), TWS Scoring Rubric: revised version (Study 31)	2
	In-service evaluation tool	Intern Keys Teacher Candidate Assessment (Study 11)	1
	Both candidate and in-service tool	edTPA (Studies 4, 5, 12, 22, 33)	1
Institutional conceptual framework (*n* = 1)	Institutional conceptual framework	Disposition Assessment* (Study 25)	1

Note. * A name assigned by the RA this was done where tools did not have a specific name.

While this section has focused on authentic tools used for assessing candidates specifically, it is worth briefly mentioning the development of authentic tools used in evaluation of practising teachers: these were mostly developed by state and county education departments (Studies 24, 28). Similarly, it is worth noting that our analysis of the development of identified emerging evaluation tools showed these were developed by independent researchers and grounded on a theoretical framework or prior academic research (Studies 3, 17, 38).

3.3.2.4 Design of the Tools

In this section, the design of the 11 authentic candidate evaluation tools is described. Specifically, we consider the focus of tools (i.e., disposition or effectiveness), the rating scales that were used (e.g., a 4-point scale), and the approach to rating (i.e., the number of raters, the strategies for deciding final ratings). Our analysis also includes an examination of the structure of the tools (i.e., domains), with a particular emphasis on their alignment with the three domains outlined in the UNESCO (United Nations Educational, Scientific and Cultural Organization) *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019): knowledge and understanding; teaching practice; and teaching relations.

Instrument focus. The question of what constitutes effective teaching remains a topic of debate (Study 27), and as a consequence, how teachers are evaluated is continually changing (Study 41). An examination of the 11 authentic candidate evaluation tools revealed that the most common focus is evaluation of instructional performance and competence (n = 8), followed by evaluation of teaching disposition (n = 3). While this suggests that performance-focused tools dominate, a closer look revealed that some instruments, aside from their main focus on teaching effectiveness, also contained elements related to disposition (e.g., the Competence Assessment* in Study 1, STP in Study 30). Tools that include dispositional assessment measure teachers' personal qualities or characteristics, including attitudes, beliefs, interests, appreciations, and values (Study 27).

Several studies problematized the utility of dispositions in candidate assessment. For instance, Study 26 examined the utility of dispositions by investigating the correlation between disposition results and the candidate's level of engagement while leading the class. The study reiterated the scepticism in Study 25 about the utility of dispositions to gauge teaching effectiveness due to the elusive nature of qualities like responsibility, respect, integrity, and caring/humanity, which lack clear definition and can result in unreliable and invalid assessments. Study 25 emphasized the high level of subjectivity attached to dispositions, which are not easily observable and quantifiable, leading to inconsistencies where raters may not be seeing the same thing. To address this, defining dispositions in *behavioural terms* in assessments was suggested (Study 26). The question of whether dispositions can be taught or whether they are inherent to the teacher, and thus cannot be taught, remained unanswered (Studies 26, 28).

Examination of the tools used in evaluation of *practising* teachers revealed an interesting picture, with no evaluation tools examining teacher disposition; rather, all evaluated teaching effectiveness (i.e., they were 'teaching focused'), assessing performance, competency, or the quality of preparation. Our examination of *emerging* tools unveiled several distinct foci, different from the dominant practices of evaluating teacher disposition or teaching effectiveness. For instance, an emerging instrument called SOCME-10 (Study 3) focuses on sustainable social development (it refers to this as being at the top level, Level IV). This tool diverges from traditional teaching-focused tools that assess 'what every teacher must do' (Level I; such as planning and learning assessment) and learning-focused tools centred on the learner experience (Level II; based on constructivism – applying learning in real-life situations). It also differs from tools focused on the learning context and social environment (Level III; based on socio-constructivism – includes creativity, pair work, collaboration, and

technology use). Instead, SOCME-10 considers an ideal for society to achieve, positioning its evaluation framework at a broader societal level (Study 3).

Another *emerging* evaluation instrument, PPK, demonstrates a traditional teaching focus in that it assesses teachers' competence (Study 38), but it differs from the traditional focus in that it assesses demonstration of content and pedagogical knowledge (Studies 27, 32). Indeed, the developers of PPK distinctively proposed the assessment of 'general pedagogical/psychological knowledge', denoting teachers' competence in creating and optimizing teaching–learning situations across various subjects, as opposed to subject-specific competence (Study 38).

Another emerging evaluation instrument, I-LAST (Study 17), is distinct from others because it places student learning at the centre of teaching effectiveness, instead of directly accounting for factors like teachers' education and experience. It was argued this is similar to what experienced teachers aim for as their practice evolves – in other words, teachers shift from the pre-service focus on themselves as instructors (i.e., surviving, optimizing lessons, giving clear instruction) to focus on the students as learners. The developers of I-LAST described it as a 'unidimensional item battery' providing diverse quantitative measures but with potential to be shortened and tailored to specific needs or used in longitudinal assessment of teachers as they progress through internships and further into their careers.

Tool format. As shown in Table 3.18, in the authentic tools for evaluating candidates, the format involved either a rubric (n = 8) or a rating scale (n = 3). Rubrics included comprehensive guidance to raters through detailed narratives for each domain and/or item. In contrast, rating scales, similar to Likert scales, lacked detailed description. The preference for rubrics is likely due to their perceived benefits in promoting consistent ratings and shared understanding during the evaluation process (Studies 30, 35). For example, while there might be consensus that 'professionalism' is an important indicator of effective teaching, interpretations of what constitutes professionalism can vary, so descriptions are significant (Study 27). Although rubrics were expected to enhance construct validity and IRR, research showed that their use does not ensure consistency (see Table 3.23).

Both formats included numeric scales and categorical scales (e.g., poor, fair, good, excellent). In terms of the scale range, 4-point scales were most common (n = 4), followed by 3-point scales (n = 3) and 5-point scales (n = 2). We identified one instance in which a dichotomous scale was used and one instance where a 7-point scale was used. The literature regarding scale range indicates problems associated with dichotomous scales and suggests that scales with five or more points provide more reliable and valid measures (Study 28).

Table 3.18

Tool Format and Scale Range

Format	Tool	*n*
Rubric with descriptive narratives	edTPA (Studies 4, 5, 12, 22, 33), Disposition Assessment,* Teacher education dispositions rating form (Study 26), PACT (Study 29), TWS Scoring Rubric: revised version (Study 31), PDQ (Study 35), STP, Competence Assessment* (Study 1)	8

Rating scale	Intern Keys Teacher Candidate Assessment, CLASS: Toddler version (Study 13), PEI (Study 15)	3

Scale range	Tool	*n*
Dichotomous	Disposition Assessment*	1
3-point scale	PDQ (Study 35), STP, Teacher education dispositions rating form (Study 26)	3
4-point scale	Competence Assessment (Study 1),* Intern Keys Teacher Candidate Assessment, PACT (Study 29), TWS Scoring Rubric: revised version (Study 31)	4
5-point scale	edTPA (Studies 4, 5, 12, 22, 33), PEI (Study 15)	2
6-point scale		0
7-point scale	CLASS: Toddler version (Study 13)	1

Note. * A name assigned by the RA this was done where tools did not have a specific name.

Approach to rating. In terms of the approach to rating, we examined the number of raters and the strategies for deciding final ratings – i.e., whether this was 'holistic', involving a single judgement for whole domains and items, or based on 'analytic scoring'. Our analysis was inconclusive due to insufficient information. Notably, the disposition assessment tool presented in Study 25 had different versions for use with candidates and teacher educators, but there was overlap between the versions.

Assessment dimensions. As has been noted, the debate continues on how to best define and evaluate dimensions of effective teaching (Study 17). Numerous tools have been developed to assess teacher education candidates, each with different foci and dimensions. It is within this context that we analysed the 11 evaluation instruments' dimensions, based on the three domains of the UNESCO Global Framework: knowledge and understanding; teaching practice; and teaching relations. Our analysis showed a strong focus on teaching practice (i.e., planning, instruction, and assessment). Knowledge and understanding, which are influenced by personal qualities and relationships, featured less prominently (Table 3.19). Study 30 concluded that in aligning the STP assessment procedure with the Chilean Ministry of Education's evaluation model for in-service teachers, researchers discovered an unintended benefit: the Samples of Teaching Performance) provided a potent way to link evaluation of pre-service and in-service teachers.

Table 3.19

Dimensions of Authentic Candidate Evaluation Tools, Categorized According to the UNESCO Global Framework

Source	Structure	Teaching knowledge and understanding	Teaching practice	Teaching relations
UNESCO	10 standards under 3 domains	Practising teachers know and understand: SI: How students learn and the particular learning, social, and development needs of their students S2: The content and related methodologies of the subject matter or content being taught S3: Core research and analytical methods that apply in teaching, including with regard to student assessment	Teachers' practices consistently demonstrate: S4: Planning and preparation to meet the learning objectives held for students S5: An appropriate range of teaching activities that reflect and align with both the nature of the subject content being taught, and the learning, support, and development needs of the students S6: Organization and facilitation of students' activities so that students are able to participate constructively, in a safe and cooperative manner S7: Assessment and analysis of student learning that informs the further preparation for and implementation of required teaching and learning activity	Teachers' professional relations include active participation in: S8: Cooperative and collaborative professional processes that contribute to collegial development and support student learning and development S9: Communications with parents, caregivers, and members of the community, as appropriate, to support the learning objectives of students, including formal and informal reporting S10: Continuous professional development to maintain currency of their professional knowledge and practice
Competence Assessment* Study 1	30 competencies (performance indicators) under 4 domains	No relevant content	Planning (S4) On-stage teaching (S4 & S7) Assessment (S7)	Professionalism (S8 & S10)
edTPA (Educative Teacher Performance Assessment) Studies 4, 5, 12, 22, 33	3 tasks, each comprising 5 rubrics, totalling 15 rubrics overall	No relevant content	Planning (S4) Onstage teaching (S4 & S7) Assessment (S7)	No relevant content

Intern Keys Teacher Candidate Assessment Study 11	Single factor structure, 'teaching effectiveness'; 72 criteria under 10 items (standards)	Incorporation of differentiated instruction (S1)	Instructional planning (S4) Implementation of instructional strategies (S4 & S7) Cultivation of a positive learning environment (S6) Creation of an academically stimulating atmosphere (S6) Use of assessments (S7) Application of assessment strategies (S7)	Effective communication (S9) Profound professional knowledge (S10) Demonstration of professionalism (S10)
Classroom Assessment Scoring System (CLASS): Toddler version Study 13	8 dimensions	Language modelling/support (S1) Regard for child perspectives (S1)	Facilitation of learning/development (S4 & S7) Behaviour guidance (S4 & S7) Quality of feedback (S4 & S7) Positive climate (S6) Negative climate (S6)	Teacher sensitivity (S10)
Profile for Evaluation of Intern (PEI) Study 15	30 criteria under 4 domains	No relevant content	Preparation and planning (S4) Instruction and classroom management (S4 & S7) Assessment (S7)	Professional development (S10)
Disposition Assessment* Study 25	17 sub-indicators (candidate version) and 23 sub-indicators (faculty version) under 6 disposition indicators	Diversity (S1) Scholarship (S3) Creative and critical thinking (S3)	No relevant content	Collaboration (S8) Caring (S10) Lifelong learning (S10)
Teacher education dispositions rating form Study 26	19 items under 6 dispositions	The belief that all students can learn (S1)	No relevant content	Responsibility (S10) Caring/humanity (S10) Fairness (S9) Respect (S8) Integrity (S10)

Performance Assessment for California Teachers (PACT) Study 29	12 rubrics (items) under 5 tasks	(Encouraging student use of) academic language (S2)	Planning (S4) Instruction (S4 & S7) Assessment (S7)	Reflecting (S10)
Samples of Teaching Performance (STP) Study 30	6 standards (dimensions)	Setting learning goals (S1)	Contextualizing teaching (S5) Lesson plans (S4) Assessment plan (S4 & 7) Decision-making and analysis of results (S7)	Reflective self-assessment (S10)
Teacher Work Sample (TWS) Scoring Rubric: revised version Study 31	24 rubric indicators under 5 key teaching and learning processes	No relevant content	Learning goal pre/post assessment (S4 & S7) Design for instruction (S4 & S7) Analysis of student learning (S7)	Contextual factors (S9) Reflection of teaching (S10)
Professional Development Qualities (PDQ) Study 35	12 disposition items	Respect for diversity (S1)	Student engagement (S5)	Collaboration (S8) Communication skills (S9) Professional appearance (S10) Professional responsibility (S10) Attendance (S9) Ethical behaviour (S9) Response to feedback (S10) Reflexive practice (S10) Professional initiative (S10) Portrays competence and confidence (S10)

Note. For each study, the elements listed are followed by the relevant UNESCO standard in brackets.

* A name assigned by the RA this was done where tools did not have a specific name.

An examination of the content of *emerging* tools identified some distinct content, not explicitly aligned with the three UNESCO *Global Framework* domains (Table 3.20) – one example was the pedagogical practice 'formation of universal values and an ethical life plan' in SOCME-10.

Table 3.20

Dimensions of Emerging Evaluation Tools

Tool	Structure	Elements
SOCME-10 Study 3	10 core pedagogical practices under the factor 'mediation of problem-based training, collaboration, and inclusion'	Motivation to achieve the expected learning Concept learning through graphic organizers and case studies Solving real problems Formation of universal values and an ethical life plan Assertive communication Collaborative work Development of creativity and innovation Application of transversality Resource management Product-based formative assessment
Item-Level Assessment of Teaching Practice (I-LAST) Study 17	94 items across 6 dimensions	Management (student and classroom management) Student accountability (teacher–student responsibility) Assessment (student evaluation) Teacher accountability (self-accountability) Individualizing instruction (tailored teaching) Literacy (literacy content and practice)
Pedagogical and Psychological Knowledge Study 38	39 items across 2 dimensions and 4 sub-dimensions	Classroom processes (knowledge of classroom management, teaching methods, assessment) Students' heterogeneity (knowledge of learning processes, individual student characteristics)

Note. Due to insufficient information, the evaluation tool EDA (Study 27) could not be included in the analysis.

Process and procedures involved in judgement-making. We analysed the process of implementation of the tool and use of results in TEPs according to seven key dimensions (Table 3.21). This included examining use of evaluation results, categorizing tools based on their use for summative decisions (impacting final outcomes, like certification), one-time feedback (providing immediate insights), and progress-oriented feedback (supporting ongoing development). We also examined evaluation approaches, differentiating between summative evaluations (assessing final outcomes) and formative evaluations (offering feedback for improvement). Additionally, we examined the evaluation site, identifying whether the context for evaluation was field experiences (i.e., practical settings, like internships) or programme courses (i.e., an academic environment) as well as the frequency of evaluations. Furthermore, we examined assessment methods, such as observation and portfolio assessment and who conducted the ratings – university-based teacher educators,

candidates, and so on. Finally, we examined whether training was provided to raters and, if so, what the coverage was (i.e., full training for all, training for some groups only).

Table 3.21

Implementation of Tools

Use of evaluation results	Tool	*n*
Summative decision	Competence Assessment (Study 1),* edTPA (Studies 4, 5, 12, 22, 33), PEI (Study 15), PACT (Study 29), TWS Scoring Rubric: revised version (Study 31)	5
One-time formative feedback to guide candidate improvement	Competence Assessment* (for all; Study 1), PEI (conditional; Study 15), PACT (conditional; Study 29)	3
Progress-oriented formative feedback and monitoring	CLASS: Toddler version (Study 13), Disposition Assessment,* PDQ (Study 35)	3
No information	Intern Keys Teacher Candidate Assessment, Teacher education dispositions rating form (Study 26), STP	3

Evaluation approach	Tool	*n*
Summative	edTPA (Studies 4, 5, 12, 22, 33), TWS Scoring Rubric: revised version (Study 31)	2
Formative and summative	Competence Assessment (Study 1),* PEI (Study 15), PACT (Study 29)	3
Progress and monitoring oriented	CLASS: Toddler version (Study 13), Disposition Assessment,* PDQ (Study 35)	3
No information	Intern Keys Teacher Candidate Assessment, Teacher education dispositions rating form (Study 26), STP	3

Evaluation site	Tool	*n*
Field experience	Competence Assessment (Study 1),* edTPA (Studies 4, 5, 12, 22, 33), Intern Keys Teacher Candidate Assessment, CLASS: Toddler version (Study 13), PEI (Study 15), PACT, STP	7
Single programme course	TWS Scoring Rubric: revised version (Study 31)	1
Multiple programme courses	Disposition Assessment*	1
Both programme course and field experience	PDQ (Study 35)	1
No information	Teacher education dispositions rating form (Study 26)	1

Frequency of evaluation	Tool	*n*
One time	edTPA (Studies 4, 5, 12, 22, 33), PACT, STP, TWS Scoring Rubric: revised version (Study 31)	4
Two times	Competence Assessment (Study 1),* PEI (Study 15)	2
More than two times	CLASS: Toddler version (6 times; Study 13), Disposition Assessment,* PDQ (Study 35)	3
No information	Intern Keys Teacher Candidate Assessment, Teacher education dispositions rating form (Study 26)	2

Assessment method	Tool	*n*
Observation and self-assessment	Competence Assessment (Study 1),* PEI (Study 15), Disposition Assessment,* PDQ (Study 35)	4
Portfolio/work sample of candidate	edTPA (Studies 4, 5, 12, 22, 33), PACT, STP, TWS Scoring Rubric: revised version (Study 31)	4
Observation	Intern Keys Teacher Candidate Assessment, Teacher education dispositions rating form (Study 26)	2
Observation and peer assessment	CLASS: Toddler version (Study 13)	1

Raters	Tool	*n*
University-based teacher educators	Competence Assessment (Study 1),* CLASS: Toddler version (Study 13), PEI (Study 15), Disposition Assessment,* PDQ (Study 35), TWS Scoring Rubric: revised version (Study 31)	6
Candidates	Competence Assessment (Study 1),* PEI (Study 15), Disposition Assessment,* PDQ (Study 35)	4
School-based teacher educators	Competence Assessment (Study 1),* PEI (Study 15), PDQ (Study 35)	3
Qualified raters	edTPA (Studies 4, 5, 12, 22, 33), PACT	2
Candidate peers	CLASS: Toddler version (Study 13)	1
No information	Intern Keys Teacher Candidate Assessment, Teacher education dispositions rating form (Study 26), STP	3

Training for raters	Tool	*n*
For all	edTPA (single rater; Studies 4, 5, 12, 22, 33), PACT (single rater), TWS Scoring Rubric: revised version (single rater; Study 31)	3
For some	Competence Assessment* (only candidates; Study 1), PEI (only mentors; Study 15), Disposition Assessment* (only candidates)	3
No training	Teacher education dispositions rating form (Study 26)	1
No information	Intern Keys Teacher Candidate Assessment, CLASS: Toddler version (Study 13), PDQ (Study 35), STP	4

Note. * A name assigned by the RA this was done where tools did not have a specific name.

Our examination revealed that evaluation results were used in both summative decision-making (n = 5) and efforts to support growth (n = 6). Among the five tools that were used in summative decision-making, three also contributed to formative evaluation, but this was aimed at providing one-time formative feedback to facilitate candidate improvement and not necessarily ongoing feedback and support for growth. It is also noteworthy that only one of these three (Competence Assessment*) provided support for all candidates, with no threshold applied in terms of candidates' results. The other two provided support conditionally, targeting students who had failed (PACT) or fallen below a specific threshold (PEI). Essentially, those tools providing one-time formative feedback were primarily geared towards supporting summative decisions rather than fostering ongoing growth. Study 39 suggests that the most essential element of formative evaluation, to make it '(in)formative', is to provide univocal, cognitively acceptable, and relevant feedback on performance throughout the learning journey, so it cannot be realized by corrective or general feedback based on knowledge of results, as happened in many cases with the examined tools.

The way formative assessments were arranged affected the realization of benefits, as confirmed by Studies 13 and 15. Three tools were used to support candidates' growth through progress-oriented formative feedback, which included monitoring and tailored interventions. Of those, two focused on improving disposition (Disposition Assessment,* PDQ) and one focused on improving teaching effectiveness (CLASS: Toddler version). The practices related to these three tools are summarized in Table 3.22.

CLASS: Toddler version was administered six times during a field experience, with weekly ratings and feedback. Disposition Assessment* was administered throughout a programme, in each semester and across each course. PDQ was implemented three times – once at the outset of a programme and twice during field experiences. We found no instances where progress-oriented formative feedback and monitoring directly contributed to or influenced summative decisions.

Table 3.22

Evaluation Practices to Support Growth

Tool	Feedback type and frequency
Classroom Assessment Scoring System (CLASS): Toddler version implementation Study 13	**Observation:** formal feedback (narrative and quantitative feedback based on CLASS scores on a weekly basis over 6 weeks) and direct supervision and individualized informal feedback (daily, during/after classroom visits) from faculty; direct supervision and immediate informal feedback from classroom teachers **Video-based feedback:** informal feedback from peer coaching groups on a weekly basis **Candidates:** a summative reflection on their semesters and assessment experiences **Use of scores:** • to monitor candidates' progress and determine the level of tiered support: o universal support, provided to candidates showing significant progress by Week 2 o targeted support, for those not making progress by Week 2 o intensive support, for candidates without clear growth by the final 2 weeks
Disposition Assessment* Study 25	**Observation-based rating:** by programme instructors each semester for every course throughout the programme **Self-evaluation:** candidates identify two disposition goals at the beginning of each semester, after receiving training about the tool. **Continuous assessment:** ongoing monitoring to track development and identify areas that need improvement **Use of scores:** • for identification of disposition status (Levels 1 to 4): o to make assessments based on the volume and frequency of identified weaknesses o to note persistent weaknesses across semesters are noted o to evaluate candidate-prepared development plans and their efficacy o to track Teacher Education Deficiency Reports filed per semester • to support strategies based on disposition status: o Level 1 – candidates make development plans to address identified areas o Level 2 – continual monitoring and support to ensure progress o Level 3 – additional support measures to accelerate improvement o Level 4 – committee investigation to determine the extent and causes of issues, leading to a tailored remediation plan using candidate self-ratings
Professional Development Qualities (PDQ)	**Observation and self-evaluation:** Each group (candidates, mentor teachers, and university supervisors) provides independent ratings three times. There is no intention to align ratings, as the evaluation is learning

Study 35	focused. The initial completion of the PDQ typically occurs in the candidates' first teacher education class and then once in the middle and once at the end of the student teaching experience in the field (this is also the end of programme). The PDQ results are stored and managed on the Watermark (Livetext) online platform. **Use of scores:** to track candidates' professional dispositions over time; to encourage discussions among candidates and those who support them in school settings about important disposition characteristics

As shown in Table 3.21, the most common setting where evaluations took place was field experience. In seven cases (Competence Assessment,* edTPA, Intern Keys Teacher Candidate Assessment, CLASS: Toddler version, PEI, PACT, STP), this was the sole setting for evaluations, and in one instance (PDQ), evaluations were carried out in both programme course and field experience settings to assess candidates' development across different stages of their education journey. For those tools used only in programme courses, one (Disposition Assessment*) was used in each course of the programme throughout each semester to determine disposition levels and one (TWS Scoring Rubric: revised version) was used to assess work samples submitted by candidates during a programme course called 'Student Teaching Seminar Course', which was taken in the last semester of the TPP (senior culminating experience).

Turning to assessment methods, Table 3.21 shows that a combination of self-assessment and observation (n = 4) and portfolio/teacher candidate work (n = 4) were the most common, followed by observation alone (n = 2). Portfolios included a collection of authenticated artifacts, such as lesson plans, student work samples, and video-recorded evidence (for three to five lessons; Study 12), and reflective commentaries (Study 4). One further case incorporated peer assessment in conjunction with observation. Notably, the ratings obtained from peer assessment were not integrated into the decision-making process concerning the candidates (i.e., to determine the level of support the candidate needs). Rather, these were considered as a valuable means for peers to familiarize themselves with the tool (CLASS) their university supervisors were using to rate them (Study 13). We found no instance of value-added measures or classroom student evaluation of candidates, despite these two forms of assessment being widely entrenched in evaluation of practising teachers.

Our examination of raters, the individuals or groups responsible for conducting evaluations, was based on data for 8 tools (Table 3.21). This revealed that rating was most typically carried out by university-based teacher educators (n = 6), followed by candidates themselves (n = 4) and school-based teacher educators (n = 3). Notably, for two tools, qualified raters (external or local raters) carried out the assessment. In one instance, ratings were gathered from candidate peers. Furthermore, while self-rating was prominent in the context of formative evaluation, this did not extend to the decision-making level. This observation aligns closely with the key role played by school-based educators in the evaluation process. Nevertheless, it is worth noting that, as demonstrated in Study 35, teacher candidates themselves acknowledged the value of 'completing' disposition assessment forms multiple times, as it contributed to their self-awareness and ongoing development of dispositions.

Our examination of training for raters was based on data for seven tools (Table 3.21). For three tools (edTPA, PACT, TWS Scoring Rubric: revised version), training was provided for all raters, but in these cases raters acted individually were required to

have a certificate (see edTPA , PACT). For three tools, training was selectively provided, either for candidates (n = 2; Competence Assessment,* Disposition Assessment*) or for school-based teacher educators (n = 1; PEI). In one instance (Teacher education dispositions rating form), no training was provided to any raters. However, it is possible that the absence of training data in relation to four tools reflects lack of training rather than insufficient information in the relevant studies.

3.3.2.5 Measures for Rating Quality and Inter-Rater Reliability in Teacher Education

In the context of teacher education, the processes involved in ensuring the quality of teacher assessments are complex, demanding, and time-consuming, leading to challenges in their integration within educational institutions. However, measures to promote reliable judgement-making in teacher education are crucial, and feasible strategies should be developed for all those involved in assessment (Study 41). Our analysis identified several strategies for rating reliability and IRR:

- **double scoring:** carried out for a sample of assessments, for quality assurance purposes. PACT portfolios underwent double scoring as per state policy;
- **additional scorers for borderline cases:** used to come to a decision where candidates fail or come close to failing an assessment; i.e., edTPA (recommended), PACT;
- **combining rating:** accumulating ratings; i.e., PEI (only where ratings were close), Competence Assessment* (direct accumulation);
- **dispute resolution through interpretative/verbal resolution methods:** PEI (only where ratings were divergent); a trio of raters – teacher candidates, a university-based teacher educator, and a **school-based teacher educator –** discussed significant differences in ratings and assessed additional documents prepared by candidates;
- **dispute resolution through committee investigation:** i.e., Disposition Assessment* (only where concern about a candidate was raised by a university-based teacher educator; this prompted referral to a committee for analysis of the candidate's scores across multiple assessments to investigate whether the candidate's weaknesses were localized to a specific course with a particular instructor or extended across multiple settings with self-assessment scores serving as a reference point); and
- **candidate feedback:** feedback by candidates about their overall experience at the end of field experience (CLASS: Toddler version). Interviews with candidates post field learning aimed to gather feedback on their experiences with **school-based teacher educators** and the assessment process.

3.3.3 Reliability

Reliability signifies consistent, replicable, and dependable results (Cohen et al., 2018). Our examination revealed that findings related to the nature of reliability of judgements of teaching effectiveness were focused within four areas: internal consistency reliability, IRR, influences on rater reliability, and proposed ways to improve reliability. Table 3.23 shows which tests were used to estimate tool reliability.

Table 3.23

Tests Employed in Estimating Rating and Tool Reliability

Internal consistency reliability	Reliability coefficient (Study 7) Cronbach's alpha (Studies 3, 11, 35) Cronbach's alpha ('classical test theory') and Rasch perspectives ('modern psychometric technique') (Study 17) Second-order confirmatory factor analysis to confirm constructs (Study 18) Kuder-Richardson 20 (Study 25) Confirmatory factor analysis (Study 26) Pairwise correlations between domains and items (Study 28) Standard deviations to compare holistic and analytic scores (Study 31) Cronbach's alpha and standard deviation and means (Study 38)
Consistency and accuracy	**IRR:** Qualitative interview data used as evidence for estimating IRR (Study 9) IRR calculated using Pearson product-moment correlation coefficient (Study 27) IRR calculated using Cohen's kappa (Studies 26, 29) IRR calculated using standard deviation (near or exceeding 1.0 deemed less consistent) from 'true scores' (Study 31) Similarity in mean scores (generalized estimating equation model) used as evidence for IRR (Study 35) Level of alignment between raters calculated with exact and partial agreement (Study 36; no mention of IRR) **Accuracy:** Accuracy calculated by comparing raters' scores with an identified 'true score' (Study 23) Chi-square tests used to examine the matching items (i.e., areas of concern) between raters (Study 25) Accuracy based on standard deviations and number of times scores differed by more than two score levels from 'true scores' (Study 31) Consensus estimates (inter-rater agreement) calculated using a combination of perfect agreement with true scores and ratings +/– one acceptable level of true scores (Study 34)

Note. IRR: inter-rater reliability

3.3.3.1 Internal Consistency Reliability

Internal consistency reliability reflects the extent to which items within an evaluation instrument measure the same underlying construct (Cohen et al., 2018) – in this case, *teaching effectiveness*. In the represented studies, researchers tended to prioritize internal consistency reliability, testing for this more often than other forms of reliability. Several studies revealed that consistency and accuracy of assessments across raters and time tended to be more prevalent in holistic scoring compared to analytic scoring (Studies 23, 26, 31, 35), suggesting scores may be more reliable when used in a holistic manner (Study 31). In Study 28, the degree of consistency among items of the T-TESS rubric considered the statistical

properties and the extent to which it differentiated teachers on teaching quality. Further, Study 23 noted that certain elements may have been harder to rate than others, but that differences in administrators' reasoning were not related to accuracy. A number of studies looked to ensure dependable and consistent results in the same setting with the same type of subjects (Studies 11, 23, 41).

Researchers used various methods to ensure that evaluation tools were reliable and provided consistent information about teacher effectiveness or other constructs of interest. Cronbach's alpha was the most frequently used (Studies 3, 11, 17, 35, 38); this tests for internal consistency, measuring how well the items in a tool (e.g., questions about teaching practices) work to assess the intended construct (e.g., teaching effectiveness; Cohen et al., 2018). However, 'classical test theory', i.e., Cronbach's alpha, was critiqued in Study 17, and the researchers used Rasch analysis, described as a 'modern psychometric technique', in addition to Cronbach's alpha. Study 38, in addition to Cronbach's alpha, used standard deviation and means to confirm the results.

Other studies explored alternative reliability measures. Study 7 used the reliability coefficient value, and Studies 18 and 26 employed confirmatory factor analysis (Study 18 used second-order confirmatory factor analysis specifically) to assess internal consistency reliability. Study 25 used Kuder-Richardson 20, and Study 28 employed pairwise correlations between the sub-dimensions and item ratings to examine internal consistency. Study 31 calculated standard deviation levels near to or exceeding 1.0.

3.3.3.2 Inter-Rater Reliability

IRR is a specific form of reliability that focuses on the level of agreement among different observers who are evaluating the same thing (Cohen et al., 2018). Several studies examined IRR, considering the consistency of the judgements of several raters. Some studies confirmed instances of inter-rater agreement and consistency (Studies 23, 24, 25, 31, 34, 35, 36). Others revealed inconsistencies and disagreements between raters (Studies 9, 23, 25, 26, 29, 31, 35). Analysis identified two notable patterns. Candidates tended to rate themselves lower than peers and school-based teacher educators rated them (Study 36), yet their self-ratings were either similar (Studies 25, 35) or lower than ratings assigned to them by university-based teachers (Study 36). The second pattern was that school-based teacher educators' ratings were almost always higher than both teacher candidates' (Study 35) and university-based teacher educators' (Studies 9, 35). The review also noted inconsistencies between school-based teacher educators and university-based teacher educators (Studies 9, 26). In Study 35, statistically significant similarities in overall mean scores between teacher candidates and supervising faculty regarding professional dispositions were found. The study also noted statistically significant higher rating from school-based teacher educator teachers over university-based teachers not only at one point in time but across time. This was deemed an important finding, as school-based teacher educators ('mentors') were 'professional teachers in the field observing the actual teaching practices and dispositions of teacher candidates' (Study 35, p. 128). However, another study interpreted 'overly positive' ratings from school-based teacher educators over university-based teacher educators as the 'weaknesses' of the school-based teacher educators in assessing the readiness of pre-service teachers (Study 9, p. 242).

A majority of studies employed descriptive statistics using either exact ('assigned same score') or partial percentage agreement ('adjacent agreement', assigned either the same or within a difference of one point) or standard deviation or comparison of raters' scores with an identified 'true score' (Studies 23, 31, 34, 36). One study (Study 25) use a chi-square test to examine the match between raters. Studies which calculated IRR based on advanced statistical techniques (i.e., Cohen's kappa, weighted kappa) were less prevalent (Studies 26, 29). One study calculated IRR using Pearson product-moment correlation coefficient (Study 27), another considered 'similarity in mean scores' (generalized estimating equation model) as evidence for IRR (Study 35), and another used qualitative interview data as evidence for estimating IRR (Study 9). Measures such as Cohen's kappa and intra-class correlations were recommended for accurate reporting and to account for chance agreement (Study 29).

3.3.3.3 Influences on Rater Reliability

Our analysis revealed that a widely shared objective in evaluation of teaching was to ensure that judgements are accurate and reliable. One study investigated whether estimates of IRR in teacher education are influenced by the statistical methods used (Study 29), and one compared pre- and post-training rater agreement to explore whether digital training influenced levels of agreement in rating of teaching (Study 34). Study 2 investigated whether pre-service teachers' judgements of an instructor are influenced by factors such as the instructor's native language, the gender of pre-service teachers, and the location (Germany and Australia). Study 21 examined whether employment supervisors exhibit bias based on new teachers' socioeconomic status and ethnicity. Studies exemplified inconsistencies and inaccuracies in judgement due to five main factors: ratees; raters; tool characteristics; deployment of evaluation; and methods used to determine reliability and validity.

Ratees characteristics and contexts. Very few studies in the context of teacher education have identified influences on rater reliability that are related to the ratees. Some studies considered the context that ratees work in. No study found that the gender of candidates being rated influenced the rating they receive. In Study 2, candidates' gender was found to be unrelated to their evaluation of a teacher's skills. Ratee characteristics influenced judgements in two studies (Studies 23, 36). And while ethnicity significantly affected outcomes like edTPA pass rates (in Study 12, Hispanic candidates in Washington were over three times more likely to fail the edTPA), this was not always a significant factor (e.g., in Study 21, which involved principals' ratings of new teachers). Nonetheless, there have been calls to address bias against teacher candidates from minority backgrounds to improve diversity in the teacher workforce (Studies 12, 15).

Candidates' alignment to the task requirements (i.e., writing ability) influenced the quality of the assessment product, consequently affecting the final grades assigned (Studies 12, 30). Studies 24 and 28 both identified that school characteristics influence the ratings assigned to practising teachers. In schools with relatively high numbers of students eligible for free meals, low-income schools, and schools with a relatively high number of special education students or ethnic minority students, teachers received lower ratings, whereas teachers in more advantageous settings (such as those teaching gifted students) received higher ratings. This trend was observed in both evaluations by principals and students (Study 24) and evaluations by externally qualified raters (Study 28).

Study 24 explored ratings of teachers working in different grade levels, finding that elementary school teachers tended to receive higher ratings than those teaching higher grades, and ratings of teachers in different subjects, finding that teachers of subjects like English typically received higher scores. Schools with a larger proportion of more experienced teachers (Study 28) and experienced teachers (Study 24) received higher ratings. Another study looked at bias of supervisor principals when rating new teachers preparedness to teach and found no bias based on the education level of the new teachers' mothers and fathers or their family income (Study 21).

Raters. Some studies looked at the effects of raters' background, characteristics, cognitive processes, beliefs, preferences, and prejudices on their judgements (Studies 11, 23, 24, 29, 31, 41). Two studies showed that some judgements deviated from the formally designated tasks, incorporating non-scoring criteria based on experience from conducting other evaluations, personal teaching and rating experience, and unconscious mental resource use (Studies 23, 29). Study 23 found that 86% of 35 trained administrators used internal criteria at least once during their scoring. Study 29 showed that the same raters produced different results at different times. Study 24 explored whether some principals had higher standards and were 'tougher' or 'more lenient' than others in rating in-service teachers; findings indicated principals are not entirely consistent in how they applied the school systems rubric Further, Study 31 uncovered that when raters found the candidate's work samples (based on TWS Scoring Rubric: revised version) were better than in previous semesters, they inflated the scores. Two studies evidenced rating errors stemming from raters' misinterpretation of the scoring rubrics and prompts (Studies 29, 31), and one of these (Study 29) highlighted to need for cognitive problem-solving skills to manage multiple simultaneous considerations (i.e., use of academic language, evidence of planning, and evidence of implementation by identifying lesson artifacts). However, the gender of pre-service teachers did not influence their judgement of the teacher (Study 2).

It was argued that by relying on subjective criteria, raters cannot accurately assess ratees' effectiveness (Study 23), and that using subjective criteria reduces consistency between raters (Study 29). One study showed that administrators' personal expertise influenced how they processed observed behaviours – in other words, whether it was the third time the rater had used the rubric or the 33rd time affected their assessment (Study 23). However, no study addressed whether the level of teacher educators' experience influences their judgements. In addition, no study addressed whether the social groups raters belong to (e.g., their ethnicity, socioeconomic status) influences their judgement-making. In addition, no study examined whether fit between ratee and raters in terms of teaching subject area influences judgement. However, Study 34 argued that lack of improvement in IRR following training could have been due to raters' lack of content knowledge in relation to ratees' subjects.

Several studies argued that consistencies in assessments might stem from raters' preferences and mindset (Studies 9, 35, 36). Even when rating scales were based on a broad range of scores, practices such as clustered rating (i.e., non-normal distribution) persisted, and certain grades were assigned disproportionately, either around the middle or high end of evaluation scales (Studies 29, 35). For instance, Study 29 found that on a 1 to 4 scale, a score of 4 was given in less than 2% of the ratings. Similarly, Study 35 observed that most raters (candidates, university-based and school-based teacher educators) assigned the highest score (3) on a 3-point scale. Additionally, binary ratings were common, with principals often

adopting a mindset of satisfactory/unsatisfactory when using a 4-point scale, to potentially avoid conflict with ratees and teacher unions (Studies 16, 21, 28). It was argued that such skewed distribution of scores, with a halo effect (Study 35) or positive mindset (Study 9), served to slant estimates of IRR upwards, producing unreliable measures of true evaluation consistency (Study 29).

Study 23 examined social dimensions of reasoning using ratings of a teacher featured in a video; even though there were no prior personal relationships between ratees and raters, the intention was to remove social influences as far as possible. This study found that secondary administrators assigned an established 'true' rating slightly more often than primary administrators (57% and 44%, respectively), but this difference was not statistically significant.

Two studies suggested rater confidence plays a crucial role in reliability. Additionally, low confidence of candidates in the rating practice itself, and excessive self-criticism, potentially impacted reliability (Studies 35, 36).

Tool characteristics. A restricted range in the rating scale (Studies 28, 35), such as having only three options, left little room for variability and could lead to consistently high ratings (Studies 28, 29, 35). This reduced the ability of raters to capture nuances in teacher preparation (Study 21) and effectively differentiate teachers' performance (Study 28). In addition, it was argued that measurability of evaluation domains and items – i.e., observable behaviours (Studies 25, 26) – and practicality and complexity of evaluation tools influenced how judgements were made (Studies 8, 29). The vital importance of validated tools for reliable judgements was highlighted (Study 41).

Deployment of evaluation. Studies highlighted that the deployment of evaluation plays an important role in reliability of ratings. The intended purpose of evaluation, particularly in high-stakes contexts, can significantly affect the dependability and validity of ratings. For instance, Study 24 illustrated the risk that ratees, to enhance their ratings, prioritize superficial behaviours or compliance with the expected behaviours rather than focusing on genuine improvement in their performance. Additionally, contextual elements such as rater training (Studies 1, 4, 9, 31), assessment methods (Studies 20, 24, 35), the number of raters involved (Studies 24, 41), and the frequency of assessments (Studies 16, 35) all contributed to rating reliability.

Methods (i.e., data collection, data analysis) of determining reliability and validity. A masking effect of quantitative data was evidenced in Study 2, which appeared to suggest that qualitative data could unearth biases in raters' judgement that were not revealed by quantitative data; it was argued that the use of multiple sources of evidence would overcome this issue.

Use of percentage agreement to measure IRR led to an inflating effect, as evidenced in Study 29: compared to use of Cohen's kappa to calculate IRR estimates, when percentage agreement was used, the agreement level rocketed up from 19% to 99% due to lack of consideration of chance agreement in percentage agreement.

3.3.3.4 Proposed Ways to Improve Reliability

The most widely taken action to improve reliability has been the standardization of sources of information, scoring, and criteria (Studies 33, 41), the fundamental idea being to exclude contextual influences (Study 33). However, standardization did not guarantee objective and reliable judgements (Study 4), and it was found that standardized assessment tools did not always align with the context of specific subject areas (e.g., art teacher Study 33), were unresponsive to programme values and candidate needs (Studies 4, 33), and disregarded the real-time context of teacher–student relationships (Study 33). However, some studies did argue that standardization can impact validity and intended outcomes (Study 41). Study 33 suggested a more radical approach, granting autonomy to university- and school-based teacher educators to choose contextually appropriate evaluation tools due to their unique understanding of the context.

Training was one of the most frequently suggested solutions for improving reliability and validity of judgements (Studies 1, 26, 27, 31), for all raters (Studies 1, 4, 26, 31) and explicitly for school-based teacher educators (Study 9). Study 31 recommended 'regular training' rather than one-off training, and it advised a session on quality control in relation to scoring. Study 34 suggested the need for more effective training materials and enough time for scorers to engage with training materials or the assignment itself, especially where training is online. Further studies specifically focused on the impact of training, using pre- and post-training tests; these found poor inter-rater agreement (Study 29) and little to no improvement in inter-rater agreement (Study 34). Study 34 concluded that IRR improved post training for some types of assessment (i.e., research papers, case studies) but decreased for others (i.e., digital portfolios).

Some empirical studies concluded training was not an effective solution (Studies 23, 29, 31), even with extensive training and sessions on quality control of scoring (Study 29). Though for most trainee scorers' (80%), self-perceived level of reliability and confidence in rating increased, and they spent more time on scoring after training (Study 34). In Study 23, school administrators exhibited a variety of reasoning strategies to justify judgements. Interestingly, findings indicated that variation in reasoning strategies did not affect the accuracy of ratings. Study 41 noted that evaluations of teaching have evolved to include methods such as peer assessment, self-assessment, portfolio assessment, and simulated teaching. The combination of supervisor observation and candidate self-reporting has been recommended (Studies 35, 39, 41), and this could be a way to validate self-evaluations (Study 41). It has also been suggested that peer rating should be integrated alongside other forms of assessment, as this would help not only the rated but also the candidates in rating their own learning (Studies 13, 36).

Other recommendations to improve reliability included having multiple raters rather than a single rater (Studies 24, 35), employing a variety of assessment methods (Study 24), and assessing multiple times (Studies 16, 17, 35). Study 17 emphasized that drawing accurate conclusions from observational data requires more extensive, frequent, and prolonged observations than the typical 1 or 2 hours per year. Study 24 highlighted that reliability can still be achieved with a single rater if the evaluation process involves extensive evidence collection throughout the year; this would give the rater much more information than would be available to, say, an anonymous reviewer relying on a video of classroom practice.

Constructing objective indicators for assessment was reported as important to mitigate potential subjectivity in judgement-making (Studies 25, 26), as some indicators were challenging to operationalize. Study 26, in particular, suggested conceptualizing dispositions as a single, global dimension rather than as a set of separate dimensions. A few studies suggested strong university–school collaboration for a common understanding of teacher education and evaluation (Study 9). Study 31 suggested adjustment of evaluation tools based on collaborative discussion among faculty, and Study 1 suggested that rather than adapting an existing instrument and mandating its use, instruments should be modified in collaboration with schools, due to potential differences in the perspectives of developers and users of the instrument.

Portfolios of student teachers' work have also been suggested as a way to improve reliability (Study 20), and these are already widely used in many TEPs (Studies 20, 41), especially in the US. However, the intentions behind the use of portfolios – i.e., to be student centred and learning oriented – has been found to be flawed (Studies 20, 22), as they are used instead for organizational needs, such as quality assurance (Study 20), or their use turns into a procedural formality (Study 22).

Active involvement of candidates in self-evaluation of their own effectiveness created opportunities for growth, in particular when self-ratings were deliberated alongside school-based teacher educators' assessments, demonstrating triangulation (Studies 15, 35). This fostered candidates' autonomy, self-regulated learning, self-reflection, and self-monitoring practices (Studies 15, 36, 39). Study 13 further found that engagement with multiple forms of evaluation feedback (e.g., direct, immediate) throughout multiple evaluation points was supportive. Study 24 found that use of multiple measures (i.e., rating by principals, student surveys, and value-added achievements) in in-service teacher assessment complemented each other. In Study 28, the use of triangulation was evident, with focus group discussions used to confirm questionnaire results.

3.3.4 Validity

Several studies explored instrument validity, and the tests used are listed in Table 3.24. In assessing instrument validity, face, content, and construct validity emerged as the most common types considered, and predictive validity was the least common type. Notably, it appears that in Study 14, content validity may have been mistaken for face validity.

Our results confirmed the claim made in Study 17 that work on validation tended to be based on a classical test theory validation framework – i.e., using confirmatory factor analysis and Cronbach's alpha. This was criticized in Study 17, which favoured modern psychometric techniques such as the Rasch framework.

Table 3.24

Tests Used in Tool Validation

Study	Type of validity	Analysis	Data collection approach and data set
1	Construct validity	Exploratory structural equation modelling – combination of confirmatory factor analysis and structural equation modelling	Quantitative: pre-existing 'ratings' of candidates by candidates and school-based and university-based teacher educators
3	Content validity	Aiken's V. acceptance criterion: V > 0.80 and lower confidence limit over 0.6	Mixed: assessment/judgement of subject experts and new teachers through a survey and verbal remarks
	Construct validity	Factorial structure through exploratory factor analysis and confirmatory factor analysis	Quantitative: pilot of an instrument to assess new teachers
4	Face validity	Thematic analysis	Qualitative: interviews with university-based teacher educators
	Consequential validity	Thematic analysis	Qualitative: interviews with university-based teacher educators
5	Consequential validity	Thematic analysis	Qualitative: textual analysis of coursework and portfolios
6	Predictive validity	Multilevel analyses of correlations (also descriptive statistics)	Quantitative: pre-existing 'evaluation results' and candidates' entry characteristics to teacher education (i.e., GPA), candidates' grades during teacher education, classroom students' ratings of their teachers
7	Content validity	Content validity index and content validity ratio	Mixed: assessment/judgement of subject experts through focus groups and a survey
	Construct validity	Confirmatory factor analysis	Quantitative: a pilot of administration to university students
8	Face validity	Descriptive statistics – percentage, mean, standard deviation, variance, weighted means score	Mixed: satisfaction and experience of school-based and university-based teacher educators through surveys and focus groups
10	Face validity	Mean, standard deviation, *t* test, analysis of variance	Quantitative: assessment of teacher educators and prospective teachers through surveys
11	Construct validity	Exploratory factor analysis	Quantitative: university teachers' 'rating' of candidates

12	Predictive validity	A simple logit model (for Equation 1, see page 383), ordinary least squares (for Equation 2, see page 383), stacked model (for teacher's effectiveness)	Quantitative: pre-existing 'evaluation results', candidate teachers' employment records, teachers' effectiveness (measured by value-added results)
	Consequential validity	Two-sample *t* test	Quantitative: pre-existing 'evaluation results', candidate teachers' employment records and ethnicity
13	Consequential validity	Thematic analysis	Qualitative: interviews with candidates
14	Content validity	Content validity index	Quantitative: assessment/judgement of expert judges through surveys
	Face validity	Cohen's kappa index	Mixed: assessment/judgement of expert judges through written comments and surveys
15	Consequential validity	Thematic analysis	Qualitative: pre-existing 'evaluation results' and focus groups with candidates
16	Face validity	Percentage	Mixed*: assessment/judgement of school administrators through questionnaire with open-ended questions
	Construct validity	Rasch partial credit model/Rasch framework (described as 'modern psychometric technique')	Quantitative: school-based teacher educators' 'rating' of candidates
17	Content validity	Percentage (2 out of 3 deemed appropriate)	Quantitative: assessment/judgement of university-based and school-based teacher educators
	Face validity	Percentage (2 out of 3 deemed appropriate)	Quantitative: assessment/judgement of university-based and school-based teacher educators
18	Construct validity	Second-order confirmatory factor analysis and structural equation modelling	Quantitative: schoolteachers' self-rating
20	Face validity	Percentage	Quantitative: assessment/judgement of candidates
22	Consequential validity	Thematic analysis	Qualitative: reflective commentary of candidates
27	Construct validity	Q-sort procedure	Quantitative: assessment/judgement of subject matter experts through a survey

30	Consequential validity	Thematic analysis and frequency	Mixed: assessment/judgement through questionnaires, focus groups, rating scores
31	Construct validity	Thematic analysis	Qualitative: feedback form regarding experience of scoring (this revealed misinterpretation of the tool construct)
33	Predictive validity	N/A	Non-empirical
	Consequential validity	N/A	Non-empirical
37	Predictive validity	Correlation coefficients and probabilities	Quantitative: results of pre-existing evaluation by employer principals, administrative records for graduated teachers' SAT scores and GPA
	Content validity	Mean scores, standard deviation	Mixed: assessment/judgement of in-service teachers
38	Construct validity	Confirmatory factor analyses together with chi-square goodness-of-fit and descriptive fit indices (comparative fit index, root-mean-square error of approximation, standardized root-mean-square residual)	Quantitative: self-rating of teacher candidates and a survey (to collect data for four variables related to teacher quality)
43	Face validity	Percentages, Pearson product-moment correlation coefficient, multiple regression analysis	Mixed*: assessment via survey (with open-ended questions)

Note. * Quantitative data collection via questionnaire with some qualitative data provided via an open-ended comment section.

3.3.4.1 Construct Validity

Construct validity involves *accurately* defining a 'construct' and *fairly* operationalizing it. Accurate definitions are supported by expert opinion, comparisons with tests that use a similar construct, exhaustive literature reviews, and grounding of the construct through relevant theory. Fair operationalization requires agreement on how the construct is measured, ensuring that instruments distinguish it from other constructs and capture only the intended construct (Cohen et al., 2018). Several studies examined the ways in which judgements of teaching reflected real situations and if instruments fully reflected what they aimed to measure (Studies 1, 3, 7, 11, 17, 18, 27, 31). Factor analysis was used in several studies to establish construct validity (Studies 1, 3, 7, 11, 18, 26, 38). Study 7 used confirmatory factor analysis specifically, and Study 11 used exploratory factor analysis specifically. Study 3 used both exploratory factor analysis and confirmatory factor analysis. Study 1 and Study 18 used a combination of confirmatory factor analysis and structural equation modelling, known as exploratory structural equation modelling, to assess the construct validity of instruments. Study 38 combined confirmatory factor analysis with chi-square goodness-of-fit and descriptive fit indices (comparative fit index, root-mean-square error of approximation, and standardized root-mean-square residual) in tool validation. Study 27 took a distinct approach, employing a Q-sort procedure, where participants categorize content based on relevance to the construct (in this case, standards) to identify items that best fit the intended focus of the tool. In Study 17, classical test theory was used, with authors noting this was one of the most commonly used tests in the field; this study also used the Rasch model to explore construct validity.

Several studies reported compromised construct validity, revealed through qualitative approaches and thematic data analysis, and inconsistent applications and misunderstandings of instrument constructs (Studies 1, 24, 31). In Study 31, raters misinterpreted the tool construct (i.e., Bloom's taxonomy) due to a lack of shared understanding; as their ratings did not reflect the intended constructs, the validity of the evaluation was compromised. Study 1 suggested that enhancing validity of an instrument would be possible through instrument revision that tackles potential differences of understanding among developers and users of the instrument. This was put forward in addition to rater training (i.e., to promote common understanding of the instrument construct) and curriculum alignment with the instrument.

3.3.4.2 Content Validity

Content validity requires that the defined content is representative of the entire scope of the construct, covering all relevant aspects of the defined content area (Cohen et al., 2018). Our examination of the content of the identified authentic candidate evaluation tools (based on 10 tools), as elaborated on in Section 3.3.2.3, showed that they are underpinned by various sources, such as evidence (n = 8), standards (n = 7), pre-existing evaluation tools (n = 4), and institutional conceptual framework (n = 1). Some tools were underpinned by a combination of sources; as shown in Table 3.16, 6 authentic candidate evaluation tools were grounded on both standards and evidence. The origin of standards was not explicit in the examined tools except for STP, which was developed in Chile with the involvement of 17 universities in collaboration with the Ministry of Education.

Though use of standards is often praised and provides a framework for assessing how far the tool captures the right competencies, some scholars criticize the way standards are integrated

in evaluation tools (Studies 32, 40). It has been argued that rather than being a way to enhance validity, the use of standards is more to do with the 'standards movement' (Study 32). It has also been argued that not all standards are well established (i.e., they are narrow, not research grounded, unclear, and not relevant to the context), so they cannot be used to provide predictive validation (Studies 32, 40). Standards are generally state-centric and used to provide assurance to the general public and other invested parties (i.e., accreditation agencies), but leave out professionalism as a concept (Study 40). It has been suggested that teacher educators should take an active role in developing institutional standards and/or customizing existing ones in order to provide a more representative grounding when it comes to developing assessment tools (Study 32). Importantly, in our examination of 11 evaluation tools, no instances were identified of institutional standards being incorporated in tool creation, though there was one instance where a tool was created based on an institutional conceptual framework (Disposition Assessment*). Where some form of standard was employed, it was professional standards set out by associations (n = 4), state and/or national standards (n = 2), or combination of both (n = 1). Notably, evidence used to establish validity came from internally generated evidence, such as a needs analysis (n = 2), or from prior literature (n = 2; i.e., academic research, theoretical frameworks, literature published by professional associations) or a combination of both (n = 4).

In relation to content, such standards traditionally encompassed two fundamental types of knowledge essential for teachers: subject matter knowledge and pedagogical content knowledge (Studies 27, 32). Study 32 argued for prioritizing pedagogical knowledge over subject matter knowledge, but Study 38 argued only 'general pedagogical knowledge' could help us to understand effective teaching (Study 38).

Several methods were used to make sure evaluation tools truly captured what they were designed to assess (i.e., that they had content validity). Studies 7 and 14 calculated the content validity ratio and content validity index, respectively, to gauge how well experts agreed on the relevance and representativeness of the content of tools for measuring the intended concept (e.g., teacher effectiveness). Study 38 incorporated feedback from practising teachers to validate the content of a tool. They used average scores and standard deviations to understand how relevant teachers found the elements of the tool. Statistical analysis of agreement was also used. For instance, Study 17 used descriptive statistics like percentages to analyse content validity (i.e., where two out of three experts agreed on the appropriateness of including a particular item, that showed content validity). Study 3 used Aiken's V statistic with a lower confidence limit to assess content validity.

3.3.4.3 Face Validity

A number of studies addressed face validity (Studies 4, 8, 10, 14, 16, 17, 20, 43), delving into the perceived suitability and effectiveness of the instrument (Cohen et al., 2018). Several studies revealed a notable sense of dissatisfaction and concern with evaluations (Studies 4, 8, 10, 16), indicating low confidence of teacher candidates and teacher educators to engage in evaluation processes actively and sustainably (Studies 4, 8). This was predominantly attributed to the perception of lack of validity and reliability of the evaluation tool (Studies 4, 8, 10, 17, 41), leading to recommendations for the adoption of empirically validated tools (Studies 8, 31). Low confidence and engagement with evaluation measures was also attributed to tools being cultural insensitive (Studies 4, 33), high-stakes consequences linked

to results (Studies 4, 12), and unclear and impractical evaluation tools (Study 8). In one study, this also led school administrators to move away from standards-based indicators to identify their own criteria for evaluation tools (Study 23). Study 4 recommended creation of collaborative networks and ongoing professional development activities to enhance understanding of evaluation tools, address concerns, and encourage active participation.

Qualitative approaches using interviews and focus groups (Studies 4, 8) were used to gather educators' and other stakeholders' views on the tool's relevance. Several studies (Studies 8, 10, 16, 17, 20) used surveys to gather feedback on the tool's clarity and satisfaction with the tool. Descriptive statistics like means, standard deviations, and weighted means were used to analyse responses. One study (Study 10) used advanced statistical tests like *t* tests and analysis of variance to determine face validity. This involved comparing the perceptions of different groups on the tool's appropriateness. Study 14 employed Cohen's kappa to assess agreement among raters on how well the tool reflects the intended construct. Study 43 used a combination of percentages, correlations (Pearson product-moment correlation coefficient), and regression analysis to explore the relationships between different aspects of the tool and its perceived face validity.

3.3.4.4 Predictive Validity

Only a handful (4 out of 45) of the studies explored whether evaluation tools could accurately predict teachers' future effectiveness – i.e., predictive validity (see Cohen et al., 2018). Three were empirical studies (Studies 6, 12, 37), while one was non-empirical (Study 33). The empirical studies employed various advanced statistical methods. Study 12 used models like logit, ordinary least squares, and stacked models to analyse the relationships between evaluation scores and future teacher performance. Study 37 used correlation coefficients and probabilities to assess predictive power. Study 6 used a combination of descriptive statistics and multilevel analyses to examine correlations, providing a more comprehensive picture.

Findings revealed insufficient evidence pertaining to tools' ability to predict subsequent teaching success. Certain measures and indicators, some based on the time prior to entering teacher education and some based on the time during teacher preparation, were found to be valuable in predicting the future teacher effectiveness (Studies 6, 33, 37). Examination of the predictive value of certain measures – such as personal traits and academic ability – for later teaching effectiveness did not provide concrete conclusions, and there were even some conflicting findings. For instance, in Study 37 academic ability did not significantly predict future teaching performance in, but in Study 6 it did have a significant correlation with efficient classroom management as part of instructional quality.

From a critical standpoint, the differences in these studies in terms of the research methodology, the contextual factors, and the specific aspect being measured as a proxy for *effective teaching* (such as instructional quality rated by school students, observational performance rated by employment supervisors, or value-added student achievement) may explain the variations in predictive value of academic ability. However, the fact remains that the studies included in the review provide inconclusive evidence for a link between academic ability and teaching effectiveness. This lack of established predictive link points to the need for careful consideration of admission criteria, reconsidering the use of disposition and personal traits other than 'agreeableness', so as not to overlook individuals who are genuinely

interested in teaching and to recognize the transformative role of TEPs in nurturing their potential (Study 37).

Some evaluation scores in certain subjects (i.e., reading edTPA) prevented ineffective teachers from entering the workforce (Studies 2 and 24), while others failed to predict teaching success in subjects such as mathematics (Study 12) and arts subjects (Study 33). Research regarding edTPA, the most frequently used candidate assessment tool identified in the review, has yet to establish predictive validity (Study 33). This is despite extensive nationwide use over two decades in the US (Studies 29, 33). One study suggested holding off on its high-stakes use tied to teacher licensure until sufficient evidence is available on predictive value (Study 33). However, a study in the German context showed some promise in terms of prediction of future teaching effectiveness. The second state exam in Germany, which assesses candidates' 'procedural knowledge', was found to be a strong predictor of future instructional quality, but the first state exam, which assesses factual/declarative knowledge, did not have predictive value (Study 6). However, despite the predictive validity of procedural knowledge, several studies noted a persistent gap in successful translation of theoretical knowledge into practical application (Studies 5, 6, 17, 30). .

In conclusion, to inform admission to programmes and to the profession, further research is necessary to determine the extent to which evaluation results, and which specific indicators, can reliably predict teaching success. This is crucial given the frequent use of evaluations as a gateway into TEPs (Study 37) and the labour market (Studies 12, 33), as well as their evidenced negative impact on the diversity of the teacher workforce (Study 12). Therefore, careful consideration is warranted in selection of indicators (Study 37) and incorporation of specific characteristics in targeted preparation programmes (Studies 6 and 37). To provide transparency, we have organized the current evidence in Tables 3.25 and 3.26.

Table 3.25

Studies Focused on the Predictive Value of Evaluation Results and Indicators of Effective Teaching

Study	Independent variable	Dependent variable	Study context and sample	Findings
6	Candidate's characteristics prior to entering teacher education – measured by their cognitive abilities ('intelligence'), academic performance ('high school GPA'), and five personality traits (i.e., neuroticism, agreeableness, extraversion, conscientiousness, and openness to new experiences) Candidate's performance on state examinations during teacher education – on Exam 1 (theoretical knowledge) and Exam 2 (procedural knowledge), which are integral for admission to the teaching profession in Germany	Candidate's later instructional quality, rated by classroom students; defined according to four dimensions – cognitive activation, classroom management, social support, instructional tempo	Students' (n = 3,768) ratings of teachers (n = 113), Germany	Among personal traits, 'agreeableness' – characterized by being altruistic, sympathetic, trustworthy, and nurturing – was the only predictive indicator for future instructional quality (i.e., creation of a supportive social environment). But other traits such as 'conscientiousness' – characterized by being organized, punctual, goal-oriented, and honest – were irrelevant. Candidates' 'intelligence' (i.e., cognitive abilities) was not a significant predictor of future instructional quality. High school GPA results were a strong predictor of teachers' instructional quality, particularly in terms of efficient classroom management ($\beta = .25$, $p = .023$). Exam results during teacher education were found to be a strong predictor of future instructional quality when the exam measures candidates' 'procedural knowledge' but not predictive the exam measures factual/declarative knowledge.
12	edTPA (Educative Teacher Performance Assessment) scores	Candidate's subsequent entrance into the workforce Candidate's later effectiveness, measured by students' reading and mathematics value-added measures	Teachers' (n = 2,362) edTPA scores and students' (n = 277) value-added measures, Washington, US	edTPA scores were a valid predictor of teacher workforce entry, with the correlation increasing notably after edTPA scores became consequential in Washington. edTPA scores in certain subjects (i.e., reading edTPA) were shown to prevent ineffective teachers from entering the workforce but failed to predict teaching success in subjects such as mathematics.

37	National Council on Teacher Quality's suggested selection standards, including SAT scores and GPA	Teacher's performance and preparation – rated by employment supervisor	Graduates (n = 1,723), California, US	There was no prediction of high school GPA and SAT on new teachers' performance.

Table 3.26

Predictive Value of Evaluation Results and Indicators of Effective Teaching

Evidence of prediction	Independent variable	Findings
No evidence of prediction	Teachers' intelligence/ cognitive ability prior to teacher education	Teachers' cognitive ability had no predictive value for instructional quality, based on classroom students' (n = 3,768) rating of teachers (n = 113) in Germany (Study 6)
	Teachers' academic ability prior to teacher education	Teachers' high school GPA had a significant predictive value for instructional quality,, particularly in terms of efficient classroom management (β = .25, p = .023), based on classroom students' (n = 3,768) rating of teachers (n = 113) in Germany (Study 6)
		New teachers' high school GPA and SAT scores had no predictive value for their teaching performance,, based on supervisor principals' rating of graduates (n = 1,723) in California (Study 37)
Mixed evidence of prediction	Teachers' personal traits prior to teacher education	Teachers' conscientiousness (i.e., being organised, punctual, goal oriented, and honest) had no predictive value for instructional quality,, based on classroom students' (n = 3,768) rating of teachers (n = 113) in Germany (Study 6)
		Teachers' higher levels of agreeableness (i.e., being altruistic, sympathetic, trustworthy, and nurturing) prior to entry to the programme had a significant positive predictive value for instructional quality, defined as creating a supportive social environment in which students feel secure and valued, based on classroom students' (n = 3,768) rating of teachers (n = 113) in Germany (Study 6)
	Teachers' scores from gatekeeping exams/ evaluations	Teachers' performance in first exam, measuring factual/declarative knowledge, had no predictive value for future instructional quality; but teachers' performance in second exam, measuring 'procedural knowledge', did have a significant predictive value for future instructional quality, based on classroom students' (n = 3,768) rating of teachers (n = 113) in Germany (Study 6)
		Teachers' passing edTPA scores had significant predictive value for their students' academic performance in reading (n = 204 teachers) but not in mathematics (n = 206 teachers) (Study 12)
Evidence of significant prediction	Candidates' portfolio scores	Candidates' edTPA scores had significant predictive value for workforce entrance, based on rating of candidates by outsourced evaluators (n = 2,362) in Washington

3.3.4.5 Consequential Validity

Several studies addressed consequential validity, which means the inferences made from an evaluation are sound (Cohen et al., 2018). These studies examined how evaluation practices affect both the assessment process and the candidates (Studies 4, 5, 12, 13, 15, 22, 30 and 33). Some studies found that certain evaluation practices contributed to candidates' growth and learning. For instance, involving candidates in their own evaluation process (Study 15) and requiring them to write reflective commentaries (Study 30) helped them become more self-reflective practitioners (Studies 15 and 30). Effective feedback, combined with tailored support based on evaluation results, also significantly enhanced candidate growth (Study 13). Study 30 found that introducing STP expanded supervisors' discussions with student teachers and adjusted their supervisory approach. In this context, rubrics accessible to student teachers were found to be important for transparency and for a common understanding of what quality teaching means.

However, some studies also identified issues compromising consequential validity. For example, Studies 12, 15, and 22 pointed to the consequences of low diversity in the workforce, which compromised the consequential validity of evaluations. Study 12 found that Hispanic candidates in Washington were more than 3 times more likely to fail the edTPA. However, White, middle-class candidates performed poorly in areas related to diversity and culturally responsive teaching, indicating challenges in terms of connecting with their students from different backgrounds (Study 15). Additionally, using evaluation as a one-time gatekeeper was found to possibly screen out candidates who could become effective teachers (Study 12). Study 12 evidenced some edTPA takers (8% of reading teachers and 14% of mathematics teachers) failing their first attempt but passing their second attempt to be placed in the high-performing teacher category (the top 20% of value added).

In an empirical study of the edTPA's consequential validity (Study 22), researchers concluded that there was no sufficient evidence supporting the consequential validity of edTPA as an assessment during student teaching, especially in social justice-oriented programmes. Participants, especially students of colour or first-generation college students, reported that the high-stakes, standardized format and external scoring of the edTPA had negative effects for them personally in the form of mental stress and financial burden. edTPA was also claimed to be encouraging inequitable practices, such as focusing on high-achieving classes and selecting curricula based on scoring criteria rather than student needs.

In Study 4, researchers suggested consequential validity could be compromised through a shift in purpose from evaluating constructs of effective teaching for teacher candidate growth and learning to evaluation for gatekeeping purposes. This leads to a shift where the evaluation itself becomes the focus of instruction (Study 33), resulting in a 'teach to the test' approach (Studies 4, 33) that could weaken validity. High-stakes decisions could also cause ratees to take actions to superficially improve their performance (Studies 24, 33), thereby diminishing educative engagement (Study 22). Study 33 found that in this context, candidates deviated from authentic, student-centred lesson planning – e.g., music student teachers tailored their lessons solely to meet the edTPA prompts. Study 1 argued that educative use of evaluation is not the responsibility of teacher educators alone; rather, this should be based on collaborative efforts by teacher educators, policymakers, and department and college administrators.

Studies addressing consequential validity mainly used qualitative methods, particularly thematic analysis (Studies 4, 5, 13, 15, 22), and addressed the impact of evaluation tools by identifying recurring themes in the data. In contrast, Study 12 used a quantitative approach, employing statistical techniques to examine the ethnic distribution of edTPA passers, with a two-sample *t* test to assess consequential validity. Study 30 employed a mixed methods approach, collecting data via a questionnaire and focus groups. Study 33 stood out by not relying on empirical data; instead, it examined consequential validity based on the author's scholarship and personal experiences.

3.4 Discussion of Key Findings From Selected Studies

Our systematic review was conducted using the PRISMA framework (Page et al., 2021) while also following established systematic review methods (Bryman, 2016). Searches were conducted in January 2023 across 19 databases relevant to or focused on teaching (Table 3.1). The criteria for selection were peer-reviewed articles, written in English or Welsh, published after 2010, and including aspects of judgement-making on teacher effectiveness (Table 3.3). The key search strings we used are provided in Table 3.2.

From this search, a total of 632 studies were identified. No research written in Welsh was found. Following the removal of duplicates and screening of abstracts and titles based on the inclusion and exclusion criteria, 46 studies were selected for detailed full-text examination. During this process, one study was excluded due to concerns about research quality, resulting in a total of 45 studies, which were summarized for data extraction (the framework we used to produce summaries is provided in Table 3.4).

The summaries underwent thematic analysis following Braun and Clarke's (2006) six-step framework. Recurring themes were identified inductively, iteratively, and collaboratively by the PI and RA (Table 3.5Table 3.), and these were subsequently added to the summaries of each study (Table 3.16).

3.4.1 Context and Themes

Three main themes of the studies included in the review were identified as: validity; reliability; and instrument development and implementation (Table 3.7). Over half the studies were carried out in US states. No research was identified from the UK home nations of Scotland, Wales, England, and Northern Ireland (Table 3.8). Where the study involved teacher candidates, this took place exclusively within the context of university-based TPPs. Most research was conducted within a single university context. Only one study involved multiple TEPs from different countries (Table 3.9). Quantitative and mixed methods were the most prominent methodologies. The vast majority of studies were empirically driven – i.e., relying systematic collection and analysis of data – and only five studies were classified as non-empirical – i.e., with no systematic data collection. Most empirical studies drew on primary data, followed by secondary data and a mix of primary and secondary data (Table 3.10). Studies that gathered original data predominantly used evaluation instruments and questionnaires, while those that used secondary data drew on pre-existing evaluation data (Table 3.11). In almost half of the empirical studies, participants were university teacher educators (Table 3.12). Of 45 studies, 27 occurred in the context of candidate evaluation during teacher preparation (Table 3.13Table 3.).

3.4.2 Evaluation Instruments and Processes Involved in Judgement-Making

Eleven *authentic* candidate evaluation instruments were identified, and these were examined in terms of development, design, and implementation (see Table 3.14). All but one instrument (STP) was created and used in the US. PACT paved the way for edTPA, the dominant teacher evaluation tool. New instruments like PEI aim to address edTPA's weaknesses, such as the practice of outsourcing raters. A mix of institutions created the tools, including single universities or a consortia of universities, research centres, state departments, and independent researchers. Five of these tools were developed by TEPs. Others were adopted directly (n = 4) or modified (n = 1) by TEPs (Table 3.15). The tools were grounded on various sources: evidence (i.e., prior literature, original data collection; n = 8); teaching standards (i.e., state/national standards, professional standards; n = 7); existing tools (n = 4); and institutional frameworks (n = 1). Some were based on a combination of these (Table 3.16). Most of the tools (n = 8) evaluated teaching performance and competence, and some evaluated dispositions (n = 3). Some included elements of dispositions alongside effectiveness (e.g., Studies 1, 30). However, concerns were raised about using dispositions in candidate assessment. (One *emerging* tool – SOCME-10 – goes beyond the traditional focus on effectiveness or dispositions to consider sustainable social development; Study 3). In the authentic instruments, rubrics were dominant; these provided detailed guidance for raters, including explanations (e.g., in the form of narratives). This was followed by rating scales, which were simpler, with limited descriptions. The most common format was a 4-point scale, while the least common were dichotomous (e.g., yes/no) and 7-point scales (n = 1 each; Table 3.18).

We examined the dimensions of *candidate* evaluation instruments according to the three domains of the UNESCO *Global Framework*. This showed a strong focus on teaching practices (i.e., planning, instruction, and assessment). In contrast, knowledge and understanding, which involved personal qualities and relationships, featured less prominently (Table 3.19). Evaluation results were predominantly used for or contributed to *summative decision-making*; use of results to support growth of teacher candidates was rare. Most of the evaluation tools (5 out of 9) were used to inform summative decisions. Of these five, three also contributed to *formative evaluation*, but this was aimed at providing one-time formative feedback to facilitate candidate improvement, not necessarily ongoing feedback and support for growth. Only one of these three provided support for all candidates. The other two provided support conditionally, targeting students who had failed (PACT) or had fallen below a specific threshold (PEI). Only three tools were used to support candidates' growth with *progress-oriented formative feedback*; this included monitoring and tailored interventions. Two of the three focused on dispositional growth (Disposition Assessment,* PDQ), and one focused on teaching effectiveness growth (CLASS: Toddler version).

Evaluation was conducted most often during school-based experiences (in 7 out of 10 tools). The most common evaluation methods were observation with self-assessment with observation or portfolio/work samples. Observation only and peer assessment were less prevalent. University-based teacher educators were the most common source of ratings (n = 7). Students (n = 4) and school-based teacher educators (n = 3) were less frequently involved in rating. One case relied on peer ratings (CLASS: Toddler version). Provision of training to raters was inconsistent. When training was available, this targeted candidates (n = 2), school-based teacher educators (n = 1), and individual raters who needed certification

(PACT, edTPA). Notably, training for teacher educators was not found in these studies. Four studies did not mention training, possibly indicating the lack of it. One study indicated that no training was provided.

Measures to improve reliability of rating (i.e., accuracy, IRR) involved various strategies:

- **double scoring:** Some studies aimed to improve quality by having a second rater review a sample of existing evaluations (e.g., PACT portfolios).
- **additional scorers for borderline cases:** Tools like edTPA and PACT used extra raters where candidates came close to a pass or fail.
- **combining ratings:** Some tools (e.g., PEI, Competence Assessment*) used a combination of scores from multiple raters, either summing or averaging the scores.
- **dispute resolution:** In cases of significant differences in ratings, some tools (e.g., PEI) incorporated discussions between raters, teacher candidates, and educators to reach a consensus. Tools like the Disposition Assessment* in Study 25 referred cases to a committee to analyse discrepancies in assessments.
- **candidate feedback:** One study (CLASS: Toddler version) gathered feedback from candidates about their assessment experience to identify potential issues.

3.4.3 Establishing Reliability of Judgements

Our examination revealed a number of findings related to the nature of reliability of judgements of teaching effectiveness. These focused on four key areas: internal consistency reliability; IRR; influences on rater reliability; and proposed ways to improve reliability. Findings on internal consistency reliability showed that consistency and accuracy of assessments tended to be more prevalent in holistic scoring than analytic scoring across raters and time. Cronbach's alpha was the most frequently used in testing of internal consistency. One study favoured that 'modern psychometric techniques', such as Rasch analysis, over 'classical test theory', such as Cronbach's Alpha (Study 17).

Findings relevant to IRR revealed that while some studies confirmed instances of inter-rater agreement and consistency, others found inconsistency and disagreement between raters. Two notable patterns emerged in these studies. The first pattern was that candidates tended to rate themselves lower than their peers and school-based teacher educators rated them, yet in comparison to university-based teachers' ratings, their self-ratings were either similar or lower. The second pattern was that school-based teacher educators' ratings were almost always higher than both teacher candidates' and university-based teacher educators' ratings. The majority of studies employed descriptive statistics, including exact ('assigned same score') or partial percentage agreement ('adjacent agreement', either the same or within a difference of one point), standard deviation, or comparison of raters' scores with an identified 'true score'. However, measures such as Cohen's kappa and intra-class correlations were recommended for accurate reporting and to account for chance agreement.

Five factors influencing rating reliability were identified: ratees; raters; tool characteristics; deployment of evaluation; and methods used to determine reliability and validity. The most common suggestions for improving reliability were standardization of sources of information, scoring, and criteria as well as training for raters. However, these would not guarantee that evaluators make objective and reliable judgements. Recommendations included creating more effective training materials, 'regular training' rather than one-off training, and including

a session on quality control for scoring. Yet some empirical studies concluded training would not be an effective solution, even with quality control sessions and more extensive training. Other recommendations to improve reliability included having multiple raters rather than a single rater, employing a variety of assessment methods, and assessing multiple times. Another suggestion was constructing objective indicators for assessment. It was also suggested that teacher candidates' active involvement in self-evaluation should be deliberated alongside school-based teacher educators' and teacher educators' ratings, thereby demonstrating triangulation.

3.4.4 Establishing Validity of Judgements

In terms of instrument validity, face, content and construct validity were the types most frequently considered, while predictive validity was considered least. Some types of validity were not understood properly. In terms of *construct validity*, several studies examined the ways in which judgements of teaching reflected real situations, whether they measured what they were intend to measure, and whether instruments fully represented what they aimed to measure. Factor analysis was used to establish construct validity. Some studies took a distinct approach, employing a Q-sort procedure (Study 27) or a Rasch model (Study 17) to explore construct validity. Further, compromised construct validity was indicated through qualitative approaches and thematic data analysis, which revealed inconsistent application and misunderstanding of instrument constructs.

For *content validity*, to ensure evaluation tools truly captured what they were designed to assess, several methods were used: content validity index; content validity ratio; and percentages of agreement. Examination of the content of *authentic* candidate evaluation tools showed that they were underpinned by various sources: evidence; pre-existing evaluation tools; an institutional conceptual framework; or some combination of these. Some studies criticized the integration of standards in evaluation tools, noting that not all standards were well established (i.e., they were narrow, not research grounded, unclear, and not relevant to the context).

The perceived suitability and effectiveness of the instruments – i.e., their *face validity* – revealed a notable sense of dissatisfaction and concern with evaluations, leading to low confidence of teacher candidates and teacher educators in evaluation and lack of active and sustainable engagement in the process of evaluation. Qualitative interviews and focus groups were used to gather educators' and other stakeholders' views and initial impressions of a tool's relevance. Several studies used surveys to gather feedback on a tool's clarity and the level of satisfaction with a tool. Both descriptive and advanced statistics were used to analyse these responses.

Only a handful of studies (4 out of 45) explored *predictive validity* and only three of these were based on empirical research. Various advanced statistical methods were used to calculate predictive validity: logit, ordinary least squares, and stacked models; correlation coefficients and probabilities, descriptive statistics; and multilevel analyses. Findings revealed insufficient evidence on the predictive value of candidate evaluations on subsequent teaching success. Certain measures and indicators, covering the period prior to entering teacher education and the period of teacher preparation, were found to be valuable in predicting future teaching effectiveness, while others were not. Examination of the predictive value of certain measures – i.e., personal traits, academic ability – on later teaching

effectiveness did not provide concrete conclusions, and there was even some conflicting evidence. Some evaluation scores in certain subjects (i.e., reading edTPA) prevented ineffective teachers from entering the workforce; others failed to predict teaching success in subjects such as mathematics and arts subjects. The review found that even for the most frequently used candidate assessment tool, edTPA, predictive validity had not yet been established. Further research is necessary to determine whether evaluation results, and which specific indicators, both for admission to programmes and the profession, can reliably predict teaching success.

In relation to *consequential validity*, certain evaluation practices were found to contribute to candidates' growth and learning: candidates being self-reflective practitioners; involving candidates in their own evaluation process; and requiring candidates to write reflective commentaries. Additionally, effective feedback, combined with tailored support based on evaluation results, also significantly enhanced candidate growth. However, some studies also identified issues compromising consequential validity: low diversity in workforce; using evaluation as a one-time gatekeeping function; shift from using evaluations to enhance teacher candidate growth and learning to using evaluations for gatekeeping. These all compromised the consequential validity of evaluations. Studies addressing consequential validity mainly used qualitative methods, particularly thematic analysis. Some studies used quantitative or mixed methods approaches, employing statistical techniques.

3.4.5 Considering Dependability

It is clear from the variety of ways researchers engage with questions of validity and reliability that there is a constant interplay between how these are approached and how they are perceived by those engaged in evaluating teaching effectiveness. For example, an evaluation tool with high content and construct validity, confirmed through advanced statistical modelling, may not be used with fidelity and thus may not yield reliable results. Additionally, IRR can be achieved regardless of the accuracy of a measure. Moreover, findings from this review support that compromised confidence in tools, process, or purpose can lead to more variability in judgement decisions. It could be argued that IRR may not be attainable, or perhaps even desirable, when judging teaching. While the same evaluation instrument may be used each time an observation of practice occurs, IRR is dependent on consistency in what is being measured. However, in teaching, the setting (i.e., classroom environment, learner complexity, different schools; Cooksey, 1996) and subjects always change. It is within these concepts that SJT emerges to guide consideration of the diverse settings in which competency is judged. Neither ecological validity (i.e., the connections between judgement criteria and the cues used to make judgements) nor cue utilization validity (i.e., the connection between the cues that are observed and the judges making decisions about student teachers) were evident in the included studies, except for Study 23.

The variability evidenced in these studies further reflects the complexity and uncertainty of the teaching endeavour and questions the desirability of standardization and high-level objectivity. Perhaps the findings on the low influence of training to improve IRR and limit potential rater bias support a rethink around how reliability and validity are determined. There is a need for a holistic and balanced judgement strategy that enables decision makers to consider various factors and does not overlook the professional judgement and personal insights of raters or the individuality of each student teacher. Perhaps moving from the

cannon of quantitative language to that of trustworthiness and dependability would be a more fitting way to judge a phenomenon that defies uniformity. In studies included in this review, it was put forward that multiple raters could help mitigate variance in understanding and implementation of evaluations (Study 24) and that multiple ratings could be concluded with a quantitative rating aggregation or interpretative qualitative approach. Ultimately, what is desired is a reassurance that decisions made about student teachers are as valid and reliable as possible; thus a broader consideration of how this is determined may be useful.

Interestingly, the creditability of collective component parts of the entire process of judging teaching effectiveness was not evident among the research examined. Findings overall indicate a needed alignment of evidence used to make to make judgements on teaching effectiveness, and there were two critical reasons (Haigh et al., 2013) for why these occur in the first place: to confirm that student teachers have the necessary personal qualities and relationships to assume independent responsibility for a classroom; and to confirm that they can plan, teach, and assess for pupil learning.

3.4.6 Challenges in Complexity

As Cooksey (1996) noted and this systematic review confirms, judgement-making appears to remain a best estimate of the right choice under specific constraints, which always runs the risk of error. Furthermore, simultaneity of influences from different levels prompt variability (Martin et al., 2019), and even small influences (e.g., how an evaluator grounds a judgement they observe) can have a cascading, consequential effect (e.g., whether or not a student teacher receives licensure). Even the simplest teacher decisions can have multiple causal pathways (Opfer & Pedder, 2011). The degree of ambiguity and variation with which decision makers are able to cope among an intertwined set of probabilistic relationships indeed varies from one setting, TEP, or education system to the next. What is considered important in investigating and establishing validity and reliability remains just as variable according to the literature. An interesting deliberation emerges to reconsider predictive validity and to continue to question whether TEPs should seek to guarantee particular outcomes (Cochran-Smith et al., 2014). There appears to be a continued need to mesh both professional standards and professional judgement when practices of student teachers are assessed, and to continue to illuminate what is or should be considered legitimate knowledge in the process of teacher education.

As Biesta (2020) observed, effectiveness is considered a process value, and effective ‘for what’ and ‘for whom’ should be a consideration of TEPs in the exploration of judging effectiveness. It may prove useful during this era of high accountability and increased empirical scrutiny to re-engage with educational purposes to better understand what is at stake for new teachers when judgements are made. To that end, Biesta’s (2015) three functions of *education*, *qualification*, *socialization*, and *subjectification* may prove applicable to navigating judgements of effectiveness made in teacher education. Qualification is the most dominant reason judgements of teaching effectiveness were made in this review (i.e., gatekeeping); however, this appears often to be at the expense of other purposes. A tension between high-stakes consequential outcomes of judgements and educative uses of evaluation for growth was revealed in the findings. TEPs may be challenged to consider if knowledge, skills, and dispositions to teach should be precise, confined, and measured analytically according to operationalized indicators or if these can be relatively broad, such as the holistic

ability to gracefully teach increasingly diverse learners. Socialization brings consideration to the ways teacher education attempts to make student teachers competent members of the profession and reproduce expected identities. Teacher educators are confronted to consider if orientation into existing traditions and standardized ways of doing is what is desired or if it is more necessary for new teachers to be transformative and for TEPs to review what could be reductive evaluation measures. Finally, Biesta (2020) reminds us that education itself always also impacts on the student teacher as an individual; thus, teacher education can serve to either enhance or restrict capacities and capabilities. TEPs may consider, therefore, in what ways evaluation processes are situated to capture important dispositional aspects of high-quality teaching, such as developing a sense of self and agency, as decisions about entering the profession are made.

3.5 Conclusion

This chapter detailed the methods and results from Phase 1 of the project, a systematic literature review seeking to better understand judgement-making on teaching effectiveness. It explored in depth the nature of judgement-making processes, how the classroom practice of pre-service teachers and in-service teachers is evaluated, the criteria and competencies used to judge teaching effectiveness, the validity and trustworthiness of evaluation instruments, the variation between rater groups, and the reliability of the judgement-making process. In Chapter 4, we investigate the development of professional standards in England, Scotland, and Wales, which have been used to define the competencies being judged in teacher evaluation and to validate the tools used in this process.

The findings of the systematic review contributed to the design of the questionnaire used in the case studies presented in Chapters 5, 6, and 7 and shaped the briefing questions for the Delphi Panel (see Chapter 8). The findings also shaped and the overall convergent cross-phase and case meta-analysis (see Chapter 9) by providing triangulation of data, in which the results of this project can be situated (see Chapters 10 and 11).

4 Professional Teaching Standards Policy Review

In Phase 2 of the research project, a comparative crosswalk analysis was conducted involving the professional standards for newly qualified teachers (NQTs) in three national jurisdictions of the UK – England, Scotland, and Wales. This comparison was necessary so that we might develop an understanding of the universal and particular aspects of standards informing judgements and evaluation tools used during school-based experiences where observations of teaching occur. Before and since the advent of devolution in Scotland and Wales in 1999, the three jurisdictions have each followed distinctive, and increasingly divergent, pathways in respect to education policy. It is often stated there is no 'British' education system, given these devolved educational powers; in what ways this has evolved needed to be explored in order to conduct collaborative research across the three nations. When working across national boundaries, there is a challenge in terms of whether standards really mean the same thing. There remains debate about how far these systems hold to common standards of what constitutes high-quality teaching and whether they are in fact still relatively similar.

This chapter details the ways in which the professional competencies for NQTs are articulated in each jurisdiction, comparatively analysing each nation's standards alongside the internationally recognized UNESCO (United Nations Educational, Scientific and Cultural Organization) global professional teaching standards (Education International & UNESCO, 2019). In 2019 the 8th World Congress of Education International passed a resolution supporting the implementation of the joint Education International and UNESCO framework on professional standards. The intention of these standards was to make clear what constitutes effective, ethical practice in the profession. A fifth set of standards, the Interstate Teacher Assessment and Support Consortium (InTASC) standards, used widely across the US, was used to further consider alignment and to enable the project team to expand reading of teacher educator judgement and allow for future scale-up opportunities, increased generalizability, and timely replication of the project in the field of comparative educational research.

Set within the theoretical framework of social judgement theory (Cooksey, 1996), which informed this project, the review of professional teacher standards guided conceptualization of the judgement problem in evaluating new teachers' effectiveness, as well as identifying the dimensions used to make said judgements.

The chapter begins with an investigation of background and contextual information regarding development and refinement of professional standards for teachers, which starts to reveal similarities and departures in processes and the discourse of standards setting in each national jurisdiction. Next, the chapter focuses on findings from the critical policy analysis and 'crosswalk' exercise involving comparison of the current standards in England, Scotland, and Wales anchored alongside the UNESCO global standards and the InTASC standards. The resulting crosswalk, the first comparison of its kind, puts forward novel insights into the devolved educational standards for teaching used to judge student teachers' performance. The chapter also explores the meaning and potential implications for teacher preparation, ongoing professional development, teaching practices, student outcomes, policy, and research.

4.1 Professional Standards for Teachers

A strategy across a number of nations to improve equity and quality in education has been articulation of professional teaching standards that specify what teachers should learn and be able to do (Darling-Hammond, 2017). Consideration of established teaching standards and criteria has been an integrated practice and component of teacher education systems, quality assurance, and accountability worldwide (National Academy of Education, 2024). Standardization, as Carter (2008) put it, is the process of legitimization, which has the power to elevate the profession. Prior literature explores the use and construction of professional teaching standards, their influence on teacher education, and associated criticisms and potential future directions within this field.

Professional standards have a number of related uses, including preparation of new teachers, recruitment and hiring of teachers, a pathway to or road map for accomplished teaching, guidance for experienced professionals, a structure for focusing improvement efforts, and communication with the wider community and education stakeholders (Danielson, 2007). Professional standards tend to serve three main functions (Council of Chief State School Officers [CCSSO], 2013), all essential to success: they can indicate a broad vision of where the profession is headed; they can provide a shared understanding of a specific 'bar' or level of performance and conduct that must be met; and they can articulate the supports necessary to ensure teachers have opportunities to meet the standards. As Charlotte Danielson (2007) pointed out, standards of professional practice are not unique to education and are well reflected in other professions (e.g., medicine, accounting, architecture). Definitions of expertise and procedures to qualify novice and advanced practitioners, Danielson noted, 'are the public's guarantee that the members of the profession hold themselves and their colleagues to high standards of practice' (p. 2).

Efforts dedicated to defining a knowledge base for teachers' knowledge, skills, and competencies have been ongoing for decades, particularly since the mid-1980s (Tigelaar & van Tartwijk, 2010). These efforts have translated into standards and criteria in the pursuit of teacher effectiveness, which serve as a foundation for teacher education curriculums, assessment, and quality assurance (Yinger & Daniel, 2010). One of the earliest examples was the introduction of the InTASC standards in the US in 1987 (Papanastasiou et al., 2012); these set out to define effective teaching for all learners and establish a progression towards sophisticated teaching practices (CCOSS, 2013). While professional standards vary greatly in detail and encompass a wide range of dimensions, they can be broadly categorized into three fundamental areas of focus: essential subject matter knowledge; pedagogical content knowledge; and professional values and dispositions. Effective teaching emerges from the synergy of these dimensions, as it hinges on imparting specific content (subject knowledge) through proficient instructional techniques (pedagogical knowledge) that are implemented through and underpinned by an overarching set of professional skills and attributes. Wyatt-Smith and Looney (2016) recognized professional standards as the 'codified representations of teachers' work' (p. 805).

Prior research has pointed out that accreditation bodies and many professional standards are government-centric and, at times, leave out teacher professionalism as a concept (Papanastasiou et al., 2012; Yinger & Daniel, 2010). Furthermore, standards that drive teacher education programmes (TEPs) exhibit diverse origins, ranging from institutional-level

constructions to national, state, and professional standards. Smalley and Retallick (2012), in a study of standards in a TEP, found that state (87%) and institutional standards (67%) were the most influential, followed by professional (47%) and national (43%) standards. Some TEPs do not have autonomy to select or customize standards, so they adapt mandated standards; others have the autonomy to select one or multiple sets of standards and customize them for their needs, though this is not a common approach. Papanastasiou et al. (2012) identified that some TEPs even create their own institutional-level standards in line with calls for needs-based standards and research-based evidence.

Professional standards are also used for assessing prospective teachers in TEPs. They can be used to guide selection of assessment criteria or to inform evaluation of prospective teachers' practice in simulated and real classroom settings (Tigelaar & van Tartwijk, 2010; Yinger & Daniel, 2010) by highlighting specific assets (i.e., skills, learning outcomes) that a teacher needs to demonstrate based on their preparation. Such standards have also influenced teacher education curricula and defined benchmarks for admission, licensure, and professional growth; therefore, standards expected of future teachers influence not only what they learn (Tillema, 2010) but also what they are taught (Tanguay, 2020). Standards are also frequently described as a guardian in achieving objectivity and consistency within the assessment of teacher candidates, as well as their use in making informed judgements about competence (Papanastasiou et al., 2012).

Prior research has examined the influence of teaching standards on teacher effectiveness. Studies conducted in the US, such as those that examined the influence of expectations set by a standards-based portfolio performance assessment (i.e., edTPA – Educative Teacher Performance Assessment), revealed that teacher educators recognize the significant potential influence of these expectations on the development and learning of novice teachers, particularly in high-stakes educational contexts (Tanguay, 2020). Tillema (2010) found that the presence of explicit standards were seen as a condition for successful self-assessment, as they often framed the difference between self-perceptions of attainment and externally set standards of competence.

Despite their advantages for candidates and programmes, professional teaching standards also face criticisms and challenges. Critics have argued that standardized assessment can induce mental and financial stress (Behizadeh & Neely, 2018) and may also narrow the curriculum and student learning, thus hindering learning opportunities (Tanguay, 2020). When imposed, a lack of consideration for programme values may also occur. What is more, high-stakes standardized assessments can shift the focus of instruction and the profession away from authentic, student-centred ways for future teachers to demonstrate their development and towards simply working for the test. This was observed in a study by Parkes and Powell (2015) with music education student teachers who tailored their lessons solely to meet standards-based assessment prompts. Papanastasiou et al.'s (2012, p. 306) study highlighted the potential for professional standards to both guide and constrain, and it problematized the standards movement, noting how quality of teacher preparation is assessed based on assumed criteria without rigorous evidence or validity. Validity has been queried particularly in relation to predictive and consequential validity when standards-based evaluations are used to assess new teachers' effectiveness (Anderson et al., 2024). While standards provide a framework for consistent evaluation, they can also impose limitations, as they may not align with the values and goals of all stakeholders. This prior research calls for a more nuanced

approach that considers the diverse educational landscapes and the needs of new teachers, better reflecting the landscape of fact.

4.2 Methods

The research design for this analysis drew on two established methodological frameworks, namely critical policy analysis and an exploratory crosswalk analysis to compare professional standards in England, Scotland, and Wales. Studies which have employed a crosswalk method to interrogate professional standards are diverse in terms of subject and scope. They fundamentally share the objective of identifying alignment, misalignment and/or discord between sets of standards or constructs with similar purposes. They include, for example, work on public health competencies (Woodhouse et al., 2010), nursing (Mahlmeister, 2015), and school-based mental health professionals (Zabek et al., 2023). In the field of education, the crosswalk is a well-established practice, often employed by public educational institutions and professional associations (particularly in the US) to map competencies and constructs across related domains of practice, to inform action (e.g., CCSSO, 2022; Commission on Accreditation of Athletic Training Education, 2020; Commission on Teacher Credentialing et al., 2020; Early Childhood Personnel Center, 2020). Yet, the crosswalk has not been widely used or methodologically codified as a comparative analysis tool in the field of educational research to date.

In line with the work of Diem and Young (2015) on critical policy analysis, the starting point for the analysis was to view the professional standards as 'constructions' – effectively, 'artefacts' of educational and policy ideologies, articulated at the point of practice (Morgan et al., 2024). As noted, the three UK nations examined in this study have plotted increasingly divergent policy paths, and this trend has accelerated since the advent of devolution. As such, a critical policy analysis lens was employed to examine *how* each of the jurisdictions articulated ostensibly similar, broadly cognate professional competencies, what inferences could be made about the underlying assumptions of the nature of teacher professionality, and how these articulations reflected the wider policy contexts within which they sat (Young & Diem, 2018). As a qualitative research technique, the process of comparison involved evaluating the standards documents to interpret them, gain an understanding of their meaning in context, and develop the information they provide.

This study's comparative analysis, therefore, happened in accordance with conventions set out in studies and the US policy tools referenced, and reflects a novel integration with critical policy analysis. The following steps were taken to carry out the standards crosswalk.

Step 1: Identify evaluators involved in the exercise. Experts were teacher educators working in teacher preparation in the constituent nations and members of the project team – one from England, one from Scotland (Principal Investigator – PI), and two from Wales.

Step 2: Assemble all relevant professional standards documents. The professional standards documents for new teachers were compiled by the PI in a password-protected shared digital project folder. These were:

- UNESCO: Global Framework of Professional Teaching Standards (Education International & UNESCO, 2019)
- InTASC: Core Teaching Standards and Learning Progressions for Teachers 1.0 (CCSSO, 2013)

- England: Teachers' Standards: Guidance for School Leaders, School Staff and Governing Bodies (Department for Education [DfE], 2011)
- Scotland: The Standard for Provisional Registration: Mandatory Requirements for Registration With the General Teaching Council for Scotland (General Teaching Council for Scotland [GTCS], 2021b)
- Wales: Professional Standards for Teaching and Leadership (Welsh Government, 2019)

Step 3: Create a template to populate the data. The PI created a crosswalk template with the UNESCO global standards and InTASC standards provided in two columns and empty columns for the standards from each of the constituent nations (see an example of the completed crosswalk in Figure 4.1 and the full crosswalk comparison in Appendix A4.1).

Step 4: Analyse and crosswalk the standards. Individually, evaluators from each nation populated their respective columns, mapping the national standards onto the equivalent UNESCO standards. Throughout the process, as evaluators searched for content and themes (Merriam & Tisdell, 2016), they noted deficiencies and gaps, unique wording or elements, and standards for which there was no clear alignment.

Step 5: Confirmation and audit of alignment. Following the initial alignment, the team members met in person to review the results, identify patterns, and deliberate on any standards for which there was not clear alignment. Critical discussions between the researchers led to development of a shared understanding of:

how each set of standards variously aligned, or misaligned, with the UNESCO 'benchmark' standards;
how the articulations of practice embedded in each set of standards reflected the divergent and unique policy ecology of each nation; and
whether or not there were significant gaps in any nation's standards when analysed against those of the UNESCO framework; or conversely whether there were any areas of practice articulated by any of the jurisdictions' standards which were not covered by the global standards.

Consideration of the language used in each of the standards was also factored into the overall analysis.

Step 6: Summarize results. The team members summarized the overall results and confirmed consensus implications.

The analysis was pragmatic in terms of its operational methodology, with an iterative, emergent approach employed for each phase (Hammersley, 2022). The UNESCO global teaching standards were used as an 'anchoring' benchmark set of standards for the crosswalk exercise, against which each nation's standards were aligned and interrogated. The three domains of knowledge, practice, and professional relations and the 10 corresponding standards in the UNESCO framework provided a system of concepts, assumptions, expectations, and beliefs that informed the overall enquiry (Maxwell, 2005) and facilitated cross-nation comparison.

Figure 4.1

Example From Standards Crosswalk Template

PROFESSIONAL TEACHING STANDARDS CROSSWALK				
UNESCO Global Framework	**SCOTLAND**	**ENGLAND**	**WALES**	**InTASC**
All Teachers	**Standards for Provisional Registration (SPR)**	**Teachers' Standards**	**Professional Standards for Teaching and Leadership (QTS)**	**All Teachers**
I. Teaching Knowledge & Understanding II. Teaching Practice III. Teaching Relations	1. Being a Teacher in Scotland 2. Professional Knowledge & Understanding 3. Professional Skills and Abilities	I. Teaching II. Personal & professional conduct	I. Pedagogy (P) II. Professional learning (PL) III. Collaboration (C) IV. Innovation (I) V. Leadership (L)	A. The Learner & Learning B. Content Knowledge C. Instructional Practices D. Professional Responsibilities
1. How students learn, and the particular learning, social, and development needs of their students (Domain 1)	*3.2.2 Engage learner participation* • value all learners and their participation, actively engaging children and young people in decision-making about their education • demonstrate care and commitment to working with every learner, embracing diversity to ensure that every learner feels welcome, included and ready to learn; • demonstrate knowledge and understanding of wellbeing indicators and childhood development; • recognise that childhood experiences impact on the learning and wellbeing of children and young people and actively respond in appropriate ways, seeking advice and collaborating as required; and • utilise strategies to nurture caring and supportive and purposeful relationships with learners and celebrate success	*2. Promote good progress and outcomes by pupils* • be accountable for pupils' attainment, progress and outcomes • be aware of pupils' capabilities and their prior knowledge, and plan teaching to build on these • guide pupils to reflect on the progress they have made and their emerging needs • demonstrate knowledge and understanding of how pupils learn and how this impacts on teaching • encourage pupils to take a responsible and conscientious attitude to their own work and study	P1. The teacher develops and demonstrates up-to-date theoretical knowledge and understanding as well as practical insight into how children and young people develop and learn. P4. The teacher demonstrates knowledge, understanding and experience of high expectations and effective practice in meeting the needs of all learners, whatever their different needs. P14. The teacher provides appropriate levels of challenge and expectations for the range of student abilities and characteristics, motivating learners to achieve.	Standard #1: Learner Development - The teacher understands how learners grow and develop, recognizing that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas, and designs and implements developmentally appropriate and challenging learning experiences. Standard #2: Learning Differences - The teacher uses understanding of individual differences and diverse cultures and communities to ensure inclusive learning environments that enable each learner to meet high standards.

4.3 Findings: Professional Teaching Standards in the Three Nations

National representations of what constitutes good teaching are shaped by particular policy and cultural contexts, and these are examined in this section. The analysis of professional standards for NQTs in England, Scotland, and Wales provides significant insights into the educational philosophies and priorities of each region and how these have developed over time. These insights underline the broader ideological differences that shape teacher preparation, professional development, and pedagogical practices.

4.3.1 Development of England's Standards

The professional standards which align with qualified teacher status (QTS; DfE, 2011) at the time of this research were introduced by the Conservative–Liberal Democrat coalition government formed in 2010. This makes them the longest-standing set of teaching standards since the first statutory teacher competencies were established in England in 1984. Following the general election in 2010, the outgoing Labour government's Department for Children, Schools and Families was reconfigured as the DfE and Michael Gove was appointed Secretary of State for Education. His stated intention was to improve the quality of teaching, and as part of his rhetoric he claimed that the existing criteria for teachers, by which he meant the qualification standards, lacked rigour (Spendlove, 2024). The revisions to the QTS standards formed part of a catalogue of changes – involving a mosaic of changes in terms of schools' policies – which impacted significantly on teacher education.

The evolution of standards for the teaching profession in England began in 1984, under Conservative Prime Minister Margaret Thatcher. In 1984 the first set of statutory teacher competencies was issued in Circular 3/84, followed by amendments in Circular 24/89 in 1989, and subsequent updates for new secondary teachers in Circulars 9/92 and 14/93 and for

new primary teachers in 1992 and 1993 as circulars for competencies presented as annexes 'appearing subordinate to the regulations' for teacher education (Smith, 2013, p. 430). It is interesting to note changes in terminology over this period. In the documents from the 1980s and 1992, the term 'student' was used with reference to student teachers completing their university or teaching college qualifications. This term was replaced with 'newly qualified teachers' in 1993. Both the change in language and the nature of the frequent updates to the competencies can be seen as 'consistent with the technical-rational approach to teacher education' (Ellis & Childs, 2023, p. 7) adopted during the 1990s, which identified specific skills and competencies required of new teachers.

The development of the competencies listed in the Circulars described above also reflects the progress towards and bringing into law of the Education Reform Act 1988, which made the National Curriculum and the associated assessments mandatory for all state schools in England. Thus, the competencies written in 1992 and 1993 relate to the requirements on new teachers for teaching and assessing pupils in line with the National Curriculum. At the same time, the regulations for teacher education were changing. Circular 24/89 directed a more school-based approach to teacher education. There was an enhanced requirement for both student teachers and their university-based lecturers to spend more time in school, and in addition staff in schools were expected to be involved in the planning, delivery, and assessment of teacher education. Circulars 9/92 and 14/93 reinforced the statutory nature of partnerships between schools and universities, with schools receiving money for training that had previously gone to universities. From 1992, initial teacher education (ITE) was also brought into the regulatory framework, through a schedule of inspections by the Office for Standards in Education, which brought new levels of state surveillance, scrutiny, and accountability into teacher education. The Education Act 1994 established the Teacher Training Agency, which had responsibilities for the provision and funding of teacher training in England and was charged with improving the careers information about teaching and the quality of routes into the teaching profession to support a raise in standards of teaching.

The election of a Labour government in 1997 happened alongside the transition from competencies to significantly more detailed 'standards' for NQTs. 'Although development of the first set of standards took place during the final stages of Conservative rule, they were finally published in July 1997, by which time Labour had been in power for almost two months' (Smith, 2013, p. 436).

Between 1997 and 2010, a swift and sweeping set of education policy initiatives were introduced by the Department for Education and Employment, which became the Department for Education and Skills and later the Department for Children, Schools and Families. The Teaching and Higher Education Act 1998 led to the establishment of the General Teaching Council for England (GTCE) in 2000 to support improvement of the quality of teaching and learning and become the regulator of teacher conduct, therefore holding responsibility for professional standards. In the Education Act 2005, the Teacher Training Agency was relaunched as the Training and Development Agency for Schools (TDA), which was directly accountable to Parliament. In line with Labour's schools policy, such as Every Child Matters, the TDA had an expanded remit with responsibility for improving the training and development of the entire school workforce. Many of these changes impacted directly on teacher education and the expectations placed on teachers by the state. New legislation, standards, and organizational infrastructure embedded the term 'initial teacher training' (ITT)

in place of ITE, and there was rapid growth in what was framed as the school-led ITT sector. Two new sets of standards were introduced in this era, in 2002 and 2007. In 2002, standards were categorized into three groups:

- professional values and practice
- knowledge and understanding
- teaching

A major change in 2007 was a newly differentiated model of teacher standards based on professional development and career stages. This meant that for the first time standards for trainee teachers (as they were then typically known) became the foundation for expectations of NQTs. The standards introduced a hierarchy of new descriptors: main scale, upper pay scale and advanced skills teachers. Despite recognizing the different career phases, this new document was more condensed than the 2002 version and was presented as a large colour poster showing career progression and related professional expectations. The new descriptors included references to reflective and reflexive practice, which Knight (2017) suggested were welcomed by ITE providers and teachers.

Following the 2010 election, under the Conservative–Liberal Democrat coalition, the newly designated DfE with Michael Gove as Secretary of State for Education undertook what it called the 'bonfire of the quangos', which led to a series of changes. In 2012 the Teaching Agency was established as an executive agency of the DfE, in place of the TDA and with some of the former GTCE roles (the GTCE was abolished). The Teaching Agency was thus responsible for ITT in England, as well as the regulation of the teaching profession. It was then merged with the National College for School Leadership to become the National College for Teaching and Leadership in 2013. A consequence of these changes included 'the loss of significant teacher education policy expertise and sector intelligence' (Spendlove, 2024, p. 48).

Amid these changes, the 2011 Teachers' Standards (DfE, 2011) were established, and these remain current at the time of this research. There are eight generic standards covering teachers qualifying to teach in primary and secondary sectors and all teachers in post. While offering a simplified document and a reduced set of standards (compared to the previous 102 separate standards), the generic nature of these standards is contentious. The same standards now apply to assessment of trainee teachers during and on completion of ITT, at the end of their 1 year as an NQT, and throughout their time in the profession.

Although the 2011 Teachers' Standards have not been altered, there has continued to be significant change in the sector. Despite the persistence of the QTS standards, it is noteworthy that DfE-designated academies and free schools can and do employ teachers without QTS (DfE, 2011). The majority of secondary schools (about 80%) are now academies – either stand-alone or within multi-academy trusts – as are almost 50% of primary schools, so the exclusion of QTS is not insignificant. A new Early Career Framework (DfE, 2019) became statutory in 2021 following pilot and early rollout phases. This meant that all new teachers were classed as early career teachers for 2 years (replacing the 1-year NQT status). The Early Career Framework sets out training content all new teachers are expected to master, and it is framed as a series of evidence statements, worded as 'learn that' and 'learn how to' statements, covering five core areas: behaviour management; pedagogy; curriculum; assessment; and professional behaviours. The framework is aligned with the

Teachers' Standards, which remain the benchmark for assessment of trainee teachers and early career teachers. Teachers in England are able to gain QTS through a wide range of ITT routes, including those offered by universities, school-based consortia, and new providers. This diverse ITT provision landscape was further consolidated following the DfE ITT accreditation process in 2022.

4.3.2 Development of Scotland's Standards

Gillies (2018, p. 108) importantly noted that Scottish education has never been integrated into a British system; it has remained separate even since the union of parliaments in 1707, a distinction which has been seen as a mark of national identity and pride. As Anderson (2018) stated, 'Scottish education has been characterized by a peculiar awareness of its own history' (p. 100). Since devolution and the opening of the Scottish Parliament in 1999, there has been a trajectory towards an increasingly outcomes-based approach and a move away from strategic issues to focus on operational matters and targets (Gillies, 2018). Teacher education in Scotland remains exclusively delivered by university providers in partnership with local authorities and schools. Fast-track or non-university-based models that have been adopted across a number of education systems have not been introduced more broadly.

Education policy is led by the Cabinet Secretary for Education and Skills, with the Scottish Parliament providing legislative oversight and scrutiny. The government's executive agency, Education Scotland, is charged with supporting both quality and improvement. Despite being directly accountable to government ministers, it is expected to operate independently and impartially (Education Scotland, 2023a). The GTCS is the teaching profession's independent registration and regulation body, responsible for teaching standards covering all stages of the professional continuum, from initial teacher preparation to principalship.

Teaching standards in Scotland were first established in 2000 (GTCS, n.d.b) and followed by a series of further standards across stages of a teacher's career (see Table 4.1). Together these form a framework continuum clarifying what it means to become, to be, and to flourish as a teacher in Scotland. The GTCS provides a side-by-side comparison of these standards, organized into two categories: benchmarks for teacher competence and what is termed 'aspirational standards' after full registration is attained (GTCS, 2021a). The standards framework is supported by the principles and values set out in the *Code of Professionalism and Conduct* (GTCS, 2012).

Since their formation, two reconceptualizations of the standards have occurred (GTCS, 2012, 2021). The first was set in motion in 2011 by *Teaching Scotland's Future* (Donaldson, 2011), referred to colloquially as the 'Donaldson Review', which marked a pivotal moment in Scottish education. The report was commissioned in response to a seeming lack of consistency in teaching quality across schools and local authorities, the observed variability in provision of mentoring and continued professional development, and a perceived compliance culture. Underpinning the review was the focus on teaching as a profession and teacher professionalism. Of the report's 50 recommendations, most can be directly or indirectly connected to teaching standards. A key recommendation was that the teacher standards framework should be reviewed in order to be 'explicit about the core knowledge, skills, and competencies that all teachers need to continually refresh and improve as they progress through their careers' (Donaldson, 2011, p. 97); Recommendations 35 and 36 specifically addressed professional standards as a strategic priority (Figure 4.2).

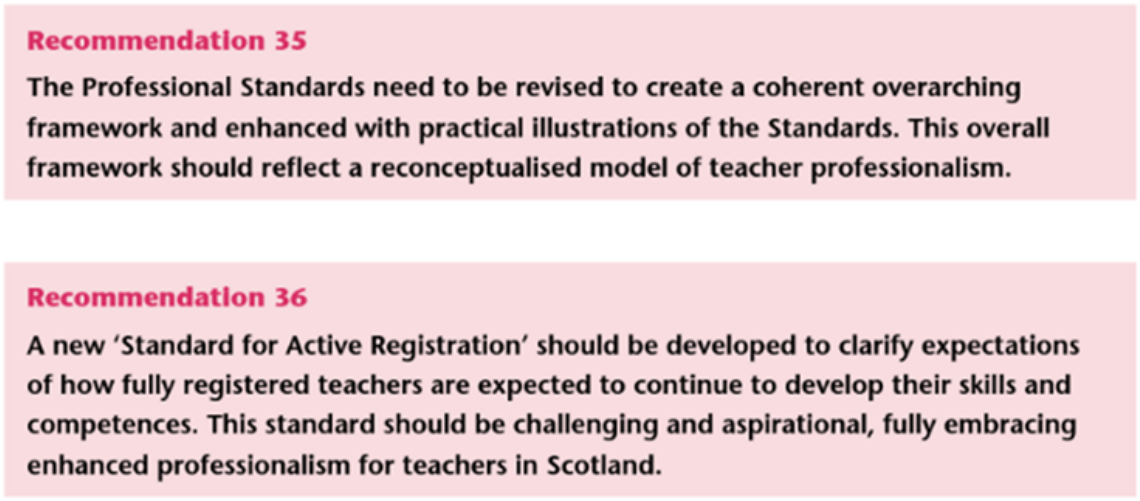
Table 4.1

Standards Across the Teacher Professional Continuum

Type of standard	GTCS standard	Career stage
Benchmarks for teacher competence	Standard for Provisional Registration	Initial teacher education
	Standard for Full Registration	Induction/probation (newly qualified teacher)
Aspirational standards	Standard for Career-Long Professional Learning	Post induction
	Standard for Middle Leadership and Management	Middle leader/head of department
	Standard for Headship	Principalship

Figure 4.2

Donaldson Review Recommendations Regarding Professional Standards



Note. From Donaldson (2011, p. 97).

A revised model in 2013, which followed on from the Donaldson Review (Donaldson, 2011, p. 26), called for clarity about the qualities and capacities of high-quality teachers. Also following on from the Donaldson Review, the National Improvement Framework was established in 2015 to evaluate how well schools meet national priorities (Education Scotland, 2023b). Drivers of improvement, reported annually, include school leadership, teacher professionalism, parental engagement (Education Scotland, 2018), assessment of children's progress, school improvement, and performance information. School- and national-level information on publicly funded schools is readily found on the online school information dashboard.

A ‘refreshed and restructured’ edition of the professional standards was enacted in August 2021 (GTCS, 2021b) following open consultation and evidence seeking from a range of stakeholders. A comparison of the 2012 and 2021 versions indicating key changes has been provided by GTCS (2023). This third version was informed by a literature review (McMahon, 2021), which suggested that Scotland’s overall approach to standards broadly aligns with similar approaches internationally. Implications further noted were the need for standards to be backed by research and, importantly, for the research base to be published as part of the standards document, the need for transparency in acknowledging contributors to standards development, and the need for careful consideration of the processes and pacing of implementation in professional practice (McMahon, 2021).

Along with recently revised professional standards, Scotland is experiencing a substantial reform agenda. This is evidenced in the myriad of recent independent, national, and international reviews, reports, and recommendations, including:

- *Scotland’s Curriculum for Excellence: Into the Future* (Organisation for Economic Co-operation and Development [OECD], 2021)
- Additional Support for Learning Action Plan: A Progress Report (Morgan Report; Morgan, 2021)
- Putting Learners at the Centre: Towards a Future Vision for Scottish Education (Muir Report; Muir, 2022)
- It’s Our Future: Independent Review of Qualifications and Assessment (Hayward Report; Hayward, 2023)
- All Learners in Scotland Matter: The National Discussion on Education: Final Report (Campbell & Harris, 2023)
- Fit for the Future: Developing a Post-School Learning System to Fuel Economic Transformation (Withers Report; Withers, 2023)

In light of these reports, a major restructuring of key agencies is underway, which will see a merging of the curriculum and assessment function of the Scottish Qualifications Authority (n.d.) and Education Scotland, and a separation of the development and support functions from the inspection function, for which Education Scotland has had responsibility (Muir, 2022).

Historically, progress in Scottish education has been marked by local autonomy in decision-making, a great deal of policy consultation, and transparent ways of working with interest groups and stakeholders (Keating, 2005). In a marked departure from this, on 15 October 2023, the then Cabinet Secretary for Education and Skills announced the formation of a Centre of Teaching Excellence, which could help make the country a ‘world leader in new approaches to learning and teaching’ as part of the afore-noted wider educational reforms (Scottish Government, 2023a). While details have been elusive, what it means to be an effective teacher clearly remains a continuing discussion in Scotland.

4.3.3 Development of Wales’ Standards

It is a reasonably well-established view that the educational landscape in Wales has seen three distinct phases of policymaking since the advent of devolution in 1999 and is by now well into what has been called the ‘third phase’ (Milton et al., 2020). Received accounts of the early years of devolution, the first phase, have traced a tendency towards an experimental

environment (Moon, 2012) where Wales was at the time keen to pursue a 'high trust' approach to education policy (Power, 2016), abolishing SATs and league tables and increasingly differentiating itself from its English neighbour's emphasis on choice and competition. Following disappointing Programme for International Student Assessment (known as PISA) results in 2009, characterized by the then Minister for Education as a 'wake-up call' (Andrews, 2011, 2014), the second phase of policymaking occurred from 2010 to around 2016; this involved the leveraging of increased external accountability back into the system, a renewed focus on literacy and numeracy, and the introduction of a categorization and then a banding system for schools, all with the ostensible aim of driving school improvement (Connolly et al., 2018). In the third phase, from around 2016, Wales has been engaged in a further ambitious and far-reaching process of reform.

This phase of policymaking, signalled by the publication in 2017 of *Education in Wales: Our National Mission* (Welsh Government, 2017b), has seen a shift away from the rhetoric of high accountability and the watchful emphasis on 'standards' back towards a narrative of trust, teacher autonomy, and re-professionalization. In a 2020 assessment of the Welsh standards in relation to the most recent reform process, the OECD concluded that 'Wales initiated a shift from what had become a managerial education system to one based on trust and professionalism' (OECD, 2020, p. 14). While such an extensive reform process is far from complete, there has indeed been a conscious and concerted effort in this direction, which has included reform of provision for pupils with additional learning needs, a review of qualifications, a refreshed professional learning offer, and, perhaps the centrepiece of this reform journey, a new curriculum. Following a review of the curriculum in 2015 (Donaldson, 2015), Wales has developed and is in the process of implementing the Curriculum for Wales, a purpose-driven, teacher-led curriculum that affords teachers high levels of autonomy and professional discretion (OECD, 2020). Wales has also made significant headway in decoupling pupil assessment from high-stakes public-facing measures of accountability via the new curriculum, and it has initiated the development of a new 'made-for-Wales' General Certificate of Secondary Education qualification to be implemented from 2025 (Qualifications Wales, 2023). The range of value-based changes in Wales across the recent years has influenced and shaped the development, structure, and content of the Welsh teacher standards.

4.4 Findings: Crosswalk Comparison

In addition to the analysis of the development of standards and changes, as part of the policy review, the professional standards themselves underwent a close investigation in this study. Analysis of standards for NQTs in England, Scotland, and Wales revealed meaningful insights into the educational philosophies and priorities of each home nation. It is important to reiterate from the policy review that while England and Wales have standards which apply to all teachers with an increasing degree of sophistication expected over time, in Scotland the standards for new teachers (including student teachers and first-year teachers) are distinct and separate from the standards for fully qualified teachers (see Table 4.1). These insights reflect broader ideological differences that shape teacher preparation, professional development, and pedagogical practices. Two key areas were explored: comparison of teaching standards in the three jurisdictions and their alignment with the UNESCO framework; and identification of agreement, gaps, or areas of overlap between national and international standards. (The full crosswalk analysis is provided in Appendix A4.1.)

4.4.1 Results of Crosswalk Comparison by UNESCO Domains

The UNESCO *Global Framework* is organized holistically into three domains, which are globally recognized by teachers as genuine, and 10 standards. The domains are: *teaching knowledge and understanding*, which has three standards; *teaching practice*, with four standards; and *teaching relations*, with the final three standards (see Figure 4.1 and Appendix A4.1). Analysis revealed that the three domains are evident across all three sets of national standards. Although classified and categorized using slightly different terminology and with varying depth and breadth, there is consistency in the sense that the overall domains are reflected in some way and the distinction between knowledge, skills, and professional dispositions is made. This was expected, even in the different policy contexts, given the global applicability of the anchoring framework.

The English standards exhibit only two domains, collapsing teaching knowledge and understanding and teaching practice into the broad category of 'teaching', with a focus on teachers' conduct instead of development and strength of relationships and with a national reference to 'not undermining fundamental British values' (DfE, 2011). The Scottish standards, with three nearly corresponding domains, align most closely with the global framework. The terminology of 'Being a teacher in Scotland' reveals singularities: the national focus recognizing Scotland as a distinct educational setting, with an inference that it is different from other places; and the qualification of 'teaching relations' by evoking specific values-centred language of social justice, trust, respect, and integrity to define the third domain. Additionally, professional commitment is indicated for language provision in the Gaelic medium. Interestingly, the Welsh standards are organized according to five domains, revealing the most distinct set of domains. While domains of pedagogy and collaboration align generally with the first two UNESCO domains, what is termed 'teaching relations' is differentiated according to the domains of professional learning, innovation, and leadership. This demonstrates a joining up of competencies across the career span of a teacher from induction to formal leader, as well as the underpinning assertion that development of the teaching profession can lead to transformation of the education system in Wales (Welsh Government, 2019, p. 4). The structure of the standards shows what competence at the next level may look like across the domains. In a similar manner to Scotland, in the case of Wales as overarching precursors to the detailed standards, but thereafter embedded throughout; values and dispositions are specifically referenced; the standards emphasize 'the central importance of the promotion of Welsh culture and language' (Welsh Government, 2019, p. 8). While Scotland and Wales both have distinct national elements in their overall domains, it is noteworthy that the English standards reference Britain instead of England. This linguistic choice indicates the combination of the three nations of the island of Great Britain, an interesting choice in the context of devolved educational powers. It is therefore in the third domain of 'teaching relations' that the greatest difference arises. The UNESCO framework states:

> Teaching is inherently constituted in relationships. As well as engaging with students, professional relationships with colleagues, parents, caregivers, and education authorities are crucial to effective teaching. Relations with the general community are also crucial to a teacher's work and to the profession as a whole. (Welsh Government, 2019, p. 5)

It appears to be the way teachers are considered as professionals and expected to engage in their professional work, with both the privileges and obligations conferred within the wider community, that emerges as a distinction in the devolved nations. The next part of the analysis, at the level of the 10 UNESCO standards, looks more specifically at the areas of collaboration, communication, and professional development, which comprise this domain of 'teaching relations' in the national standards.

4.4.2 Results of Crosswalk Comparison by UNESCO Standards

In addition to comparison with the three broad domains, analysis considered the alignment of the three sets of standards to the 10 UNESCO standards: learners; content, research; planning and preparation; instructional strategies; learning environment; assessment; collaboration; communication; and professional development (see Appendix A4.1). Overall, the broad standard areas are more alike than different across the three nations; however, many key differences emerged. While professional standards in Scotland and Wales align with all of the 10 global standards, England's standards have a significant gap in two areas: analysis revealed no professional standards for teachers in England in relation to UNESCO Standard 3: research, or Standard 10: professional development. The remaining standards for England could all be aligned. In addition, there are specific standards unique to Scotland and Wales which could not be aligned and thus mark a distinction. Unique to Scotland is Standard 1.1 professional values; this sits within the domain of 'being a teacher in Scotland' (GTCS, 2021b, pp. 4–5), which describes the professional standards that outline 'what it means to become, to be and to grow as a teacher in Scotland' (p. 4). Clearly articulated is the overarching commitment that 'Scotland's teachers help to embed sustainable and socially just practices in order to flourish as a nation' (GTCS, 2021b). This distinct focus on national flourishing and specific values is exclusive among the standards. In the Welsh standards, there is arguably a focus on more cultural, rather than civic, expressions of nationhood, with explicit commitments to Welsh culture and the Welsh language (Welsh Government, 2019). In relation to the standards for Wales, two standards do not align to the global framework:

> P15. The teacher demonstrates a willingness to seek, listen to, and take account of the views of learners in order to engage and encourage them as active participants in their own learning.
>
> P19. The teacher raises awareness of how high-quality learning experiences and performance outcomes lead to improved learning and a heightened sense of well-being. (Welsh Government, 2019, p. 37)

These standards reflect a distinctive Welsh policy ecology which has foregrounded wellbeing and learner voice though legislative and policy instruments, such as the Wellbeing of Future Generations Act (2015), and the provisions of The Curriculum and Assessment (Wales) Act (2021), which make the promotion of knowledge and understanding of children's rights compulsory. There are also specific nuances in the standards of each home nation. While these are evidenced in Appendix A4.1, a summary of the key findings is provided next for each set of professional teaching standards.

4.4.3 Features of England's Standards

The English standards are characterized by a directive tone, mandating specific professional practices. This approach aligns with a vision of the teacher as a practitioner who follows

sanctioned guidelines and procedures rather than as an autonomous professional making informed decisions. The emphasis on curriculum knowledge and behaviour management, with a focus on technical competencies over reflective and research-informed practices, further underscores this directive approach. The portrayal in English standards of the child as a passive subject of pedagogical practice contrasts with the more dynamic and context-sensitive approaches in Welsh and Scottish standards. While there is a nominal acknowledgment that pupils should be 'involved', the prominence of pupil voice is notably absent. This sort of passive view has been the subject of criticism from the sociology of childhood, which suggests that the education system may be more focused on shaping children into 'little adults' rather than recognizing and supporting their developmental phases and capacities. Additionally, the absence of a focus on research and continuous professional development in the English standards highlights a potential gap in relation to promoting a culture of ongoing learning and adaptation among teachers. This omission suggests a static view of professional competence, which may limit the ability of teachers to respond to evolving educational challenges and innovations.

4.4.4 Features of Scotland's Standards

Scottish standards are noted for their succinct and clear tone, which may aid their usability and implementation. However, this brevity raises questions about whether more complex details are embedded within the specific competence descriptors and connected to the main statements. Interestingly, the descriptors are marked as 'professional actions', and the preamble of each standard begins with 'you are required to' (GTCS, 2021b), which is language worthy of note given the overall greater degree of agency and professionalism indicated across the standards, even for novice teachers in their probationary year. Overlap of descriptors and competencies does occur numerous times for similar concepts; for instance, teaching practices related to 'digital technologies' are listed in three different standards. This can make it difficult to suss out the nuances of each standard and understand exactly what is expected of the new teacher. Like the Welsh standards, Scottish standards prioritize engagement with research as a critical component of professional competence, reflecting a commitment to evidence-based teaching practices. Distinctive aspects of Scotland's standards include the expectation of an enquiring stance, support of Gaelic language provision (GTCS, 2018), commitment to the United Nations Convention on the Rights of the Child (Education Scotland, 2023c), promotion of practitioner enquiry (GTCS, n.d.-c), and provision of play-based and outdoor learning, as well as the classing of Learning for Sustainability, beyond a responsibility and professional commitment, as a 'way of being' (Anderson & Tonner, 2023, p. 164). The most recent standards include a new section on professional values of social justice, trust, respect, and integrity, and they place more emphasis on the significance of professional learning. The recommendation for systemic support of and investment in mentoring and professional learning has been continually confirmed. Scottish standards, like their Welsh counterparts, underscore the centrality of professional learning to ensure teachers remain current and effective in their practices.

4.4.5 Features of Wales' Standards

The professional standards in Wales are complex and multifaceted, mirroring the intricate nature of educational practice. This complexity, while reflective of the diverse and context-dependent nature of teaching, raises questions about the usability and accessibility of these

standards for teachers. An example of this complexity can be found in the use of 'behaviours' (plural) in Welsh standards, which acknowledges the contextual nature of student behaviour. This contrasts with the more limited binary framing of behaviour in English standards, which tends to categorize behaviour as either positive or negative. A key aspect of the Welsh standards is the emphasis on research-informed pedagogy, which aligns with both UNESCO and Scottish standards. This focus underscores a vision of the professional teacher as one who actively engages with research, reflection, and enquiry to inform their practice. This is in stark contrast to the English standards, which appear to advocate for a more de-professionalized version of teachers, who are seen more as technicians implementing prescribed practices than as autonomous professionals making informed decisions. Welsh standards also highlight the importance of understanding children's cognitive, emotional, and social development, suggesting a holistic approach to education. Furthermore, professional learning is prominently featured, emphasizing the necessity of continuous professional development to maintain current and effective teaching practices. This comprehensive approach seeks to balance the practical demands of teaching with the need for ongoing professional growth and adaptation.

4.5 Implications

The divergent professional standards for NQTs in England, Scotland, and Wales have significant implications for teacher preparation, professional development, and educational outcomes. Understanding these implications can help policymakers, educators, and stakeholders make informed decisions to enhance the quality and effectiveness of both teaching and teacher preparation. The implications are outlined next, and Table 4.2 provides a high-level summary of the implications.

4.5.1 Teacher Preparation

England: The prescriptive nature of England's standards suggests that TEPs may need to be highly structured, focusing on specific mandated practices. This could result in less flexibility in teacher training, potentially limiting the development of critical thinking and adaptive skills.

Scotland: The clear and succinct standards imply that teacher preparation can focus more directly on key competencies, with an emphasis on integrating research into practice. This approach may streamline teacher training, making it more focused and efficient.

Wales: The multifaceted and research-informed standards suggest that TEPs need to be comprehensive and rigorous. Programmes must equip teachers with the skills to interpret and apply complex standards, fostering a deep understanding of pedagogical theories and practices.

4.5.2 Professional Development

England: The lack of emphasis on continuous professional development in England's standards may result in fewer opportunities for teachers to engage in ongoing learning. This regulatory approach could lead to stagnation in teaching practices and a lack of adaptation to new educational challenges and research findings.

Scotland and Wales: Both countries place a strong emphasis on continuous professional learning, highlighting the need for ongoing professional development opportunities. This

supports a culture of lifelong learning, encouraging teachers to stay current with educational research and innovative practices.

4.5.3 Teaching Practices

England: The directive and prescriptive nature of England's standards may lead to a more uniform approach to teaching, with a focus on achieving specific outcomes. This could limit teachers' ability to innovate and adapt their methods to suit individual student needs.

Scotland: The integration of research into teaching practices supports evidence-based approaches. Teachers are likely to be more reflective and innovative, continuously improving their methods based on the latest research.

Wales: The holistic approach encourages teachers to consider the broader context of student behaviour and development. This can lead to more personalized and effective teaching strategies that cater to the diverse needs of students.

4.5.4 Student Outcomes

England: The focus on curriculum knowledge and behaviour management may lead to improved academic performance in standardized assessments. However, the lack of emphasis on understanding child development and continuous professional learning could limit the overall effectiveness of education in addressing the holistic needs of students.

Scotland: The emphasis on evidence-based practice can result in high-quality teaching that effectively addresses student needs, promoting positive academic and developmental outcomes.

Wales: The context-sensitive and holistic standards aim to foster an educational environment that supports the overall development of students, with the intention of creating better social, emotional, and cognitive outcomes.

4.5.5 Policy and Practice

England: Policymakers may need to reconsider the balance between prescriptive standards and professional autonomy. Increasing opportunities for continuous professional development and integrating research into practice could enhance teacher effectiveness and adaptability.

Scotland and Wales: The emphasis on research and continuous professional development in both regions suggests that policies should support and fund ongoing professional learning opportunities for teachers. Additionally, policies should encourage the integration of research into teaching practices, promoting a culture of enquiry and innovation.

The findings suggest that the different standards impact teacher preparation and professional development. In England, the focus on performance metrics can create high-pressure environments that may limit innovative teaching practices. In Scotland, the emphasis on holistic development supports a more nurturing educational environment, but may face challenges in demonstrating measurable outcomes and robust or meaningful consideration of accountability. Wales' balanced approach attempts to harness the strengths of both models, though it must navigate the complexities of integrating these philosophies effectively; such an

approach requires a degree of sophistication with respect to not only the pedagogical implications but also the political challenges entailed in these terrains.

Table 4.2

Implications of Findings From Comparative Analysis

Implications	England	Scotland	Wales
1. Teacher preparation	Expand a highly structured programme focused on specific practices Incorporate critical thinking and adaptive skills to incorporate competencies related to research and continuous learning	Focus on key competencies through clearly expressed descriptors Further integrate research into practice through collaborative enquiry	Reflect the comprehensive rigorous standards Focus on critical skills to interpret standards and advance across the profession
2. Professional development	Revise standards to codify opportunities for continuous professional learning	Continue to support a culture of lifelong learning and expand opportunities for professional growth and innovation	
3. Teaching practices	Employ a pupil-centred application with uniform standards	Strengthen evidence-based approaches with teachers through practitioner enquiry	Further develop personalized pupil learning
4. Student outcomes	Balance standardization with a holistic, pupil-centred approach	Address pupils' specific contextual needs in a rights-based approach	Focus on social, emotional, and cognitive pupil outcomes
5. Policy and practice	Focus on rebuilding professional autonomy, enquiry, and evidence-based approaches	Advocate for support and funding for professional learning, innovation, and further co-developed research initiatives and integration of research into practice	

4.6 Conclusion

This chapter detailed the differences and similarities in professional standards across England, Scotland, and Wales, reflecting regional educational priorities and ideologies. It also situated these national standards within the international comparative context and drew out implications regarding a variety of power dynamics. While Wales and Scotland emphasize research-informed, reflective practices and holistic development, England's more prescriptive approach focuses on technical competencies and mandated practices.

The results of this phase of the study informed the development of instrumentation for the video-based task used in case studies of judgement-making in TEPs in the three nations (see Chapters 5, 6, and 7). The results also helped to shape the briefing questions for the Delphi

panel (see Chapter 8) and the overall convergent cross-phase and cross-case meta-analysis (see Chapter 9) by providing a common understanding of dimensions for judgement of teaching effectiveness. In this way, the results contributed to answering the research questions and putting forward recommendations (see Chapters 10 and 11).

5 Case Study 1: University of Glasgow, Scotland

This chapter presents a case study of judgement-making in the initial teacher education (ITE) programme in the School of Education (SoE) at the University of Glasgow in Scotland. This descriptive case study includes empirical data collected through a video observation task, a questionnaire, and focus groups and interviews with university-based teacher educators, school experience tutors/associate tutors, and school-based mentor teachers. It is one of three cases in a descriptive, multi-case approach that comprises Phase 3 of this project. The case study approach allowed for contextualization and data collection from several sources to provide a multidimensional account. The chapter starts by describing the provision of teacher education at this institution, including school experiences and evaluation processes. Then case-specific methodological information is detailed and results are presented. The chapter concludes with a discussion of key findings.

5.1 Context of Case 1

To provide context for this case, this section offers background information and describes the environment in which the research took place. It is essential within a study guided by social judgement theory (SJT; Cooksey, 1996) to consider the decision-making environment and understand the conditions in which judgements of new teachers' practices are made. This includes the educational landscape, relationships among stakeholders, programme provision, criterion measures, and the types of cue information that are available to judges (e.g., visual and auditory cues in an observation). These aspects can facilitate comparisons designed to highlight the nature of judgement activities. Additionally, understanding the professional teaching standards that inform judgements and the evaluation tools employed during school-based experiences, where observations of teaching occur, is valuable.

5.1.1 Teaching and Teacher Education in Scotland: An Overview

Scotland is a relatively small nation within the UK, with a majority of the 5.4 million population (70%) located in the central belt, a corridor that includes the cities of Glasgow and Edinburgh. Although perceived as relatively affluent, one in four of Scotland's children are officially recognized as living in poverty. That number increases to 32% of children in Glasgow, with the highest rates of relative child poverty being found for children from lone parent households and those from ethnic minority households (Glasgow Centre for Population Health, n.d.). Glasgow has actually been referred to as the least peaceful location in the UK (Nesterova & Anderson, 2024).

Educational policy in Scotland is determined by the Cabinet Secretary in the Scottish Government. The government's executive agency, Education Scotland, is charged with supporting both quality and improvement and is directly accountable to government yet expected to operate independently and impartially; among other duties, Education Scotland is responsible for carrying out school inspections (see Education Scotland, 2023a). There are 32 local school authorities (or councils) with responsibility for public services, including education. Schools are mostly public non-denominational, with some Roman Catholic (approximately 15%) and independent (c. 5%) schools (Scottish Government, 2022).

Scotland's national curriculum, the Curriculum for Excellence, contains eight curricular areas; additionally, literacy, numeracy, health and well-being, and learning for sustainability are recognized as the 'responsibility of all' (Education Scotland, n.d.). The Curriculum for Excellence's broad general education has five levels. The final level, the senior phase, enables pupils to extend and deepen their learning through qualifications and also through opportunities for personal development, work placements, and volunteering. Local authorities, the Scottish Government, and the Scottish Qualifications Authority (SQA) all have important – and sometimes related – responsibilities. Qualifications at the senior phase are provided by the SQA (see SQA, n.d.), and the number of qualifications a pupil attains can vary drastically. It is important to note that although SQA is responsible for exams, it is not responsible for the curriculum. In 2021, the Organisation for Economic Co-operation and Development (OECD) published a review of Scotland's Curriculum for Excellence which specifically pointed out a disconnect of curricular aims through general education and the exam system (OECD, 2021).

Alongside these stakeholders is The General Teaching Council for Scotland (GTCS), the teaching profession's registration and regulation body, responsible for teaching standards as well as the accreditation of teacher education programmes (TEPs), which is conducted every 5 years with one interim evaluation. It is therefore the GTCS that outlines the content of the teacher education curriculum. They determine the overall aim of teacher education as preparing student teachers to become competent, thoughtful, reflective, and innovative practitioners who are committed to providing high-quality learning for every pupil. The GTCS are independent from government and receive no funding for their role of registration and regulation; rather their work is funded by fees paid by teachers and lecturers. In 2012, the GTCS became the world's first independent professional and regulatory body for teaching. Chapter 4 provides a detailed examination of the professional standard for new teachers in Scotland, the GTCS's *Standard for Provisional Registration* (SPR; GTCS, 2021b; see also Appendix A4.1). These standards encompass the dimensions used in judgement-making in this case study.

Navigating between the government and the GTCS as the accrediting body is the Scottish Council of Deans of Education (SCDE), comprising leaders from the 11 ITE providers . The SCDE maintain a number of standing committees which ensure representation on government-, organization-, and stakeholder-led groups, and they work to ensure members from each institution are engaged and leading across committees and able to take initiatives forward. The current Education Bill consultation is an example of efforts towards collaborative decision-making with teacher education providers. Current proposals include replacing the SQA and exam approach, and maximizing the role inspection plays in providing assurance and supporting teachers to improve Scottish education, including through legislation. The role of the Scottish Government Strategic Board for Teacher Education is to oversee and evaluate reforms. For this year, the Board has identified three workstreams – career-long workforce planning and retention; status of the profession; and a framework for teacher education from ITE to the induction year and into leadership – for which the SCDE will assume joint responsibility.

The 11 university-based ITE providers vary from small institutions to large Russell Group universities (GTCS, n.d.a). In addition to accreditation standards, teacher preparation programmes must align with the Quality Assurance Agency for Higher Education's (QAA's) requirements and ensure these are met. The Scottish Government's Teacher Workforce Planning Group, through Education Scotland, has the responsibility to predict the number of teachers needed and notify university programmes of the total number of student teacher places available each year. For the 2023–2024 academic year, there was a decrease of 200 places. This was due to an oversupply of teachers in the primary sector and pockets of shortages in the bursary-supported secondary subjects. Tuition is paid by the Scottish Funding Council, which means there is no tuition cost for those seeking to be a teacher. Additionally, there is a £20,000 incentive for career changers for teachers in shortage areas (i.e., Physics, Maths, Technical Education, Computing Science, Chemistry, Home Economics, and Gaelic).

Scotland is currently experiencing a substantial educational reform agenda. This includes myriad recent independent, national, and international reviews, reports, and recommendations focusing on: vision; scope of the national curriculum; assessment and qualifications; high-quality learning and teaching; learner centredness; additional support for learning; and multiple pathways to educational success (Anderson, 2023). Furthermore, due to pay disputes, Scotland recently saw national teacher strike action on a level not experienced since the Thatcher era. This resulted in multiple days of closure of nearly 75% of schools across the country. In April 2023, teachers received the largest pay package in over 20 years, an uplift of 14.6%. Funding of master's-level professional development for teachers has also been supported in recent years in an effort to emulate countries such as Finland and Norway (Cochran-Smith, 2021). However, just a few months after the pay increases were awarded, the Scottish Government announced it was unable to offer financial support for teachers to engage in master's-level learning during the 2023–2024 year. When queried, officials noted funding would be reinstated when a future budget could accommodate it (McEnaney, 2023). While the recent pay increases were somewhat larger than those awarded in other constituencies of the UK, over time there has been a modest but significant decline in educational performance, as measured by the Programme for International Student Assessment (known as PISA; Scottish Government, 2023b; Sibieta & Fullard, 2021). And despite the very different approaches taken in the differing political constituencies of the UK, the outcomes for children from the most disadvantaged backgrounds have remained broadly similar (Social Mobility Commission, 2021). Indeed, some reports would suggest that Scottish children do much better in primary school than their English counterparts, but fall behind (especially in Maths) at secondary level (Scottish Government, 2023b). It is within this educational environment, somewhat befuddled and contradictory, that this research project was undertaken.

5.1.2 Initial Teacher Education at the University of Glasgow

At the University of Glasgow, ITE is delivered via three routes: two undergraduate routes – the Master of Education (MEduc) and the Master in Design and Technology Education (MDTechEd) – and the Postgraduate Diploma in Education (PGDE). The undergraduate programmes are 5-year integrated master's programmes preparing students for the teaching

profession in the primary (MEduc) or secondary (MDTechEd) sector. The PGDE is a 1-year programme that prepares students who already have an undergraduate degree for either the primary or the secondary sector. A requirement for accreditation of the programmes is evidencing the potential to prepare graduates to meet the SPR (see Chapter 4).

Programmes in the SoE are ostensibly designed to embody the vision statement, setting a direction for planning and execution of teaching, scholarship, and research. Hence:

> The School of Education (SoE) is committed to social justice in education and to education research and practice of the highest quality. We aspire to be a world leader in addressing the contemporary educational issues of our times and to making a difference for society's most vulnerable and educationally disadvantaged.

We consider the MEduc and PGDE programmes only; no participants from the MDTechEd programme were involved in this study. The MEduc programme (approximately 600 students) consists of four overarching course strands which thread progressively through all 4 years as part of a spiral curriculum (Bruner, 1960). These strands are: Education in Practice; Curriculum Enquiry; Electives; and Education and Society. A visual representation of the MEduc programme is provided in Figure 5.1. To align with QAA requirements and levels (QAA Scotland, 2023), the four course strands build each year alongside school-based experiences in Years 1–4 and a dissertation in the fifth year.

Figure 5.1

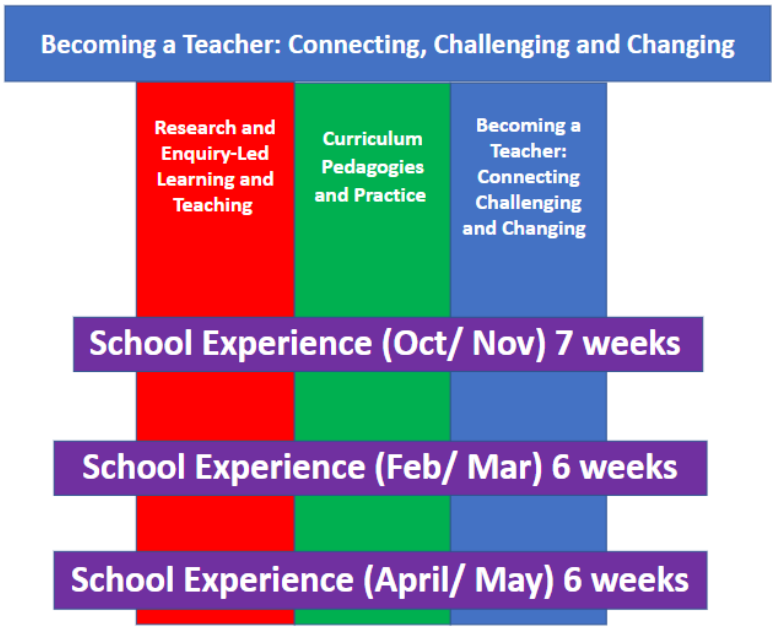
University of Glasgow MEduc Programme Structure

Year 1	**Fundamentals of Education (40 credits) Semester 1 & 2 Level 7 20 credits 1A 20 credits 1B**	Education in Practice (30 credits) Semester 1 & 2 Level 7 15 credits: 1A 15 credits: 1B	Curriculum Enquiry (30 credits) Semester 1 & 2 Level 7 15 credits: 1A 15 credits: 1B	WTPE?/ Theology (20 credits) Semester 1 Level 7	School Experience (Formative) Semester 1 & 2	Catholic Teacher Formation	120
Year 2	Education & Society (20 credits) Semester 2 Level 8	Education in Practice (30 credits) Semester 1 & 2 Level 8 15 credits: 2A 15 credits: 2B	Curriculum Enquiry (40 credits) Semester 1 & 2 Level 8 20 credits: 2A 20 credits: 2B	What If?/Theology (20 credits) Semester 1 Level 8	School Experience (10 credits) Semester 1 & 2 Level 8	Catholic Teacher Formation	120
Year 3	Education & Society (30 credits) Semester 2 Level 9	Education in Practice (20 credits) Semester 1 & 2 Level 9	Curriculum Enquiry (30 credits) Semester 1 Level 9	Elective (30 credits) Semester 1 & 2 Level 9	School Experience (10 credits) Semester 1 & 2 Level 9	Catholic Teacher Formation	120
Year 4	Education & Society (30 credits) Semester 1 Level 10	Education in Practice (20 credits) Semester 1 Level 10	Curriculum Enquiry (20 credits) Semester 1 Level 10	Elective (30 credits) Semester 1 Level 10	School Experience (20 credits) Semester 2 Level 10	Catholic Teacher Formation	120
Year 5	Education & Society (30 credits) Semester 1 Level 11	Professional Enquiry and Decision Making (30 credits) Semester 1 Level 11	Dissertation (60 credits) Semester 2 Level 11				120

The PGDE programme has approximately 220 students – around 120 for the primary sector and 100 for the secondary sector. It consists of three courses and one school experience course, which occurs in three separate parts across the year (see Figure 5.2). In June 2021, the PGDE achieved unconditional reaccreditation (meaning no changes to the planned curriculum, structure, or documentation are required for a period of up to 6 years). The refreshed PGDE programme took inspiration from Korthagen's (2004, 2017) holistic approach to teacher education. It aims to help students to reflect on their beliefs, values, and positionality in relation to important issues in education and to build their professional identity as reflective and enquiring teachers.

Figure 5.2

University of Glasgow PGDE Programme Structure



5.1.3 University of Glasgow Practices and Processes for Judging Teaching Effectiveness

experience, students are assessed according to the SPR, with consideration given to the stage they are at in their ITE programme. The judgements made by the evaluators are captured using the 'End of Placement Report Form', with ratings being either satisfactory or unsatisfactory (see Appendix A5.1). The form is agreed and utilized by another provider of ITE in a collaborative effort to provide consistency in practices of evaluating new teachers; thus the form reflects the name of both institutions at the top. Schools working with candidates from either university located in Glasgow will therefore capture their judgements using the same form. The purpose of the assessment is to evaluate students' teaching practices in the classroom according to the SPRs for each placement of school experience. This purpose is communicated to candidates in handbooks and through the 1-page brief for each placement, and it is reviewed with mentor teachers at a beginning-of-year online training session each year. Further, course leaders and school experience tutors review the

expectations and passing requirements with students. This assessment provides evidence of students' progress and includes a written narrative of what the student has demonstrated. Evaluations occur at specified intervals during preparation:

- MEduc 1: placements are not summatively assessed;
- MEduc 2: a joint formative assessed visit (FAV) is carried out by the end of Week 3;
- MEduc 3: a joint FAV is carried out by the mid-point of each placement;
- MEduc 4: a joint FAV is carried out by the end of Week 5 and an interim report is prepared by Week 6; and
- PGDE: a joint FAV is carried out by the end of Week 4 in each placement.

The End of Placement Report has both a formative and a summative purpose. The formative purpose is evident in the feedback comments provided on the FAV proforma (Appendix A5.2), which students are encouraged to use to inform their practice and identify areas for development and next steps. A narrative of what the student has demonstrated is included for each SPR and used to guide a professional dialogue with the school experience tutor and mentor teacher. After each placement, student teachers, in collaboration with their school experience tutor, complete a Personal and Professional Development Plan (Appendix A5.3) with development targets for their next placement or their probation year.

Data are generated in the format of satisfactory or unsatisfactory ratings for each of the eight standard areas. Performance data are used to provide the student with feedback on their demonstration of SPR as well as to make decisions about progression in the programme and any need to redo a placement. Results guide conversations between the student, the school experience tutor, and the mentor teacher to foster growth. Results of school placements are reviewed by the course leader, the coordinator of school placements, the programme external examiner, and the final exam board. The final evaluation decides if candidates will be able to begin their probation year or not; as the consequences relate to entry into the profession, they are quite impactful. Those who have failed the placement set targets for improvement. Performance data from school experiences are not connected to academic assessment data from courses.

Following ITE, all newly qualified teachers must complete a period of probationary service. Two routes are available: the 1-year Induction Scheme provides a guaranteed 1-year full-time post in a local authority to every eligible graduate. There is also the Flexible Route for those want or need to work part-time or desire to complete the probationary period outside of Scotland. It is at this point that ITE is considered attained and the GTCS standards for full registration begin. It is within this provision of teacher education this study was conducted.

5.2 Case-Specific Methods

Methods applicable to the entire research project are presented in Chapter 2; this included the theoretical framework of SJT, strategies to ensure trustworthiness of results, and the ethical approach taken. Methods which relate to all three case studies in the multi-case design are also presented in Chapter 2, Section 2.7 (the case study protocol is provided in Appendix

A2.3). Therefore, this section only includes considerations specific to recruitment and data collection applicable to this case.

Seventeen participants took part in the video task and questionnaire: five university teacher educators; four school experience tutors/associate tutors; and eight school-based mentor teachers (Table 5.1). Participants were selected through purposeful sampling (Cohen et al., 2018). The goal was to select participants who reflected the various roles of individuals who conduct observations and evaluate teaching effectiveness during educator preparation and could best contribute to answering to the research questions. These participants demonstrated a perspective within a defined context and had enough information for in-depth exploration (Merriam, 1998).

Table 5.1

Case Study 1 Participants

	Teacher educators	Associate tutors	Mentor teachers	Overall
Potential participants	39	47	Unknown	N/A
Video task and questionnaire	5	4	8	17
Focus group/interview	4	1	4*	9

Note. * A total of 4 mentor teachers participated in interviews following adjusted recruitment parameters.

A recruitment script (see Appendix A5.4) was sent to university-based teacher educators who were full time staff. As the Principal Investigator (PI) was in the position of administrative line manager to a majority of staff working in ITE, recruitment was conducted via email invitation by a Co-Investigator and focus groups were arranged and conducted by the Research Associate. While teaching courses on ITE programmes and assessing course assignments, none of the staff were actively involved in observing teaching effectiveness of students during placements. Recruitment occurred during the autumn term of 2023. An initial request was sent with a reminder script sent out approximately 2 weeks after the first.[1] A total of five teacher educators completed the video task and questionnaire. Of these, four agreed to contribute to a 45-minute focus group; due to scheduling constraints, two focus groups were conducted with two participants in each group.

At the time of this study, nearly all school experience tutors held the position titled ‘associate tutor’ in the SoE. This position is a part-time, non-permanent teaching role at the equivalent academic level of lecturer. The role is flexible and can involve delivering instruction, marking, providing instructional support for students, and supervising students while in

[1] It is important to note this study was carried out during University and College Union (UCU) strike action which occurred from November 2022 to October 2023, inclusive of a national marking and assessment boycott from 20 April to 6 September 2023.

school-based experiences (i.e., on placement). School experience tutors report to one of three senior associate tutors who coordinate their workload and contribution to ITE programmes. The PI met with the senior associate tutor to discuss the project; the senior associate tutor then emailed the recruitment script to all school experience tutors in January 2024 with one follow-up reminder 2 weeks later. In total, four associate tutors completed the video task and questionnaire; one agreed to an individual interview.[2]

Recruitment of school-based mentor teachers was facilitated by the SoE Coordinator of School Placements and Partnerships, whose role involved organization and administration of school experiences with local authority representatives. Agreed communication protocols dictated that head teachers directly communicate with mentor teachers, not the university staff. It is therefore unclear how many head teachers forwarded the script to mentor teachers in the schools in which placements were made in May and June of 2023.[3] Initial recruitment resulted in eight participants completing the video task and questionnaire. Of these, one agreed to a focus group and, as such, an individual interview was conducted. Given the low response rate to the focus group, the team decided to attempt to capture mentor teachers' opinions and experiences through a modified process. The SoE Coordinator of School Placements and Partnerships agreed to select a small group of mentor teachers to directly send a request to participate in a focus group in January 2024. Due to scheduling constraints for focus groups, individual interviews were conducted with three mentor teachers. These mentor teachers did not complete the video task questionnaire but instead were presented with summary information about the task and adjusted focus group questions to accommodate for them not completing the task. Therefore, mentor teacher focus group responses reflect four individual interviews. It is important to be clear that these mentor teachers were not the same participants who completed the video task and questionnaire. Therefore, mentor teacher focus group responses reflect four individual interviews.

We did encounter a relatively high drop off of participants with the video task questionnaire. A number of individuals completed the informed consent and demographic questions, but when the first judgement item in the task was presented and a rationale queried, these individuals did not continue. The completion rate for each group of participants is included in Table 5.2.

[2] The decision to delay recruitment until this point in the project was due to the UCU strike action in addition to employment changes at the university level, which saw all associate tutor contracts renegotiated between April and October 2023.

[3] The script was sent in the window of time after the national pay dispute had been resolved and before the end of the school year (May–June 2023).

Table 5.2

Case Study 1 Completion Rates

	Teacher educators	Associate tutors	Mentor teachers
Began video task and questionnaire	20	10	23
Completed video task and questionnaire	5	4	8
Completion rate	25.0%	40.0%	34.8%

While definitive reasons for the survey dropout rate remain unknown, plausible explanations can be attributed to both survey design and participant-related factors. Although participants were informed of the survey's duration in both the recruitment script and consent form, it is acknowledged that survey fatigue may have contributed to abandonment. Moreover, given the study's focus on capturing judgement decisions and policies, the requirement for qualitative responses to open-ended questions was essential but may have imposed an additional cognitive burden on participants. The complex nature of the judgement-making process itself, which is the central focus of this research, may have presented challenges for participants in articulating their thoughts and reasons.

5.3 Video Task and Questionnaire Results

5.3.1 Participant Demographics

Seventeen participants, all of whom were current or former teachers, took part in the video task and questionnaire. The sample comprised five university teacher educators, four school experience tutors (referred to as associate tutors), and eight school-based mentor teachers. While mentor teachers and associate tutors were actively engaged in student teaching evaluations at the time of the study, teacher educators were not involved in student observations during placements. Each participant group brought a unique perspective to the task. A detailed overview of participant roles, qualifications, and experience is presented in Table 5.3.

Table 5.3

Participant Demographics for the Video Task and Questionnaire

		Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)	Overall (n = 17)
Gender	Female	3	3	7	13
	Male	1	1	1	3
	Non-binary/third gender	0	0	0	0
	Prefer not to say	1	0	0	1

Overall experience in education	Under 25 years	1	0	5	6
	25 to 29 years	1	0	1	2
	30 to 39 years	3	2	2	7
	40 to 49 years	0	2	0	2
Year of experience in current role	Under 25 years	5	4	6	15
	25 to 29 years	0	0	0	0
	30 to 39 years	0	0	2	2
	40 to 49 years	0	0	0	0
Route into teaching	Undergraduate	1	1	2	4
	Postgraduate	4	2	6	12
	No qualifications	0	0	0	0
	Others	0	1	0	1
Teaching qualification	Nursery	1	1	0	2
	Primary	2	3	1	6
	Secondary	3	1	7	11
	Specialist	2	0	0	2
	None	0	0	0	0
	Other	1	0	0	0
Country where teaching qualification was obtained	Scotland	4	4	8	16
	England	1	0	0	1
	Wales	0	0	0	0
	Northern Ireland	0	0	0	0
	Other	0	0	0	0
Highest level of qualification	Below bachelor's degree	0	0	0	0
	Bachelor's degree	0	1	2	3
	Postgraduate	0	2	2	4
	Master's degree	3	1	2	6
	Doctorate	2	0	2	4

Most participants were female and with substantial years of experience in the field of education. Participants had been working in the role of teacher educators from 6 to 23 years. Many of the participants qualified as teachers themselves through the postgraduate route (70.5%), with only four undertaking the undergraduate programme for a teaching qualification. Collectively, 65.0% of the participants had experience teaching secondary education. All but one participant obtained their teaching qualification in Scotland. All but three participants had attained qualifications beyond the bachelor's level; two of these were mentor teachers and one was an associate tutor.

5.3.2 Results from the Video Observation and Judgement Task

Participants' range of responses and patterns of consensus and dissensus on observed teaching effectiveness are presented in Tables 5.4, 5.5, and 5.6. Participants were asked to watch the 15-minute video which simulated the natural process of observation used in teacher education and then provide judgements in each of the seven dimensions of the United Nations Educational, Scientific and Cultural Organization (UNESCO) *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019; see Chapter 2) and an overall judgement of the teaching effectiveness; they were also asked to indicate which dimensions were most and least difficult to judge (see Appendix A2.1). In addition, they were asked in open-ended prompts to explain *how* they made these judgement decisions in order to capture the cues utilized, their judgement policies, and potential influences. Results are presented according to role in the judgement-making process as well as the overall results.

The overall judgement by teacher educators of teaching effectiveness was 3.4 out of a possible 5.0, indicating above satisfactory teaching was demonstrated. The judgements made by teacher educators varied, with some dimensions being considered highly effective to nearly unsatisfactory. These judgements did not yield any immediate pattern. The dimension rated highest was 'learning environment' and the lowest was 'instructional strategies'. The dimension where the highest standard deviation was found was 'learning environment'; due to the spread of scores, it is this area which reflects a higher degree of inconsistency among the raters. This was followed by 'planning & preparation'. Mean scores for the seven individual areas ranged from 2.8 to 3.6 (R = 0.80).

Table 5.4

Teacher Educators' Judgements on Seven Elements of Observable Practices of UNESCO Professional Teaching Standards

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(n = 5)							
Q1. Learners	0	3	0	2	0	4	3.20	0.98
Q2. Content	0	1	3	0	1	3	2.80	0.98
Q3. Research	0	2	1	1	1	4	2.80	1.17
Q4. Planning & preparation	1	0	3	0	1	3	3.00	1.26
Q5. Instructional strategies	0	1	2	1	1	3	2.60	1.02
Q6. Learning environment	2	1	0	2	0	2, 5	3.60	1.36
Q7. Assessment	0	2	1	1	1	4	2.80	1.17

Q8. Overall rating	1	1	2	1	0	3	3.40	1.02

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.

The overall judgement of teaching effectiveness by associate tutors was 4.25 out of a possible 5.0, indicating effective teaching was demonstrated. This was the highest rating among all three groups. Judgements made by associate tutors also varied from highly effective to nearly unsatisfactory (2). The lowest scoring option of unsatisfactory (1) was not given by any associate tutor. Two areas were rated highest including 'learning environment' and 'content'. The lowest rated area was 'research', which was also the area of highest deviation followed closely by 'assessment'. Across all areas, there was less deviation among the ratings of associate tutors than the teacher educators. Mean scores for the seven individual areas ranged from 3.25 to 4.50 (R = 1.25).

Table 5.5

Associate Tutors' Judgements on Seven Elements of Observable Practices of UNESCO Professional Teaching Standards

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(n = 4)							
Q1. Learners	1	2	1	0	0	4	4.00	0.71
Q2. Content	2	2	0	0	0	4, 5	4.50	0.50
Q3. Research	1	1	1	1	0	2, 3, 4, 5	3.50	1.12
Q4. Planning & preparation	2	1	1	0	0	5	4.25	0.83
Q5. Instructional strategies	0	2	1	1	0	4	3.25	0.83
Q6. Learning environment	2	2	0	0	0	4,5	4.50	0.50
Q7. Assessment	1	2	0	1	0	4	3.75	1.09
Q8. Overall rating	2	1	1	0	0	5	4.25	0.83

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.

Results of the task for mentor teachers are presented in Table 5.6. The overall judgement of teaching effectiveness by mentor teachers was 3.38 out of a possible 5.0, similar to the teacher educators, indicating above satisfactory teaching was demonstrated. Judgements made by school-based mentor teachers varied from highly effective to unsatisfactory. The area rated highest was 'learning environment' and the lowest was 'assessment', followed closely by 'instructional strategies'. Across all areas, there was less deviation among the ratings of mentor teachers than those of teacher educators, but slightly more than the associate tutors, with the most variation occurring in the area of 'assessment'. The highest

range of scores across the seven dimensions was demonstrated by mentor teachers, with mean scores ranging from 2.50 to 4.25 (R = 1.75).

Table 5.6

Mentor Teachers' Judgements on Seven Elements of Observable Practices of UNESCO Professional Teaching Standards

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(n = 8)							
Q1. Learners	0	3	4	1	0	3	3.25	0.66
Q2. Content	0	6	1	1	0	4	3.63	0.70
Q3. Research	1	2	4	1	0	3	3.38	0.86
Q4. Planning & Preparation	1	5	1	1	0	4	3.75	0.83
Q5. Instructional Strategies	0	1	3	4	0	2	2.63	0.70
Q6. Learning Environment	3	4	1	0	0	4	4.25	0.66
Q7. Assessment	0	2	1	4	1	2	2.50	1.00
Q8. Overall Rating	0	4	3	1	0	4	3.38	0.70

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.

The learning environment was the highest rated dimension by all three groups. The lowest rated areas varied by group; no single area was consistently seen as a weakness in teaching observed. Teacher educators rated instructional strategies the lowest, associate tutors rated research the lowest, and mentor teachers rated assessment and instructional strategies the lowest; however, all were still satisfactory. The areas of highest deviation were also variable across groups. There was more variation in the ratings of teacher educators than the ratings of associate tutors or mentor teachers.

5.3.3 Results: Strategies and Rationales for Ratings

Along with the ordinal judgement provided for the observed video lesson, participants were asked an open-ended question for each of the seven dimensions: 'How did you decide what level of performance was demonstrated?' This was done in order to capture cues utilized, judgement policies, and potential influences. This question was asked for all seven areas which were rated, and the qualitative responses were analysed using the constant comparative method of data analysis for each of the three groups of participants. Findings are presented in Tables 5.7, 5.8, and 5.9 according to roles, and indicative statements of participants are provided. A coding analysis of how evaluators reached judgements for levels of performance on the seven areas of teaching practices was conducted. According to SJT (Cooksey, 1996), the ways (i.e., strategies) judges use available cues to make decisions is termed 'cue

utilization validities'; these are judges' attempts to understand the teaching observed. If a strategy was used even once, it was recorded. Prevalence and distribution of strategies and rationales, or warrants, for judgements are presented. We have included quotes from participants to illustrate and provide credibility to findings; participant codes from the analysis processes are included.

5.3.3.1 Teacher Educators

Results of our analysis suggest that university-based teacher educators used four strategies to determine an observation rating: (a) classroom cue utilization; (b) suggestions for lesson improvement; (c) internal expectation criteria; and (d) no identified strategy. As teacher educators reasoned with a given strategy, they employed a specific rationale or backing for the strategy being used. There were three types of justifications evident: professional judgement; personal judgement; and indeterminate judgement. We now describe the strategies and warrants in detail with typical examples provided. The most recurrent justification was professional judgement, and the most used strategy was classroom cue utilization. Many of the judgement cues used to assess the student teacher's performance were the observed actions of the teacher, multiple examples of demonstration, and observed pupil actions.

Table 5.7

Teacher Educators' Judgement Strategies and Rationales

Professional judgement		Personal judgement	Indeterminate judgement
Classroom cue utilization (n = 67)	Suggestions for lesson improvement (n = 16)	Using internal expectation criteria (n = 4)	No identified strategy (n = 7)
Observed teacher action (21) Observed pupil action (16) Physical environment cues (7) Multiple general examples of evidence to support rationale (6) Context cues (5) Learning materials (4) Pupil learning (3) Teacher and pupil interaction (3) Explanatory rationale (2)	Lesson improvement (10) Observed omissions (6)	Internal criteria (4)	Unable to explain (3) Need more to make judgement (2) None (2)

Note. Total codes from qualitative questionnaire statements: Q1–8; n = 94.

Classroom cue utilization (rationale: professional judgement). In their decision-making, participants utilized perceived aspects of a student teacher's observable practices and cues considered relevant from the classroom. This strategy accounted for approximately 71% of cues coded, indicating what judges looked to most when making a decision. Their attention

was directed to multiple cues, some of which were interdependent. The most common cues were from the teacher's actions, the pupils' actions, and the classroom learning environment. Both positive and negative occurrences of these cues were noted. A few examples of observed teacher actions included:

- She knows individual names. She directs questions to particular children. She walks around the classroom ensuring she is engaging with learners and is checking in on them (Q1b. Learners)
- Communication was very strong in places but also weaker in other aspects (Q2b. Content)
- She circled the tables; her presence was strong (Q3b. Research)
- Stand before the board to dictate key information (Q5b. Instructional strategies)
- Time management appeared to be very good and this is an important factor (Q4b. Planning & preparation)

Pupils' actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. A few of these were:

- Pupils were clearly engaged in the tasks. At one point a group was asked if they had evidence (Q2b. Content)
- Pupils were assessed when being asked to explain what the questions meant (Q3b. Research)
- Students were aware of objective but some unsure of how they could apply what they had learned from their resources (Q4b. Planning & Preparation)
- there is evidence of a good relationship between teacher and pupils (Q6b. Learning Environment)

Another main strategy involved a statement of what the teacher did but was specifically followed by multiple examples as evidence to support the main statement. For example:

- Make the subject matter more accessible (Use of ICT, show me boards, visuals to support their mastery of the content with the students, round robin feedback to ensure that the students have mastered the skills collectively; Q2b. Content)
- The teacher reinforced the context of the lesson and then selected students to reinforce their knowledge of her research. This was also applied as a walk and talk. This can sometimes alienate students (put them on the spot) as student could not really answer the question and then the teacher asked someone else (Q3b. Research)

Judges also observed the physical environment of the classroom. This included how the desks were arranged, what was on the blackboard, if the room appeared crowded as well as materials used for learning. Rationales supporting identification of these cues included:

- Evidence of planning & preparation on desk and on board (Q4b. Planning & preparation)
- Seemed a little disorganized – tables full of materials, not supporting an effective working space. Board a mess, although full of interesting meaningful questions! (Q4b. Planning & preparation)

- Given the restrictions of the rather crowded classroom, the organization was effective (Q6b. Learning environment)
- Jotter work and group work was also utilized (Q7b. Assessment)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy builds from the evaluators' professional judgement and reflects their role as individuals responsible for new teachers entering the profession as well as their own experiences with teachers, student teachers, and pupils in multiple classrooms and schools. From this perspective, with years of experiences and prior knowledge, evaluators used professional judgement by indicating what was not observed and suggesting how the lesson might be changed to improve the quality and rating assigned. For example, when rating the 'learning environment', one participant explained, 'there is an educational deficit where alleged group work, with markers and flip chart paper, takes the place of meaningful discussion' (Q6b). Similarly, a further participant reasoned that 'The teacher needed to work more on creating a calm space', which could have resulted in a higher level of performance. Further examples from the data to support this reasoning strategy include:

- I would suggest taking time to explore aspects in group and then the student could be scaffolded in any difficulties they encounter (Q3b. Research)
- The contentious question 'can truth change' is not explained (Q2b. Content)
- There is little to no evidence of order, structure and appropriate curricular content (Q2b. Content)
- The learning objectives were not clearly articulated (Q4b. Planning & preparation)
- The assessment task was not consistent (Q7b. Assessment)

Using internal expectation criteria (rationale: personal judgement). This rationale for respondents' judgements appeared to involve underlying personal constructs such as the evaluator's beliefs, value systems, expectations, or even emotions. While relatively uncommon among the strategies used (4%), perceptions that come from within the judges themselves were evident. It is important to note that strategies coded as internal criteria may have developed through professional experience; the scope of the data collected did not provide any indication as to whether or not internal criteria were based on professional knowledge or personal preferences. Statements given from participants included:

- It seems to be an activity for the sake of an activity (Q1b. Learners).
- I found some of the delivery a little brusque (Q5b. Instructional strategies).
- I liked the targeting of questions (Q5b. Instructional strategies).
- A sense that active learning has been misconstrued as something loud, full-on, and frenetic (Q6c. Learning environment).

No identified strategy (rationale: indeterminate judgement). Some participants indicated the basis for their judgments was not exactly known or not based on a strategy, or they were unable to articulate or establish a rationale. Some stated they were not in a position to provide a judgement. Indicative responses included:

- Impossible to answer this question (Q1b. Learners)

- I do not have the knowledge base to respond objectively to this section – therefore, I do not consider my response an accurate judgement of the teacher's practice (Q2b. Content)
- Mixed evidence here, and without any familiarity with the course content, it is difficult to make a properly informed judgement (Q2b. Content)
- I cannot comment on this as I am unaware of what fair, valid, and reliable assessment involves in this discipline (Q7b. Assessment)

5.3.3.2 Associate Tutors

Table 5.8 indicates the range of evidence participants drew on to judge teaching effectiveness. Associate tutors in the role of school experience tutors at the University of Glasgow used three strategies to determine an observation rating: (a) classroom cue utilization; (b) suggestions for improvement; and (c) internal expectation criteria. As associate tutors reasoned with a given strategy, they employed two rationales for backing strategies: professional judgement and personal judgement. We now describe the strategies and rationales of raters in detail, with typical examples included.

Table 5.8

Associate Tutors' Judgement Strategies and Rationales

Professional judgement		Personal judgement
Classroom cue utilization (n = 67)	Suggestions for lesson improvement (n = 11)	Internal expectation criteria (n = 1)
Observed teacher action (21) Observed pupil actions (14) Context cues (7) Teacher and pupil interaction (6) Multiple examples as evidence to support rationale (5) Learning materials (5) Physical environment cues (5) Pupil learning (2) Specific named strategies (2)	Lesson improvement (10) Observed omissions (1)	Internal criteria (1)

Note. Total codes from qualitative questionnaire statements: Q1–8; n = 79.

Classroom cue utilization (rationale: professional judgement). In their decision-making, participants utilized perceived aspects of a student teacher's observable practices and cues considered relevant from the classroom. This strategy accounted for approximately 85% of cues. The most common cues were from the teacher's actions, the pupils' actions, interactions between the teacher and pupils, and contextual cues. A few examples of classroom cues from observed teacher actions were:

- clarified for almost every group (Q1b. Learners)
- Student teacher has a sound knowledge of her subject, able to pinpoint details of expected responses to specific parts of lesson/board content/questions (Q2b. Content)

- Used various AifL [Assessment is for Learning] strategies such as Pair & Share/Group/Show Me Boards to share learning, encouraging all of the students to get involved (Q7b. Assessment)

Pupils' actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. A few of these were:

- There was too much information at the one time, causing confusion among the students (Q1b. Learners)
- Her question techniques and pupil responses inform her of pupil understanding of learning intention and task and success criteria (Q3b. Research)
- Children were motivated and engaged and allowed to do it in their own way (Q6c. Learning environment)

Another main strategy involved a statement related to the context cues from the classroom and the lesson being taught. For example:

- Familiar routines – recapping on task, clarifying what is required, timed elements (Q2b. Content)
- Board content prepared in advance and available for use by student teacher and pupils (Q2b. Content)
- Established routines/familiar structure for the learners to focus on the task (Q4b. Planning & preparation)
- Clearly a respectful environment – students confident to contribute, speak out, seek assistance from others when struggling; manner, interaction, relationships were very good (Q6c. Learning environment)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy, which accounted for 14% of the strategies coded, was a focused on lesson improvement, building from the associate tutors' professional judgement. This reflected their role in teacher education, which specifically involves supporting students on placement in schools, conducting observations, and completing the end of placement reports and development plans. When using the lesson improvement strategy, ratings were justified by referencing what could have been done differently to support a different rating or clarification of what should have been done (i.e., an omission). Examples from the data of the basis for this reasoning strategy included:

- Noticed a need for explanation of the task having to be repeated (Q1b. Learners)
- The learning objectives were skimmed over far too quickly at the beginning (Q4b Planning & preparation)
- A need to stop and draw the students eyes to the front when discussing new learning, then set to task (Q5b. Instructional strategies)
- Instructions, particularly around the focused task were rushed; class were already keen to start and talking; this meant that the teacher had to go round each group reinforcing, clarifying and reassuring students (Q5b. Instructional strategies)

Using internal expectation criteria (rationale: personal judgement). The rationale for this judgement was based on personal judgement, particularly a subjective feeling. There was

only one identified occurrence of this strategy. The statement given from the one participant was: 'Felt she was feeding the children the answers' (Q3b. Research).

5.3.3.3 Mentor Teachers

Table 5.9 shows the range of evidence mentor teachers relied on to judge teaching effectiveness. School-based mentor teachers used four strategies to determine their observation rating using the evidence. These four strategies are: (a) classroom cue utilization; (b) suggestions for lesson improvement; (c) using internal expectation criteria; and (d) no identified strategy. As mentor teachers reasoned with a given strategy, they appealed to specific justifications (i.e., backing) for the strategy being used. There were three types of justifications to which mentor teachers appealed: professional judgement; personal judgement; and indeterminate judgement.

Table 5.9

Mentor Teachers' Judgement Strategies and Rationales

Professional judgement		Personal judgement	Indeterminate judgement
Classroom cue utilization (n = 102)	Suggestions for lesson improvement (n = 41)	Using internal expectation criteria (n = 6)	No identified strategy (n = 7)
Observed teacher actions (34) Observed pupil actions (23) Teacher and pupil interaction (12) Context cues (12) Multiple examples as evidence to support rationale (11) Learning materials (10) Physical environment cues (7) Pupil learning (2) Formative assessment results (1)	Lesson improvement (27) Observed omissions (14)	Internal criteria (6)	Need more to make judgement (5) Uncertainty (2)

Note. Total codes from qualitative questionnaire statements: Q1–8; n = 156.

Classroom cue utilization (rationale: professional judgement). In decision-making, mentor teachers drew on perceived aspects of the student teachers' observed practices and cues considered relevant from the classroom. This strategy accounted for approximately 65% of cues. The most common cues were from the teacher's actions, the pupils' actions, interactions between the teacher and pupils, and contextual cues. A few examples of classroom cues from observed teacher actions were:

- Is able to direct pupils to key phrases and content that will allow them to be successful in their learning (Q2b. Content)
- Questioning was not skilful enough to determine the levels of individual pupils (Q3b. Research)

- She mentioned ‘connecting’ learning and was encouraging pupils to use previous lesson content which highlights that she is effectively planning her lessons to allow pupils to make connections with their learning (Q4b. Planning & preparation)
- The teacher had a good relationship with pupils but could also appear abrupt in her mannerisms towards pupils (Q6b. Learning environment)

Pupils’ actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. These included:

- She went through each individual group to check for understanding, but pupils had the same question, which implies as a class perhaps there was not full understanding on the task (Q1b. Learners)
- Given the difference in competence shown by the verbal response of learners, the teacher needed to support pupils better in the choice of task and in the expected outcomes (Q5b. Instructional strategies)
- There appears to be good learning conversations going on and learners appear to be self-motivated and get on with the task set (Q6b. Learning environment).
- There was visible formative assessment as the teacher asked questions at the start of the lesson and during group chat, but not every student participated (Q7b. Assessment)

Another key strategy appeared to involve a statement of what the teacher did, but specifically followed by multiple examples as evidence to support the main statement. For example:

- The teacher was able to help the pupils engage cognitively with the lesson as they reasoned with her about the intention of the lesson – to provide a summary. The teacher listened carefully to every question she was asked and stated the learning intention of the lesson – summarizing the key details with textual evidence and academic evidence. One pupil asked about discrimination, and she directed the question back at the pupil – ‘have you got evidence?’ – therefore taking the pupil to the next level of understanding. (Q1b. Learners)
- The teacher was engaging as many students as they could by targeting questions. Getting students to read from the board seemed ineffective but asking for an explanation in their own words was effective. Allowing students time to think and trying to show how previous lessons impacted on the one being taught was good practice. (Q3b. Research)

Judges also observed the physical environment of the classroom. This included how the desks were arranged, what was on the blackboard, if the room appeared crowded as well as materials used for learning. Rationale supporting identification of these cues included:

- Too much information for them to take on board (Q1b. Learners)
- Effort had clearly been taken to write the day’s work on the board (Q4b. Planning & Preparation)
- The learners have lots of paperwork and notes regarding the subject so they know what they are doing (Q4b. Planning & preparation)
- Pupils were comfortable enough to contribute and participate in the lesson (Q6b. Learning environment)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy, which accounted for 26.5% of the strategies, was a focus on lesson improvement building from the mentor teachers' professional judgement. It reflects their role in teacher education as the practising teacher currently in the classroom. When using the lesson improvement strategy, the mentor teachers justified ratings by referencing what could have been done differently to obtain a different rating, or by clarifying what should have been done (i.e., an omission). Examples from the data to substantiate this reasoning strategy included:

- She failed to make the subject matter accessible (Q2b. Content)
- The teacher did not use any digital technology during her lesson (Q5b. Instructional strategies)
- Slowing the pace of speech at points may also have fostered a slightly calmer working environment (Q6b. Learning environment)
- Would have been interesting to see what the learners produced by the end of the session (Q7b. Assessment)

Using internal expectation criteria (rationale: personal judgement). The rationale for these judgements, based on personal judgement, involved underlying constructs such as the evaluator's preferences or expectations. This was relatively rare among the strategies used (4.5%), yet perceptions that are personal to the judges themselves were evident. These strategies, coded as internal criteria, may have developed through professional experience; however, identifying this was outside the scope of the data collected, which did not provide any indication of whether the internal criteria were based on professional knowledge or personal preferences. Statements from participants included:

- I was struggling to understand the complex questions that were being read by students (Q1b. Learners)
- The information on display was quite visually overwhelming (Q4b. Planning & preparation)
- I wouldn't have asked readers of different levels to read in front of the whole class in the way she did (Q4b. Planning & preparation)
- I also thought the way she questioned the learners was too direct at times, and it would have made me as a learner feel uneasy (Q5b. Instructional strategies)

No identified strategy (rationale: indeterminate judgement). Some participants indicated that the basis for their judgements were not exactly known or were not based on a strategy, or that they were unable to articulate or establish a rationale. Some said they were not able to make a judgement based on what was seen in the video. Indicative responses included:

- Was unsure if learners knew what the final assessment outcome was. Were they being assessed on the poster or on how they present it? (Q3b. Research)
- This was perhaps difficult to determine from the video (Q5b. Instructional strategies)
- I didn't feel the social interactions were particularly positive (Q6b. Learning environment)
- I don't think this was visible in the clip (Q7b. Assessment)

5.3.3.4 Comparison of Judgement-Making Strategies

Comparative analysis was used to examine the pattern of rationales among the groups of judges. Overall, participants relied heavily on the available perceived cues to make judgements of teaching effectiveness, thus demonstrating similarity with attempts to understand teaching performance ('cue utilization validities'; Cooksey, 1996). Of the 329 rationales coded from qualitative data across the three groups, 236 (71.7%) reflected the strategy of classroom cue utilization. The same top four strategies occurred across all groups, reflecting little variation in the way decisions to assign a level of performance were justified. Additionally, a further 20.7% of strategies (n = 68) involved suggestions for lesson improvement. Mentor teachers engaged suggestions for lesson improvement more so than teacher educators or associate tutors, sometimes indicating what they would have done themselves. Together with classroom cues, these strategies of judgement-making demonstrated a majority of backings founded on professional judgement. Only a few instances of warrants made on personal judgement were identified (3%). This was similar across teacher educators and mentor teachers, with only one instance occurring with associate tutors. While there were a small number of cases for teacher educators and mentor teachers in which the participants could not explain a rating, decided they needed more than what was in the video to make a judgement, or were uncertain, this was not demonstrated by the associate tutors. The exhibition of indeterminate judgement was quite low overall (4.2%).

Table 5.10

Participant's Perspective on the Easiest and the Most Difficult Element to Judge in UNESCO Professional Teaching Standards

	Teacher educators (n = 5)		Associate tutors (n = 4)		Mentor teachers (n = 8)		Overall (n = 17)	
	Most difficult	Easiest	Most difficult	Easiest	Most difficult	Easiest	Most difficult	Easiest
Learners	1	0	1	0	1	1	3	1
Content	2	0	1	0	1	0	4	0
Research	0	0	0	0	1	0	1	0
Planning & preparation	1	0	0	2	1	2	2	4
Instructional strategies	0	0	0	0	0	2	0	2
Learning environment	0	5	1	1	1	3	2	9
Assessment	1	0	1	1	3	0	5	1

Note. Questionnaire: Q9–10; only one choice for most difficult and easiest could be selected.

5.3.5 Results: Easy and Difficult Dimensions of Judgement

Participants were also asked to indicate which of the seven UNESCO dimensions they found most difficult to judge in the teaching video and which they found easiest to judge.

Additionally, participants were prompted to explain *why*. Nominal responses are indicated in Table 5.10 for all participants according to their role, and this is followed by findings from thematic analysis.

Taking an overview of the judgement-making process, it appeared that some 53% found the 'learning environment' the easiest to judge. This was demonstrated by responses from both teacher educators and mentor teachers, with no area of consensus emerging for the associate tutors. This was due to the visibility of cues to the evaluator, such as conditions, group settings, and relationships. Moreover, it is likely that the material conditions are constitutively easier to assess than, say, the dispositional, epistemic, or relational characteristics of what is, after all, a highly complex setting. One participant noted it was the due to the subjectivity of their own personal preferences. Interestingly, all teacher educators agreed on the easiest item. There was less consensus regarding the area which was most difficult to judge; a high degree of variability emerged with some areas of weak agreement; 29.4% indicated 'assessment' was most difficult, followed by 'content' (23.5%). Instructional strategies and research were not considered either most difficult or easiest. Assessment was the dimension mentor teachers found most difficult to judge; for associate tutors it was 'planning & preparation'. These were found difficult to judge due to not having accesses to lesson documents or materials, the short length of the observation, ambiguity of the learning aims and classroom context, and, as stated by participants, lack of expertise of the evaluator themselves in the discipline being observed.

In the questionnaire, participants were next presented with the prompt 'When making judgements on teaching effectiveness, I ...'; they were given four options to select from, based on prior research regarding judgement-making (see Table 5.11). The table shows the starting point for making a judgement, showing that a majority of evaluators assessed the teaching demonstrated according to learning outcomes based on teaching standards. This was indicated by 8 of 17 participants (47.0%). The second most common rationale was to look for strengths first and then weigh these against identified weaknesses, reflecting on whether the positives are more important than the negatives. This was used by 5 out of 17 participants (29.4%). No evaluators started from a point of failure and looked for instances to challenge that decision.

Table 5.11

Starting Point for Participants' Judgement-Making

	Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)	Overall (n = 17)
Start from a point of failure and look for instances to challenge that decision	0	0	0	0
Look for strengths first and then weigh these against identified weaknesses, reflecting on if the positives are more important than the negatives	1	0	4	5

Consider the teaching demonstrated against the learning outcomes based on teaching standards	1	3	4	8
Other	3	1	0	4

Note. Questionnaire: Q11.

Four participants selected the 'other' option; this included three out of four university teacher educators. The qualitative comments were reviewed and the participants' responses were summarized. One participant (a teacher educator) indicated that judgement is grounded in the teacher's ability to select, use, and apply content appropriately in learning contexts. To make a judgement, the cues looked for included choices and techniques, aesthetic awareness and high standards (e.g., of pupil work), crossing disciplinary thresholds in a way that allows reframing of problematic or challenging knowledge, evidence of dialogic assessment that empowers pupils, and the ability to bring to life ongoing history and development of the world and its humans. Another teacher educator stated that 'university staff should not make judgements based on individual lesson observations; the teacher who works alongside the student is best placed to assess performance'. A third teacher educator noted that 'judgements are made on a range of things, but a significant dimension is professional judgement that involves experience and wisdom from 30 years of education'. Also noted was a strong element of subjectivity. The fourth participant who selected 'other' (an associate tutor) indicated they consider the standards first, as this is the shared expectation, then look for strengths and consider whether they outweigh the negatives, as well as thinking about what can be built on to address the negatives.

5.3.6 Results: Views on Judgement-Making

The second part of data collection included a questionnaire regarding aspects of judgement-making and influencing factors derived from prior research (see expanded results by role in Appendix A5.5). Participants were asked to rate their level of agreement or disagreement with statements about judging teaching effectiveness. These items were rated on a 7-point scale from strongly agree (7) to strongly disagree (1), with a neutral option (4). The responses to the Likert scale items are summarized in Table 5.12.

Table 5.12

Participants' Level Of Agreement With Statements Related to Judging Teaching Effectiveness

	Teacher educators (n = 5)		Associate tutors (n = 4)		Mentor teachers (n = 8)		Overall (n = 17)	
Statement	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Q12a. It is important that judgements of teaching effectiveness are accurate.	5.80	1.17	6.50	0.50	6.50	0.50	6.29	0.82

Q12b. It is important that judgements of teaching effectiveness are consistent.	6.20	0.75	7.00	0.00	6.25	0.66	6.41	0.69
Q12c. It is important that different evaluators reach consensus.	5.80	0.40	6.75	0.43	5.63	0.48	5.94	0.64
Q12d. It is important that evaluators use evidence to make judgements.	6.80	0.40	6.75	0.43	6.75	0.43	6.76	0.42
Q12e. It is important that professional judgement is used when judging teaching effectiveness.	6.60	0.49	6.50	0.50	6.50	0.71	6.53	0.60
Q13a. It is important that judgements about teaching effectiveness are made by more than one evaluator.	6.00	0.89	5.50	1.12	6.25	0.66	6.00	0.97
Q13b. It is important that potential sources of evaluator error are addressed.	6.80	0.40	6.50	0.50	6.13	1.05	6.41	0.84
Q13c. It is important for the teacher to understand how judgements about their teaching effectiveness are made.	6.80	0.40	6.75	0.43	6.88	0.33	6.82	0.38
Q13d. Judgements are always related to particular teachers at particular points in time and in particular situations.	5.00	1.10	6.75	0.43	4.63	1.58	5.24	1.51
Q13e. It is important that judgements about teaching effectiveness are considered fair by stakeholders.	6.80	0.40	6.75	0.43	5.75	1.39	6.29	1.13

Note. Questionnaire: Q12–13; 7 = strongly agree and 1 = strongly disagree.

Overall, there seems to be a high level of agreement among all participant groups on the importance of several aspects of judging teaching effectiveness. There were no areas in which

participants noted disagreement. All groups strongly agreed on the importance of accurate, consistent, and evidence-based judgements of teaching effectiveness. This is indicated by high mean scores (close to 7) and low standard deviations, in particular for questions Q12a, Q12b, and Q12d across all groups. The use of evidence to make judgements (Q12d) had the highest agreement rating. While there was general agreement on the importance of consensus among evaluators (Q12c), the level of agreement was slightly lower compared to other items. The role of professional judgement (Q12e) seemed to be valued across all groups. The importance of having multiple evaluators (Q13a) and addressing potential sources of evaluator error (Q13b) was also generally agreed on, although there was slightly more variation in opinions among associate tutors. There was strong agreement on the importance of the individual being evaluated understanding the evaluation process (Q13c) and also that the judgements made are considered fair (Q13e). The item with lowest agreement was related to judgements being about particular points in time and in particular situations (Q13d); this item was rated between neutral and somewhat agree. The view that judgements are context specific (Q13d) seemed to be more strongly held by associate tutors and mentor teachers compared to teacher educators. Associate tutors tended to have slightly higher agreement scores on most items compared to other groups; in fact, the only occurrence of perfect agreement occurred for associate tutors in relation to the importance of judgements being consistent (Q12b). Mentor teachers showed a wider range of opinions on some items (indicated by higher standard deviations), particularly regarding the contextual nature of judgements and the importance of fairness.

While there was general agreement, some differences emerged when comparing the groups. Associate tutors tended to have the highest agreement scores across most items, indicating a strong emphasis on the importance of evaluation criteria. Teacher educators, while also showing strong agreement, exhibited scores slightly lower than associate tutors, suggesting a more nuanced perspective on certain aspects of judgements. Mentor teachers exhibited a wider range of opinions on some items, particularly regarding the contextual nature of judgements and the importance of stakeholder fairness, which might reflect their practical experience in different teaching contexts. Overall, the data suggest a strong consensus on the core principles of judging teaching effectiveness with some nuances in the opinions of different groups.

5.3.7 Questionnaire Results: Agreement on Influencing Factors

Participants were further asked to rate their level of agreement or disagreement regarding factors which may influence how evaluators judge. These items were rated on a 7-point scale from strongly agree (7) to strongly disagree (1), with a neutral option (4). The responses to the Likert scale items are summarized in Table 5.13.

Table 5.13

Participants' Level of Agreement With Statements Related to Factors Influencing Judgement

Judgement-making is influenced by …	Teacher educators (n = 5)		Associate tutors (n = 4)		Mentor teachers (n = 8)		Overall (n = 17)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Q14a. Clarity of the judgement criteria	6.60	0.49	6.75	0.43	6.00	1.00	6.35	0.84
Q14b. Tension of using judgements for both professional growth and accountability	5.20	0.98	5.75	1.09	5.50	0.87	5.47	0.98
Q14c. Clarity of procedures for making judgements	6.40	0.49	6.25	1.30	6.50	0.50	6.41	0.77
Q14d. Individual understanding of effective teaching	6.40	0.49	6.75	0.43	6.25	0.83	6.41	0.69
Q14e. Contested nature of what defines effective teaching	5.40	1.36	5.75	0.43	5.88	0.78	5.71	0.96
Q14f. Professional teaching standards	6.60	0.49	6.25	1.30	6.50	0.71	6.47	0.85
Q14g. Power relationships between universities and schools in teacher education	4.20	0.40	4.75	1.30	4.38	1.80	4.41	1.42
Q14h. Personal intuition about what happens in a classroom	5.20	0.75	5.50	1.50	5.13	1.69	5.24	1.44
Q14i. Perceived levels of importance of different dimensions of teaching	5.20	1.17	5.25	0.83	4.63	1.65	4.94	1.39
Q14j. Complexity of the classroom environment in which judgements are made	6.00	0.63	6.25	0.83	5.88	1.05	6.00	0.91
Q15a. Evaluator tendencies toward leniency or severity	6.00	0.63	4.50	0.50	5.25	1.39	5.29	1.18

Q15b. Personal biases and beliefs of the evaluator	6.20	0.75	4.00	1.87	5.13	1.69	5.18	1.72
Q15c. Experiences of the evaluator from observing other teachers	6.20	0.40	5.00	0.71	5.13	1.27	5.41	1.09
Q15d. Prior interactions between the teacher and the evaluator	5.20	1.47	4.00	1.00	4.50	1.50	4.59	1.46
Q15e. Holding a pre-observation discussion	5.60	1.02	5.25	0.83	5.75	0.97	5.59	0.97
Q15f. Level of involvement of the individual being evaluated in the judgement process	6.20	0.75	5.00	0.71	5.50	0.50	5.59	0.77
Q15g. Training of evaluators to use observation criteria for making judgements	6.20	0.98	6.00	1.00	6.38	0.48	6.24	0.81
Q15h. Observation skills of the evaluator	6.60	0.49	6.00	1.00	6.25	0.19	6.29	0.67
Q15i. Perceptual information (cues) available to the evaluator	6.20	0.40	5.00	1.00	5.88	0.93	5.76	0.94
Q15j. Policies regarding evaluation of teaching effectiveness	6.20	0.40	5.50	0.87	5.50	1.00	5.71	0.89
Q15k. Quality of the reasoning strategies used to make decisions	6.40	0.49	5.75	0.43	5.63	0.86	5.88	0.76

Note. Questionnaire: Q14–15; 7 = strongly agree and 1 = strongly disagree.

There was general agreement across all groups that the factors presented influence the judgements of teaching effectiveness. This is reflected in the relatively high mean scores for individual items when considered for each group and overall. There was strong agreement on the importance of clear judgement criteria (Q14a) and procedures (Q14c), as well as the significance of professional teaching standards (Q14f), which was recognized by all groups. The importance of evaluator training (Q15g) and observation skills (Q15h) was also acknowledged.

There were also a number of factors with varied agreement. Items related to personal intuition (Q14h), personal biases (Q15b), and the complexity of the classroom environment (Q14j) showed more variation, suggesting that these factors are perceived differently by

different groups. The influence of power relationships between universities and schools (Q14g) was seen as an influencing factor mainly by associate tutors and mentor teachers, and less so by teacher educators. Evaluator tendencies towards leniency or severity (Q15a) and prior interactions between teacher and evaluator (Q15d) also showed significant differences between groups.

Teacher educators tended to have higher agreement scores on items related to clarity, structure, and professional standards. Associate tutors showed more variation in responses, particularly on items related to power dynamics, personal biases, and evaluator behaviour. Mentor teachers had a more nuanced view of several factors, with higher agreement on some items and lower agreement or neutrality on others compared to the other groups. The item with the highest agreement involved judgement criteria and standards, and the lowest agreement was related to the influence of power dynamics. The data suggest that while there is a shared understanding of some key factors influencing judgement of teaching effectiveness, there are also significant differences in perspectives among the three groups. This highlights the complexity of the evaluation process and the need to consider multiple perspectives when developing and implementing evaluation systems.

While there is general consensus across all three groups on the importance of several factors influencing judgement of teaching effectiveness, there was some variation in responses. Teacher educators tended to have a stronger focus on the formal aspects of evaluation, such as clear criteria, procedures, and professional standards. They also appeared to place a higher value on evaluator skills and the influence of training. Associate tutors expressed a more critical perspective on the evaluation process, highlighting issues such as power dynamics, personal biases, and the complexity of the classroom environment. Mentor teachers' responses appeared to reflect their practical experience in classrooms, as they had a high level of agreement with the importance of factors related to the classroom context, such as personal intuition and the complexity of the teaching environment. They also showed a more balanced perspective on various aspects of evaluation.

5.3.4 Questionnaire Results: Why Consistent and Reliable Judgements Matter

Finally, participants were presented with an open-ended question related to the overall research aim. They were directly asked: 'Why does it matter that judgements of teaching effectiveness are consistent and reliable?' Responses were analysed using the constant comparative method (Glaser & Strauss, 1967) to develop themes, and findings are presented in Table 5.14. The question was purposely presented after the video task, which engaged participants in a judgement-making exercise and included a questionnaire that considered influencing factors on judgements.

Table 5.14

Participants' Reasons for Why Consistent and Reliable Judgements Matter

Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)	Overall themes (n = 17)
• Fairness (3) • Standards in education can be undermined (2) • Implications of results (1) • Setting a minimum level of competency (1)	• Fairness (2) • Consistency of achieving standards when coming into the profession (2) • Implications of results (1) • Credibility of the profession (1) • Professional obligation (1) • Moral obligation (1) • Provide support for growth (1) • Quality workforce (1)	• Fairness (5) • Maintaining standards in education (3) • Fostering improvement and growth (3) • Setting an expected level of competency across the profession (2) • Informs areas for support (2) • Protecting the profession (1) • Implication of results (1) • Agreed criteria are adhered to (1) • Consistency (1) • New teacher confidence (1)	• Fairness (10) • Standards (5) • Competency (4) • Growth (3) • Support (2) • Credibility (1) • Consistency (1) • Confidence (1)

Note. Questionnaire: Q16.

5.3.4.1 Fairness and Equity

The most frequent theme to emerge was fairness, which was mentioned multiple times across all three groups. One teacher educator noted that 'it comes down to fairness for the student teacher', and another stated that 'all teachers [are] judged against the same standards'. A third teacher educator observed: 'The stakes are high; judgements need to be fair and supported by evidence.' Fairness was confirmed as a reason by associate tutors as well. As one participant articulated: 'In order that all students are treated fairly, no one is favoured or disadvantaged.' Another associate tutor expressed that it is important to, 'make it equivalent and fair for all teachers'. The theme of fairness was corroborated by mentor teachers, who stated: 'This is important to allow for ineffective teachers not to progress to probation and in doing so show that judgements made have been made fairly and against agreed criteria'; 'To make it fair and

consistent to all that are involved where the teacher being evaluated or the person who is doing the judgements'; and 'It's quite an unfair process at times and there is room for improvement.' Notably, another mentor teacher said:

> Individual teachers might be judged differently by different judges, and this can be unfair but also might lead to a huge variety of teachers being deemed effective. Overall, students, parents and schools need to be confident that the teachers are effective so there needs to be a standard which is met.

5.3.4.2 The Profession

Consistent and reliable judgements of teaching effectiveness are also seen as important for protecting the profession's credibility and setting an expected level of competency for new teachers to maintain standards in education. One teacher educator noted that judgements matter because 'overall, standards in education risk being seriously undermined'. This rationale was confirmed by associate tutors who stated: 'For the credibility of the profession. To ensure there is a consistency of standards coming into the profession'; and 'to ensure there is consistency in achieving the standard required'. In agreement, mentor teachers noted: 'The standards expected of teachers in the first few lessons/months/years seem to be the same as the standards being achieved by experienced teachers.'

Participants indicated that when judgements are seen as fair and accurate, it helps to maintain the public's trust in the teaching profession and to encourage new teachers to enter and stay in teaching. As one teacher educator said: 'The stakes are high; judgements need to be fair and supported by evidence.' Another stated: 'Because we are dealing with people's lives here. We need to get it right!' These sentiments were reaffirmed by other participants. An associate tutor stated:

> Professionally and morally, I think it would be remiss to allow a student who is less than effective to continue their journey without a professional discussion offering support/advice to improve their experience and that of the pupils they will teach.

Further contributing to this sentiment, mentor teachers acknowledged it matters in order, 'to maintain professional standards and to ensure NQTs [newly qualified teachers] entering the profession are both capable and confident', as well as 'to keep standards the same for everyone'.

5.3.4.3 Continued Professional Learning

Informing areas for supporting growth and continuous professional learning, both for new teachers and mentor teachers, emerged as a theme identified by both associate tutors and mentor teachers. This theme revealed a dual purpose of judgements: to determine effectiveness of teaching and also to support continual professional development (CPD), which over time fosters new teacher confidence and competence. A mentor teacher noted judgements matter: 'In order that student teachers get clarity of understanding of their strengths and areas for development. To help student teachers and their mentors see improvement and growth'.

5.3.4.4 Comparison Among Evaluators

Overall, the data suggest that there is a general agreement among all three groups about the importance of consistent and reliable judgements of teaching effectiveness. These themes encompass the need to avoid biases, discrepancies, and unfairness in the evaluation process, promoting a standardized and equitable approach. Consistent evaluations contribute to establishing and upholding a baseline of competence within the teaching profession. The need for fairness was noted in relation to both the individual being judged and the one doing the judging. These themes all point to the importance of consistent and reliable judgements to ensure teachers are meeting the necessary standards and receiving the support they need to improve their practice.

There were also some slight differences in how the groups view the significance of fair judgements. Consistent and reliable judgements of teaching effectiveness were deemed important because student teachers should be treated with equity (i.e., in a way that is right and reasonable). Teacher educators seemed to be more focused on setting standards to ensure fairness. Fairness was seen as crucial since results of judgements set a standard for minimum level of competency for educators entering the profession, and the professional standard for teachers could otherwise be undermined

Associate tutors appeared to be more focused on growth and maintaining standards. Participants noted it is important that judgements of teaching effectiveness are consistent and reliable, because there is a professional and moral obligation to ensure a quality workforce. Fairness is important because results of judgements establish credibility of the professional and consistency of teaching, which impacts on pupil learning. Additionally, they identified the need to recognize areas to support development of new teachers.

Mentor teachers also highlighted the importance of fair and consistent assessment to ensure new teachers are judged against the same criteria. It was seen as important that judgements with regard to teaching effectiveness are consistent and reliable because it helps mentors and student teachers identify specific areas for improvement and track progress over time. Mentor teachers foregrounded the importance of consistent evaluations for fostering improvement and growth. Ensuring consistency in judgements was seen as a way to prevent ineffective teachers from progressing to probation or advancing within the profession without meeting the necessary criteria. This safeguards the quality of education provided to pupils. There was a recognition that without consistency, there is potential for misunderstandings about the required standards, leading to ineffective judgements and potentially compromising pupil learning. Clarity in criteria and expectations helps in making assessments more effective. Acknowledging the potential for growth, some mentor teachers suggested a longer preparation period and additional support for new teachers. Mentor teachers also highlighted that high standards achieved in specific observed lessons might not be representative of overall teaching abilities, which develop over time. Mentor teachers expressed concern about the impact of inconsistent evaluations on student teachers, emphasizing the potential advantages or disadvantages. They also noted variations in placements and argued for a more

transparent and equitable evaluation process. Mentor teachers emphasized the need for fair and consistent evaluation to prevent biases or discrepancies in judgements. Consistency ensures that all new teachers, regardless of experience or background, are evaluated using the same standards. This concluding question allowed us to get a clearer picture of the significance of consistency and reliability and potential consequences if there is inconsistency in judgements or the results of judgements are unreliable.

5.4 Focus Group Results

Focus groups and individual interviews were carried out to facilitate discussion concerning results of the video observation task and to corroborate judgement strategies and rationales identified through initial analysis as presented in this chapter. Detailed methods are provided in Chapter 2 (see Appendix A2.3 for the case study protocol). Data included responses from a small group of teacher educators (n = 4), an associate tutor (n = 1), and mentor teachers (n = 4), all actively involved in preparing future teachers as per their various roles at the time of the study. Results from the constant comparative method of analysis for each of the questions are presented next.

5.4.1 Reasons for Consistency or Inconsistency

Participants were first asked, in relation to judging teaching effectiveness: 'What could be reasons for consistencies (or inconsistencies) between raters?' Responses for each group are presented in Table 5.15. Participants provided many possible reasons for the inconsistency among raters captured by the questionnaire, and these were organized into four categories.

According to four teacher educators who participated in the focus groups, there were many (n = 29) potential reasons why there could be a great deal of inconsistency in judging the quality of a student teachers' teaching. These were organized into four main categories to explain the inconsistencies found among participants in the video task and questionnaire. These focused on the evaluator, the student teacher, aspects of the teaching lesson being observed, and processes associated with making judgements of teaching practice.

As one teacher educator stated, 'There might be very different understandings and interpretations of the assessment criteria and that's bound to result in widely divergent opinions and judgements' (F1I1). Another teacher educator articulated:

> I kept coming back to the fact that I don't know what those learners should be learning. I don't know how engaged they are. I don't know learning about them. I don't know anything about the teacher. I don't know key elements of the lesson. And it's very difficult to gauge, or difficult for me to gauge. (F1I2)

The one associate tutor who participated also provided possible reasons for the inconsistency among raters captured by the questionnaire, and these were organized into two categories focused on variability of the evaluator and their attention, focus, mindset, and likes. The associate tutor stated:

> I think there's very much a gut reaction. You can go in there and you can get a gut reaction very, very quickly as to whether or not something is good or effective. And I

> suppose it depends on what the focus is … if people are focusing very much on just what the teacher is saying, but ignoring the fact that there's maybe disruption going on there or people aren't listening. (F16I1)

According to the three mentor teachers who agreed to a focus group and one who took part in an individual interview, there were 42 potential reasons why there could be a great deal of inconsistency in judging the quality of a student teachers teaching. These were organized into four main categories to explain the inconsistencies among mentor teachers who judge student teachers' effectiveness through classroom observation. These focus on the evaluator, the student teacher, aspects of the teaching lesson being observed, and processes associated with making judgements of teaching practice. The focus was on the variability of the mentor teacher as an evaluator, the shared understanding of what constitutes good teaching, and the process by which a valid judgement is made. One mentor teacher noted, 'the effectiveness of a lesson will always be open to a fair degree of interpretation' (I12I1). Another mentor teacher acknowledged:

> I could have been wrong because I was making assumptions based on how I saw the learners reacting. And maybe somebody else with a different kind of mindset would not want to make that kind of presumption. They may even, it may even have been a different teacher who did that learning and not the one that we were watching. (F4I1)

Results highlight the complex interplay of factors influencing consistency and inconsistency in judging teaching effectiveness. Evaluator-centric factors, such as knowledge, experience, and biases, play a significant role. Participants also noted that observation conditions, including clarity of focus and availability of information, also impact judgement accuracy. Process-related factors, such as clear criteria and calibration of raters, appear significant.

Table 5.15

Reasons for Consistency and Inconsistency Between Raters

Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)
Evaluator centred: • Evaluator's understanding of judgement criteria • Personal disposition of the evaluator • Prior experience making judgements • Experience of the evaluator from judging other teachers (both high- and low-quality teaching) • Physical state of the assessor • Evaluator's knowledge of the subject/content • Own prior experiences as a learner • Evaluator's ability to assess • Considering how others would rate • Multiple raters • Evaluator's views and biases • Different interpretations of theory • Observation skills of explicit and inferential cues • Different ideas on what constitutes good practice • Mistakes can be made **Aspects of the teaching observation:**	**Evaluator centred:** • Mindset people are maybe going in with • What they're actually looking for • There's very much a gut reaction • Teaching approaches the evaluator likes **Aspects of the teaching observation:** • The cues that are being attended to • Depends on what the focus is	**Evaluator centred:** • Different perceptions of what constitutes good practice • Background • Experience • How up to date the mentor teacher is with educational research • Continuous reshaping of what good teaching looks like • Subjectivity • What the evaluator focuses on • Personality • Tendency towards severity or leniency • Capacity to deliver given a negative judgement • Bias • Outside input from other staff • Difference in training of the mentor and the student teacher • Personal expectations • Viewpoint of evaluation as an inspection • Training of the mentor teacher in specific areas of teaching • Individual view of the rater

- Understanding the whole context in which the teaching occurs
- Cues available to the evaluator during a short lesson observation (i.e., pupil engagement levels)

Student teacher centred:

- Natural variability in performance
- Student response to high-stakes evaluation with a dual purpose of growth and accountability
- Evaluator's relationship with/knowledge of the student teacher
- Relationship with student and mentor teacher

Processes:

- Clarity of expectations (criteria) for making the judgement
- Different definitions of consistency
- Specific aspects of teaching in an observation that can be judged or not
- Different ways of doing things
- Descriptors of quality are difficult to articulate
- Clear aim of the evaluation (what you see or holistic performance)
- Tension of judgement for both growth and accountability
- Mistakes can be made
- Making assumptions based on how the learners reacted
- Mindset of reviewers
- Personal preferences as to how to teach
- Individual nature of judgement

Aspects of the teaching observation:

- Pupil learning is not always visible
- Some aspects are more observable than others
- Dynamics with pupils
- Learners' interactions with the student teacher
- Pupils' learning
- Subject area being observed

Student teacher centred:

- Disconnect between what student teacher has learned and what mentor has learned
- Level of relationship with the student teacher
- Relationship with pupils
- Knowing how to respond and apply skills
- Taking the lead from the student teacher themself

Process:

- Shared understanding of good practice between mentor teacher and university
- Vague standards
- Lack of written criteria

• Implications of the use of the evaluation results	• Judgements are quite qualitative rather than quantitative • A degree of interpretation is inherent • Lack of calibration of what constitutes good practice • Variability in school contexts • Using just 1 lesson to make a judgement • Using criteria

Note. Focus group: Q1.

5.4.2 Possible Ways to Gain Consistency

Next, participants were presented with the question: 'What would make judgement among evaluators more consistent?' Participants provided several suggestions as to how greater consistency among judgements of teaching effectiveness could be gained. Responses according to each group are presented in Table 5.16.

Teacher educators in the focus groups suggested several ways that judgements might be made more consistent. Suggestions encapsulated the full judgement experience, including preparations that occur before making a judgement, the reason by the judgement occurs, what is being judged, how the judgement is actually being made, and how the results of the judgement are communicated. One participant questioned whether consistency is an appropriate goal. One teacher educator stated: 'if consistency is what you're after, then you just reduce the assessment form to a tick box. But that invalidates the whole process. It makes the process redundant and worthless, certainly less meaningful' (F1I1). The participant further added:

> I think these are very high-level skills that we, as teacher educators generally have. It's a lot of tacit knowledge we have about what needs to be said and how it needs to be said in a supportive way. But it doesn't lend itself necessarily to this kind of will o wisp of consistency.

Another teacher educator (F2I1) said, 'having fewer criteria would help … sometimes the forms are quite complex with lots of different subcategories, whether fewer would help to be more focused and maybe some of the big picture issues'. The participant also suggested

> moving away from that kind of one-off visit to successions of smaller interventions … with fewer criteria … they still have to be at a particular standard, but maybe the journey to get to that standard could be done in a more nuanced way.

The associate tutor contributed three potential ways to make improvements in processes and criteria, all which were reflected in suggestions from the other groups. They articulated in the interview:

> there's a consistency of expectation on our part then as well. So, what we think is effective, you know, and I think it is very important to realize that that will be different. Because a second year who is out of school 2 years and is 19 years old or something in a classroom
> What you would expect to see as an effective lesson from that student in second year would be different from what you would expect to see from a student in fourth year … maybe
> something a wee bit more prescriptive might help to ensure that consistency. (F16I1)

Table 5.16

Strategies to Gain Consistency in Judging Teaching Effectiveness

Teacher educators (*n* = 5)	Associate tutors (*n* = 4)	Mentor teachers (*n* = 8)
Preparation: • Tacit knowledge of a teacher educator making a judgement being made explicit • Ensure prior teaching experience of the evaluator matches the context in which they are making judgements (primary to primary and secondary to secondary) **Why the judgement is being made:** • Ensure value of the evaluation for professional learning • Change how the results of judgements are used (growth versus accountability) **What is being judged:** • Not judging content knowledge during the observation • Limiting which aspects of teaching are judged through observation and when during teaching education • Fewer criteria **Making the judgement:** • Tacit knowledge being used by evaluators to make judgements	**Preparation:** • Agreeing key features being observed without being too prescriptive • Understanding of expectations at difference levels of progression in teacher development **What is being judged:** • Providing guidance regarding 'look fors' (descriptors)	**Preparation:** • Opportunities for more formal dialogue for everybody involved **Why the judgement is being made:** • Clear aim of making the judgement **What is being judged:** • Clear expected level of proficiency for a novice teacher (e.g., first placement or the last) • Criteria for making the judgement **Making the judgement:** • Include student teacher explanations of their practice • Moving away from sticking a number on something **Communicating judgements:** • Discussion with the student teacher after the lesson

- Use input from the student teachers
- Use clear criteria for making judgements
- Trust the judgement of school-based mentor teacher

Communicating judgements:

- Using more than just one-word summaries
- Knowledge sharing as a part of assessment (feedback)
- Improved forms for capturing judgements

Note. Focus group: Q2.

Mentor teachers offered a number of ways judgements might be made more consistent; these included similar suggestions to the teacher educations encompassing the same five thematic areas. One mentor teacher noted that 'one way that a teacher could actually get a better perception would be to speak to the student teacher first' (F9I1), and another suggested to, 'have an opportunity for different mentor teachers to be given time to discuss with other mentors or even with partners in the university or educational professionals elsewhere … for more formal dialogue for everybody involved in the process' (I12I1). Overall, participants suggested consistency could be gained by clarifying and standardizing the evaluation process, enhancing evaluator preparation and development, and shifting the focus of evaluation to be more holistic and student teacher centred along with improved communication and feedback.

5.4.3 Perceptions of Inconsistencies from Video Task Results

The third question posed to participants related to initial findings from the video task. This queried perspectives regarding the domain of teaching which yielded the greatest degree of variation of ratings. The question was: 'What are your thoughts on the finding that [name of domain] had the most inconsistent rating?'

Teacher educators noted the following views on the finding that 'learning environment' had the most inconsistent rating and also was considered the easiest to rate:

- exemplified the difference in individual understanding of effective teaching
- personal beliefs and bias of the evaluator are impactful on consistency
- showed a need of rationales/basis for judgements
- demonstrated that multiple perspectives of the evaluator come into place, as a teacher or learner or teacher educator
- consideration of the influence of content of the lesson
- clear criteria for making judgements are needed
- available cues are used as evidence
- affective components involving many different physical, sensory, and emotive cues

The teacher educators articulated that with personal beliefs, perspectives, and bias apparent, there is a need for clear criteria for making judgements and using criteria to provide rationales. They affirmed that judges should consider the influence of the content of the lesson and available cues, including the influence of affective and sensory cues, for making judgements. One teacher educator noted, 'sometimes maybe it cannot be verbalized, how you feel when you're somewhere' (F2I2). Another participant stated:

> I think that highlights just how very different people's views are about what constitutes a good learning environment. I'm not really surprised...that illustrates the need to dig [for] deeper insights and say, 'okay, why did you think that was a good or a bad learning environment?' (F1I1)

For the associate tutors who completed the video task, there was consistency in the domain of 'instructional strategies' being identified as the easiest to judge and the domain of 'learners' being the most difficult. The one associate tutor who was interviewed noted: 'I think the issue is that we all have different strides, we can all identify things maybe a wee bit easier than

others. Some people find the academia and the theory side of things far easier than the practical' (F16I1). The following thoughts on consistency and potential for variability of responses were shared.

- differences in individuals' experiences
- consideration of relationships and classroom interactions
- fine details that are being observed
- observation skills of the evaluator
- evaluators' own strengths and weaknesses as an educator
- engaging in the classroom with pupils during the observation
- formative assessment of the teacher candidate
- some aspects of teaching are much more evident than others (if used or if not used)
- the content being taught and the match with the content area of the evaluator

There was no clear pattern for which domain mentor teachers found the easiest and most difficult to rate. This variability was shared with the mentor teachers interviewed, who then shared the following thoughts on what might be the reason for the finding:

- confidence of the mentor teacher
- experience of the mentor teacher (e.g., number of student teachers observed, number of classes observed)
- emotional intelligence
- inference/interpretation skills of the evaluator
- difference in individual understanding of effective teaching
- effect of the evaluator being out of their subject area

Participants articulated that this finding regarding high variability of what was most and least difficult to evaluate could be due to confidence, experience, emotional intelligence, inferencing skills, and subject area of the evaluator, which is ultimately reflected in individual understanding of what constitutes effective teaching. To exemplify these views, one mentor teacher indicated a number of reasons:

> That's to do with confidence and experience of the mentor. I think you will find things, presumably, it's to do with emotional intelligence and what it is that you think you can see in something. I think every single individual probably would find something the most difficult and the least difficult, just really depending. And that's got to be so individual based on your experience, the number of student teachers you've seen, the number of classes you've seen, and also maybe being out of your subject area. And, you know, that might be make a difficulty, a barrier that you wouldn't have in your own subject area. (F4I1)

Based on these findings, several strategies were put forward to improve consistency in rating teaching effectiveness. This included developing clear and shared criteria for evaluating different domains of teaching, and providing comprehensive training for evaluators on observation skills, evaluation criteria, and bias awareness. Additionally, a focus on

observable behaviours while acknowledging the importance of subjective elements and considering the impact of content knowledge on ratings was acknowledged.

5.4.4 Professional Judgement and Professional Standards

The fourth focus group/interview question participants were asked was: 'What are your views about using professional judgement and professional standards to judge teaching effectiveness?' Responses for each group are presented in Table 5.17.

According to the teacher educators, there is a need to use both professional standards and professional judgement when judging practices of student teachers. For example, one participant shared: 'professional standards are very important as a part of a profession; we do need to have standards to refer to and to justify our judgements. We make our judgements in reference to a shared set of standards' (F1I1). The focus group discussions revealed the view that standards provide the basis and rationale for the judgements that are made, and these judgements are made by experienced and vetted professionals who know what the standards are and can contextualize their application. As one participant stated:

> implicit in being a professional is a sense of not so much autonomy but a sort of freedom to make a decision that's the right decision as far as you can judge it given the circumstances … we're not autonomous in that strict sense of the word because we are bound by professional codes of conduct. But there must be space somewhere to say, right, given the circumstances, given extraneous factors that we didn't see coming, this student has performed well. (F2I1)

During any lesson observation, some standards may not be observable or met, and a good evaluation of whether someone is meeting the standards requires collaboration among professionals who are providing teacher education. Although standards are very important, participants found these to ultimately be an ideal without a shared understanding, with the target of attainment and at what level remaining ambiguous. A tick-box summary of whether a standard is met or not is insufficient when this is used as accountability measures. Some elements of teaching require professional judgement, especially those that are non-negotiable and demand consistency (i.e., affective criteria such as integrity and respect); others can be more nuanced.

The associate tutor brought forward the complex interplay between professional judgement and formal criteria when evaluating teaching effectiveness. This response emphasized the importance of a nuanced approach to evaluation that combines the strengths of experienced-based judgement and formalized criteria. The associate tutor noted that 'professional judgement just comes down to, more to do with experience sometimes', and that 'professional judgement is the gut reaction', and 'you kind of confirm that judgement by using the professional standards' (F16I1).

Table 5.17

Participants' Views on Professional Judgement and Professional Standards

	Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)
Standards	• Are very important • Are an ideal • Referred to when making judgements • Used to justify judgements • Allow for context and professional judgement in the process • Do not include how well or to what extent • Clear target of attainment of the standard is needed (level) • Tick-box summaries are insufficient • Some standards are observable in a lesson observation and others are not • Purpose of standards for accountability and gatekeeping • Items that require professional judgement • Some standards are non-negotiable and require consistency; others can be more nuanced • Some criteria will always not be met • Affective criteria are the basis for good pedagogy	• Can be used to confirm/support professional judgement • Can help minimize bias	• Open for interpretation • Helpful to have • Help students know what they are being evaluated on • Some standards are observable in a lesson observation and others are not • A benchmark • Give an insight into what we should be looking for • Fairly useful when completing placement reports • Closest thing to a sort of checklist for overall competency • Decent way of measuring • Gives an idea of what to look for • Used as a guide • Are not looked at often by teachers; don't have enough time to consult the standards • Used for self-evaluating • Once they (teachers) get qualified, there is a sort of dip in performance • SPR [*Standard for Provisional Registration*] is what we would judge the student teachers against • Quite a comprehensive guideline • Helpful to look at the progression from the SPR to full registration and then continuing professional development

	• Clarity in descriptors for a shared understanding		• Fantastic documents for a student teacher or for any teacher actually to look at and understand where they should be • Standards are invaluable to making a judgement • For telling student teachers what they should aspire to in the wider scheme of being a teacher • Helpful if you're having to have challenging conversations with students • Areas to guide and focus on • Need to cover the standards to get full registration • Really effective for us to highlight areas • Students find it really beneficial because then they know what their areas are, kinds of weakness • We don't go too much detail about the standards; more general of the areas that this is what you should be doing and to make sure that you get there • Teachers should try and keep abreast of standards
Professional judgement	• Individual understanding of standards for judgement • Trustworthy subjectivity of the judge • Must be carried out by a vetted professional • Must know what the standards are • Purpose of the judgement being made – growth model or accountability • Requires collaboration in the profession of teacher education	• Equates to 'gut reaction' • Is developed from knowledge from years of observation and experience in schools • Supported by professional standards	• Judging effective teaching is an art rather than a science • Requires relationship of the mentor teacher with the pupils and with the student teacher • Requires experience • Personal beliefs are a part of professional judgement • Requires subject knowledge • Requires the ability to observe skill development over time • Requires knowing what you are looking for (i.e., formative assessment, relationships, etc.) • Reflect own biases

- Influenced what the evaluator wants to see
- Will be different when compared to someone else's
- Consequences of relying on professional judgement are important
- Involves gathering a lot of evidence
- Requires being open about what needs improvement
- When it is off, can hinder a student teacher from progressing
- Need teachers to do the overall judgement
- Requires a double-check
- Experienced teachers particularly are trusting their professional judgement more than others
- [Judgements] quite organic
- They change every so often
- Can be attributed to why inconsistencies arise
- Making judgements at different stages in our own career
- Judgements are constantly trying to reflect new government policy, new educational policy, new and ongoing changes
- A judgement is difficult because what we've been told to look for in a successful lesson is probably ever changing
- Professional judgement is used to know what to bring up with a student teacher, when, and in what degree of depth
- Comes with experience

Need for both	• A balance of objectivity and subjectivity • Professional judgement and a degree of subjectivity is necessary • Consideration of the teacher and the learners • Many ways to interpret consistency • Consideration of context and response to classroom dynamics is vital • How results are used influences how/when both standards and professional judgement are used • Judgement involves the use of standards to make a decision and freedom for professional judgement based on being 'a professional' • Purpose of the judgement being made – growth model or accountability • Unity in diversity

Note. Statements from focus group: Q4.

Regarding using their own professional judgement when judging teaching effectiveness, mentor teachers identified both strengths and challenges. Mentor teachers acknowledged professional teaching standards as a valuable tool, particularly for student teachers. One mentor teacher described their experience working with a student teacher and how both standards and professional judgement were relevant:

> using professional standards. Again, they're open to interpretation. There's lots of them. I think it would be helpful to have … if you're working together and have certain professional standards that they know [what] you're going to be looking for in the planning, they can show you where they think they will display them. And I find that's the most helpful way to use those standards, because you can't look for them all, although you might find that there are other ones that haven't been planned that are being used. (F4I1)

The standards are seen to provide a framework for understanding quality teaching and offer a foundation for evaluating practices. The standards outline a clear path for professional growth and facilitate communication during feedback and challenging conversations. One mentor teacher noted that the 'standard is really invaluable to not just me when I'm making a judgement in a student but also for telling student teachers what they should aspire to in the wider scheme of being a teacher' (I12I1). However, there are challenges associated with their consistent use by experienced teachers, as they may not refer to the standards frequently. Mentor teachers also emphasized the broad areas of the standards rather than specifics; they acknowledged the importance of using professional judgement of experienced teachers and recognized the potential limitations of subjectivity and the need for evidence-based practice.

Collectively, participants viewed the value of professional standards as providing a clear framework for judging teaching and serving as a benchmark to gauge competence. The standards also help student teachers understand expectations and areas for improvement and promote consistency in evaluation across different mentors. The role of professional judgement and the need to take a holistic view and often make quick, intuitive assessments was also brought forward. Professional judgement is also seen as evidence of the value of teaching experience. Professional judgement and standards appear to be seen as complementary to one another, both being essential when judging teaching effectiveness.

5.4.5 Universities and Schools Working Together

The fifth question posed to participants was: 'How might schools and universities work together to gain greater reliability in evaluation teaching effectiveness?' Teacher educators outlined changes to systems, practices, and understandings as ways that schools and universities could work together to gain greater reliability in evaluation teaching effectiveness. The statements that follow were captured in thematic analysis.

Systems change:

- sustained relationships or schools, departments, and specific teachers working with the university
- work with carefully selected hub schools who demonstrate excellence

- small group of students in the subject at hub schools
- university members embedded in schools
- redefine relationships between university and school
- partnered approach

Practices:

- productive communication
    - professional conversations
    - Socratic dialogue
- tutors (university) and classroom teachers working together in a joint process for decision-making
- increase reliance on the mentor teacher to make the judgement
- portfolio of multiple sources of evidence
- joint writing and research

Understandings:

- a shared understanding of what is good practice in subjects
- develop understanding of school-based mentor teachers' role in teacher formation
- reciprocal learning process – mutually beneficial
- schools' understanding of university-based teacher education (not just practical)

Participants shared that schools and universities need to redefine and build a shared understanding of their partnered approach to teacher education. This requires trusting relationships, understanding each other's roles, expectations for good practice, and reciprocal learning. As one teacher educator noted, 'if I have to put my money somewhere, that would be professional respectful conversation built over time making a professional relationship that can discuss safely and respectfully the strengths and areas in need of development' (F1I1). This might be accomplished through sustained relationships with 'carefully selected hub departments' (F1I2) in which small groups of students work together alongside a university staff member who is embedded in the school and teachers who have demonstrated excellence. As one teacher educator noted, 'there's got to be that relationship there' (F2I1). On a practical level, suggestions would require ways of working noted by the participants, such as close communication, joint decision-making, an increased reliance on the classroom mentor teachers, and a portfolio approach to judge teaching effectiveness, which hub departments could facilitate. Joint research and writing with hub schools were also put forward as what could occur in the space of mutually beneficial relationships.

The associate tutor who was interviewed outlined the following ways that schools and universities could work together to gain greater reliability in evaluation of teaching effectiveness:

Practices:

- clear expectations about the experience
- a handbook to reference for each party

- consistency of [associate tutors] implementing high-quality support [for] students and classroom mentor teachers
- regular contact with mentor teachers and students

Understandings:

- disparity between expectations of schools and university
- discussions with the schools
- strengthen relationships

They noted from their personal experience as a former headteacher that: 'There was that kind of idea of the university wanting to pass them, and we just don't think they are capable and you're kind of having to jump through hoops. There's a kind of disparity between the expectations sometimes' (F16I1).

Mentor teachers outlined the following ways that schools and universities work together to gain greater reliability in evaluation teaching effectiveness:

- a half-hour pre-meeting between a tutor and a mentor
- shared understanding of what to look for in different placements
- closer relationship with the school experience tutor and the mentor teacher
- the standards do work as a guide
- could have clip of someone teaching for 10–15 minutes or whatnot, and then it could be highlighted throughout, like, 'this is good practice, this'
- opportunity for more discussions and dialogue
- a kind of moderation approach
- opportunities for people to get involved in CPD opportunities for moderation of teaching standards
- stronger relationships with the university tutors
- communications and reminders
- clear expectations of the student provided
- work together to support students when issues arise

Participants acknowledged the current efforts towards collaboration but do see potential for improvement. They emphasized the importance of clear communication between the university tutor and mentor teacher, shared understanding of expectations for each placement, and a supportive and growth-orientated environment for student teachers. One mentor teacher stated: 'it'd probably be useful just to have maybe closer ties with the university, and certainly a, this is what we're looking for, kind of prior meeting to the placement might be useful' (F9I1). They also recognized the challenges associated with time constraints and logistical hurdles, though they desired opportunities for continued professional development in their role as mentor teachers in ITE. As another participant reflected:

> I think maybe opportunities for people to get involved in CPD opportunities for moderation of teaching standards or student teacher standards, I think, would be helpful just to see how they do it. I wouldn't be confident in saying I know that these standards here in this school and this faculty are even the same in the next

> department, which is 10 metres away, let alone the school that's a mile down the road, so I think that might be something that would help. (I12I1)

Overall, the findings indicate that a strong partnership between schools and universities, characterized by open communication, shared goals, and mutual respect, is essential for improving the reliability of judgements of teaching effectiveness.

5.4.6 Barriers and Assets for Working Together

The sixth focus group/interview question was: 'Is there any barrier or asset you would like to raise attention to that would impact working together?' The barriers and assets identified by participants that impact schools and universities working together for reliable judgements are provided in Table 5.18.

Teacher educators identified a few assets that could influence the consistency and reliability of judgements. These included good relationships in a partnered approach, clear understanding of roles, and processes for addressing disagreements in joint decisions. There were significantly more barriers within systems and practices noted in relation to working together to increase the reliability of judgements. One participant brought forward a potential barrier around a partnered approach:

> something that I've seen, and maybe haven't thought about until you said that you go into a lot of teachers' classrooms and they feel very judged, as if when you're going to see the student, not all, I think the ones that are the most open, you know, are most likely. I feel more relaxed. But sometimes you go into somebody's classroom to watch a student, I'd say more in primary schools rather than secondary schools, I've seen, they feel as if they're [mentor teacher] being judged, as well as the student teacher, their effectiveness. (F2I2)

There was a perceived lack of a shared vision of teacher education and the relationships that are needed to best understand roles and responsibilities of all parties. This is needed so that classroom teachers do not feel judged and so that the affective qualities of good teaching, which are more difficult to judge, can be better evaluated. The goal is promoting positive growth in people; this is very difficult to capture and explain, making a shared vision and good relationship imperative. As one teacher educator reminded, 'the process involves the nurture of human beings' (F1I1).

The associate tutor highlighted a barrier to effective collaboration between university-based teacher educators and school-based practitioners as a perceived lack of practical experience among university staff. This reflected a perceived lack of credibility, particularly among those teacher educators at the university with limited recent classroom experience who may be viewed with scepticism by school-based practitioners. The associate tutor shared:

> I'll tell you something else that comes up, and I shouldn't really say this, right, but there is a feeling, you know, that somebody's retired or they see you coming in. However, if there's people coming in from the university who maybe haven't been in a classroom for 20 years, there's a wee bit of, not doubt, but a bit of cynicism about,

well, how would you know? … And I think it's important that people recognize each other's strengths, but also recognize when they're lacking in a particular area. You know, somebody that's taught for 2 years in a school and then 30 years at university, then they're going to come into a school and folk are going to treat them with a little bit of cynicism, as if, well, how can you come in and tell me how to do it? I think that's the problem, that they're seen as two separate entities. It's how we can merge that together a little bit to get us working together a little bit better. (F16I1)

Table 5.18

Barriers and Assets in Collaboration

	Teacher educators (n = 5)	Associate tutors (n = 4)	Mentor teachers (n = 8)
Barriers	**System:** • Goal is promoting positive growth in people; this is very difficult to capture and explain • Different vision about teacher education • The affective qualities of good teaching are more difficult to judge • Classroom mentor teachers feel judged • Poor relationship between the school and the university • Lack of understanding of roles Practices: • University having non-subject-specialist tutors assessing secondary teachers • Lack of understanding what to look [actionable descriptors) in observation evaluations • Lack of a shared understanding of what constitutes good practice • Different ways to interpret consistency	• Years in university disconnected from schools • Conflict between associate staff and university-based teacher educators ['whose knowledge'] • University teacher education and schools are seen as two separate entities	• Time to be able to work with universities • Government support of collaboration time (more funding needed for release time) • University has the final say in the evaluation • Time • enable teachers to be given time off timetable • Not having the previous placement report • Need more information about the student and their development
Assets	• Good relationship between the school and the university • Clear understanding of roles	• Closeness to the classroom of AT [associate tutor] staff	• Sustained relationships with tutors • Reciprocity in understanding interpretations of judgement criteria

	• Partnership, a joint approach • Processes for addressing disagreement are in place	• Virtual pre-meetings • Students themselves bring up their previous placement and if there have been any issues or any things that [they] need to work on

Note. Focus group: Q6.

There appears to be an implied power imbalance observed by the associate tutor, with school-based practitioners feeling they have more authority or knowledge due to their ongoing classroom experience as well as a perceived gap between the theoretical knowledge of university staff and the practical realities of the classroom.

Furthermore, the mentor teachers consistently noted time as a substantial barrier and that mentor teachers need to be given time off timetable. There were mixed views regarding access to prior reports, but consensus on the need to know more about the student to help contribute to their growth and development. As one participant noted:

> What usually works though is when a student comes in and we speak to them because we always have kind of a certain meeting about like what's your strengths and what is it you want to work on. The students themselves usually bring up their previous placement and if there have been any issues or any things that need to work on, which is good. (F8I1)

Assets to leverage could be virtual pre-meetings and the students' self-reflection and sharing on their targets. Additionally, it was noted that the university has the final say in the evaluation, which indicates the judgement is not truly a joint decision and the collaborative approach is limited in some regards. Assets to leverage include dialogue that builds understanding of judgement criteria and sustained relationships with tutors at specific schools.

5.4.7 Additional Insights

Finally, participants were asked: 'Is there anything you would like to add about reliability and consistency or inconsistency in judging teaching effectiveness from your perspective?' In response, teacher educators shared several final thoughts regarding consistency and reliability in judging teaching effectiveness:

- misgivings around a goal of consistency
- opportunity cost of the tension of using judgements for both professional growth and accountability
- teaching as a sacred duty – a mission statement
- coming together of teacher educators
- defining teacher effectiveness will always be variable
- there are some aspects of teaching that are non-negotiable
- partnered process of making judgements
- need more than one perspective
- there are multiple ways to be effective

One teacher educator stated: 'I think there has to be inconsistency, and I think that's the nature of all knowledge' (F1I2). Another added: 'there's always debate over what we mean by teacher effectiveness and effective teaching; there's always going to be some form of related debate over what that looks like (F2I1). Another teacher educator shared:

> I don't think we spend enough time really focusing on these things that are very valuable, these things that are central to our, to our practice … I would reiterate my misgivings around pursuing consistency. Because I just think that pursuing consistency carries with it an unacceptably high opportunity cost. What you would lose is potentially measurably valuable and unequivocally important compared to what you would [be] gaining in this consistency. (F1I1)

The associate tutor who was interviewed did not add any further comments; however, the mentor teachers further noted:

- different opinions will occur
- bring opinions close together
- a degree of inconsistency is ok
- focus on describing what could be improved
- look for improvements
- avoid diametric opposites
- professional development of the judges is needed
- multiple judges need to understand each other and their rationales

As one mentor teacher stated: 'the people who are doing the judging need to try to understand where each other is coming from and why they're making those judgements' (F4I1). Overall, the complexity and uncertainty of judging teaching effectiveness, and the necessity of a system of making judgements to match this complexity, was emphasized. The nature of teaching as a profession challenges TEPs to reduce disagreement and cope constructively with a shared mission of teacher preparation with school partners.

5.5 Discussion

This case study has explored the nature of judgement-making processes regarding ITE students' teaching effectiveness and illuminated inherent complexities of evaluating teaching quality, as evidenced by the findings from the video task, questionnaire, focus groups, and interviews. Our analysis has underscored critical considerations related to evaluators' roles and responsibilities, the intricacies of assessing student teachers during their preparation, and the influence of the multifaceted nature of consistency in the judgement-making process. These insights are instrumental in addressing the research questions posed in this project and developing informed recommendations.

5.5.1 The Evaluators and Their Task

The findings from the case study revealed a high degree of congruity between the respondents with respect to their judgement of teaching effectiveness and their approaches. The data highlight the importance of accurate, consistent, and evidence-based judgements as a shared value among all groups. The role of professional judgement in teacher evaluation was also emphasized across participant groups. There was a general consensus on the importance of student teacher understanding of the evaluation process and ensuring fair judgements are carried out by evaluators. While there is a strong foundation of shared beliefs about effective teacher evaluation, the nuances in group responses provide valuable insights

into the complexities of the judgement-making process and the need to consider different perspectives when developing and conducting processes that evaluate teaching.

Interestingly, there was a great degree of variability among the ratings of the seven dimensions of teaching in the video task (see Section 5.3.2), but not in the final overall rating beyond associate tutors giving a higher overall rating (i.e., teacher educators = 3.40; associate tutors = 4.25; mentor teachers = 3.38). Some interesting questions emerge when examining ways in which the groups of evaluators distributed and aggregated scores. Hence, when looking at teacher educators' rating of 'assessment' (see Table 5.4), the modal score was 4 and the mean score was 2.8, but these modal and mean scores, which present a moderately reassuring average, mask quite wide divergence in participants actual judgements, with some rating the student as competent and others as potentially causing students harm (i.e., 1 = unsatisfactory). Taking the small sample size into account, we do see a wide range of judgements among the dimensions which held true for each group of participants; it will be interesting to see how this pattern holds with a larger sample size. The ways in which judgement results are collected, aggregated, and communicated clearly matter, in particular if determining a minimum level of attainment is an aim. This potential masking effect of lower rated individual dimensions of teaching has been found in related research by Dewaele et al. (2021), who identified a masking effect of quantitative data over qualitative data. Their findings suggested that utilization of qualitative data could unearth biases in raters' judgement in some cases, thus arguing for multiple sources of evidence, a suggestion also brought forward by participants in the focus groups as a strategy to increase reliability.

Thus, TEPs should carefully consider the extent to which it is appropriate to aggregate scores at all, most particularly where there are multiple dimensions evaluated or when tripartite assessment with multiple raters occurs, which is a frequent approach considered to increase reliability (Chaplin et al., 2014; Saltis et al., 2020). Where there are significant discrepancies among evaluators, these should likely trigger an automatic review and discussion. In a study by Brown et al. (2015), only when raters had exact-adjacent or adjacent scores was the final score aggregated. When scores were at least not adjacent, additional documentation was reviewed and discussions held until agreement was reached, a response found beneficial to student teachers' development and for increasing the reliability of results. A similar approach may be favourable when considering aggregation of dimensions into a holistic rating; it is requisite in a growth model that areas of strength and weakness are clear to inform professional development. This was clearly emphasized by participants in the case study.

To inform ways of achieving greater consistency in evaluating teaching effectiveness, it is important to understand why such disparities occur. Participants' responses to *how* they determined a rating helped us understand more fully the judgement processes in evaluating effective teaching. In this study, participants relied heavily on the available perceived cues to make judgements on teaching effectiveness (see Section 5.3.2), consequently demonstrating similarity with attempts to understand teaching performance and provide a rationale for decisions founded on professional judgement (i.e., a synthesis of knowledge of the profession, experience, tacit knowledge, and practical wisdom). A common starting point for making judgements about student teachers' practices was to consider the teaching

demonstrated against the learning outcomes based on professional teaching standards. This suggests that there is a strong emphasis on student teachers' ability to meet the expectations of the standards. The second most common rationale was to look for strengths first and then weigh these against identified weaknesses, reflecting on whether the positives are more important than the negatives. This suggests there is also a focus on student teachers' strengths, and these are considered important in making judgements about overall performance. There are a variety of other rationales used by participants for making judgements about student teachers' practices. This suggests that there is no single 'right' way to make these judgements, and that different teacher educators may have different priorities.

Differences in responses appear to reflect the particularities of perspectives associated with roles in teacher preparation. For example, associate tutors found research the most difficult domain to assess, and their job responsibilities do not include doing research, though they need to be research-informed. University-based teacher educators, on the other hand, found the learning environment most challenging to judge. Teacher educators typically have not been classroom teachers for a number of years and, as is true for those in this study, have not been involved in classroom-based school experiences for some time. This lack of 'closeness' to the classroom was brought forward by some participants as an asset associate tutors and mentor teachers contribute. By the same token, teacher educators demonstrated the most variation in their ratings overall. While we can only speculate at this stage, it is worth considering the possibility that teacher educators are perhaps more ideologically freighted than teacher mentors or associate tutors and, consequently, more likely to make judgements on these ideological attachments. Given their more proximate relationship to the exigencies of day-to-day practice, the judgements of mentors and tutors are more likely to be governed by prosaic considerations such as classroom order, student industry, and low-level disruption.

While having these different perspectives would appear to suggest some incoherence, perhaps we should look at these differences more constructively. The different parties do indeed bring different lenses that heighten expertise. This includes diversity of experiences, such as experience of teaching different learners, working in and observing practice in many schools, and a litany of instructional practices and research perspectives. One of the strengths of more 'clinically' shaped teacher preparation programmes (such as were practised in the Universities of Melbourne and Glasgow; McLean Davies et al., 2015) was the requirement to meaningfully facilitate these sometimes divergent voices; to accommodate and learn from them rather than blandly homogenize them. Hence the evidence itself becomes richer and more multifaceted and synthesized rather than aggregated. Moreover, in such models, student teachers can participate in such a way that they too learn what it is to make a professional judgement. All of this is likely to conduce to balancing out subjectivism, bias, and the context-blindness of raters and encouraging collaboration. Consistency and reliability could therefore be enhanced through the amplification of expertise found in collective professional judgements. This would stand in stark opposition to the oft-used practice of 'learning with Nelly', which relies rather too much on intuitions and dispositions.

Results from analysing judgement-making strategies and warrants indicated that a small number of teacher educators and mentor teachers (4.2%) stated they needed more than what

was in the video to make a judgement, could not explain a rating, or were simply uncertain in their decision (see Section 5.3.3). Instrumentation and piloting of the video task used in this study to capture judgements and policies was carefully designed and conducted and included selecting dimensions of teaching which could reasonably be observed through perceptual information (cues) in a teaching video (see Section 2.7 and Table 2.2). However, participants did determine that not all dimensions would be demonstrated or visible in any given lesson. Additionally, in concert with SJT, the video observation task itself was designed to simulate the process used in teacher education. This brings forward the necessity to further consider better alignment between the type of evidence gathering utilized (i.e., observation) and what may actually be observable from perceptible cues. In addition to the choice of method, this finding adds focus to construct validity and the need to ensure in ITE the formats of judgement-making and the tools used speak to what they are intended to measure (i.e., standards). Prior research confirms this necessity; an intensive exploration of assorted domains and dimensions for judgement and 11 authenticated evaluation tools is provided in the systematic literature review in Chapter 3, with appropriate references.

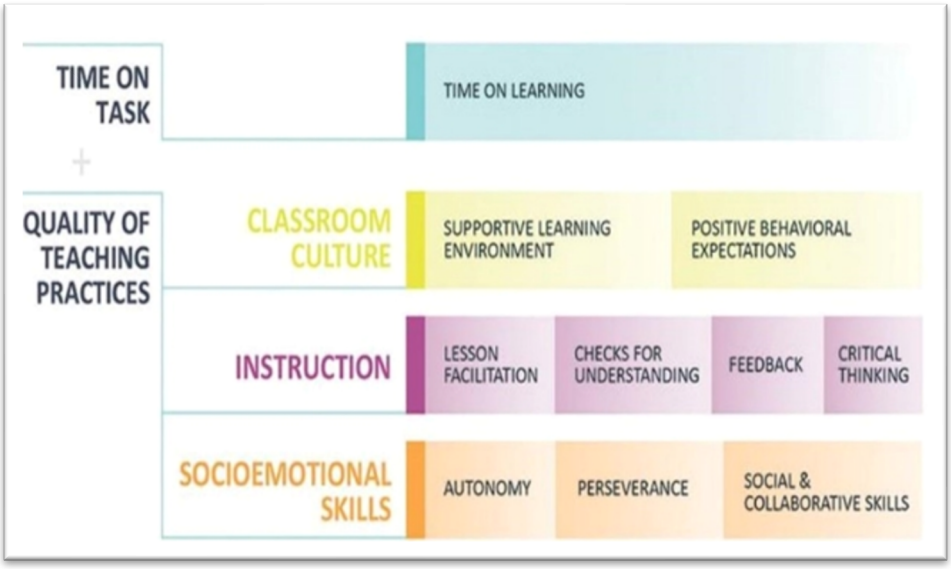
For an additional example of a substantiated approach to reliable classroom observation, it is salient to look at The World Bank's Teach Primary (and Secondary) framework (World Bank, 2022; Figure 5.3), which focuses explicitly on teaching practices. The lesson observation sheet captures three measures: time teachers spend on learning and which pupils are on task; the quality of teaching practices that help pupils develop socioemotional and cognitive skills; and aspects of the learning environment (e.g., accessibility of the classroom, materials available). The tool allows users to create additional elements determined relevant for the local curriculum and standards and has the option to exclude irrelevant elements. It is the frequent collection of these formative teaching 'snapshots' that are used to collectively understand the quality of teaching.

Assessments should continue to be both formative and summative, recognizing that effective summative evaluation requires a longer period of time and a more discerning process. It is undoubtedly the case that the World Bank frameworks offer a useful heuristic and attention to them may well be helpful. The broad categories, similar to the UNESCO *Global Framework* (Education International & UNESCO, 2019), offer a shape to observation and might be useful adapted to/overlaid on existing frameworks. However, like all such frameworks, given its performative and mildly neoliberal character, it has significant limitations. First is an inability to capture what Conroy (2004) has described as the eruptive spaces of the 'in-between' or liminal spaces. Much of what is transacted in a classroom that is important, or out of which interesting, novel, or unexpected opportunities arise, is spontaneous. These can be moments of laughter and chaos, imaginative distractions or absurdity; even, on occasion, moments of tension and disruption. As one teacher educator in this study conveyed:

> [its] the spaces rather than shapes in a lesson … experiencing the transfer of knowledge maybe in a way that I wouldn't anticipate … it's a moving concept that goes several layers deeper to find you in a space where things simply happen sometimes, and where opportunities can be found. And even if things are going very badly indeed, sometimes a space opens up, and it offers a whole perspective. (F1I2)

Figure 5.3

World Bank Group Teach Framework



Note. Figure captured from the World Bank Group website. World Bank Group. (2022, August 30). *Teach primary: Helping countries to measure effective reaching practices.* The full observation sheet is available at: https://www.worldbank.org/en/topic/education/brief/teach-secondary-helping-countries-track-and-improve-teaching-quality

One of the characteristics of the good teacher is their ability to adapt in the moment. The 'on-task' emphasis of the World Bank approach is likely to militate against both the recognition and valorization of such moments and the teacher's role in facilitating as well as responding to the momentary and fragmentary. Schools are, after all, not factories (Davis et al., 2020, and teaching is not a solo act.

Second, much of the World Bank's clientele in these matters is to be found in developing economies, where universal educational suffrage is itself still developing. These are often systems where advisory and support systems are also in relatively early stages of evolution and the support of such frameworks can be helpful. While they offer a useful point of comparison and, indeed, a complementary resource, they are not writing on tabluae rasae.

Third, the classroom may indeed be the location of the teaching and learning transactions, but it is not hermetically sealed from the sociocultural, political, and economic circumstances in which it is located. In making judgements about a particular early career teacher's performance, one should have some grasp of the life outside the school. It is difficult to imagine going into a school in West Belfast in the 1980s to assess a student teacher without having a keen sense of its political geography and demography, and of the complex, conflicted life of both students and teachers. Likewise, evaluating student teacher practice in modern-day Glasgow, identified as the least peaceful major city in the UK as it experiences extremes of social and health inequalities and injustices (Nesterova & Anderson, 2024), requires nuanced understanding. Taking a more synthetic approach to evidence gathering, as

suggested above, is more likely to match and respect the actual landscape and complexity of teaching. It can better address the how and why any particular instance of teaching was effective or not, beyond a simple rating, and it can allow for exploration of the experiences of the pupils and the teacher – factors that affect a given learning experience – and facilitate a more subtle understanding of the discursive practices of the classroom (Kerry, 1980).

5.5.2 Consistency of Processes

This chapter began with an explanation of practices and processes of judging teaching effectiveness in the University of Glasgow SoE in order to better understand and contextualize the judgement-making experience and responses of our case study participants. Overall, participants in this case emphasized the need for a transparent, standardized, and fair evaluation process. Emergent themes from analysis of focus groups transcripts revealed an emphasis on how various processes are a source of inconsistency and a potential barrier to a partnered approach to evaluating teaching effectiveness (see Tables 5.15 and Section 5.4.5), yet are also an area where strategies to gain consistency and reliability could be realized (see Table 5.16). Participants provided suggestions encompassing the full judgement-making experience, from before placements are made to reflection once completed. The processes involved in the ecosystem that is ITE school-based clinical experience, in which judgements of teaching effectiveness occur, are extensive and intertwined, so much so that space for change and improvement can be limited. It is helpful to elucidate exactly what these processes are given the prominence of processes to impact consistency and fairness.

With the university-based ITE provider being the accredited body providing ITE, many processes stem from this complex organizational matrix. Within the SoE, there are roles and leadership positions assigned to work with the local authorities and the GTCS to administer the student placement process. The student placement processes for ITE in Scotland is a national placement programme that includes a system that carries out automatic matching of student teachers to school placement offers throughout the country. The GTCS have operated and maintained the system on behalf of all 11 providers of ITE, local authorities, and schools. However, the national system reached its end of life in June 2024 and is now in transition to a new systems operator. The system has been identified as a source of frustration (Kennedy et al., 2023), in particular the way in which it allocated placements without being able to fully recognize and accommodate variability of programmes and students. As it was a national system, individual providers were limited in capacity to make changes to placement processes.

Processes also involve preparation and training of those involved in the assessment and judgement-making processes, namely the students (i.e., future teachers), classroom-based mentor teachers, and school experience tutors. This includes clear roles and responsibilities, university and school protocols (e.g., attendance, addressing disagreements), requirements for the different placements with increasing time and intensity as preparation progresses, preparation of evaluators (e.g., university assessment protocols, calibration of raters), and conveying the purpose of the evaluation. Communication processes are also a significant consideration: determining what information is shared by whom, at what time, in what

format, and through what communication system (e.g., email). The observation process in which judgements are made must also be examined; this includes establishing a shared understanding of expectations and what constitutes good teaching, planning and carrying out the observation (i.e., time, pre–post discussions, joint assessment and decision-making), formative and summative formats, practices of judgement-making, and conveying the outcome and next steps to the student. In all of this, there are technical and logistical processes, including how to access and submit reports, using university platforms such as Moodle, SharePoint, and MyCampus, and completing expense reports. There are many co-dependent elements to take into account when considering recommendations to improve processes.

Another interesting dimension of consistency emerged regarding processes. Participants in the case study evidenced consistency in their process of reasoning (Q11; see Table 5.11) and the qualitative justifications they gave for ratings of levels of performance (see Tables 5.7, 5.8, and 5.9). Participants were asked to respond to the statement 'When making judgements of teaching effectiveness, I …', choosing from four options. A majority of participants responded that they consider the teaching they have observed against the learning outcomes based on teaching standards as well as looking for strengths first and then weighing these against identified weaknesses, reflecting on if the positives are more important than the negatives. No evaluators started from a point of failure and looked for instances to challenge that decision. This approach to judgement-making was confirmed in the strategies used to determine an observation rating, which resoundingly relied on classroom cue utilization and suggestions for lesson improvement substantiated on professional judgement. This finding suggests that a standardized evaluation framework, based on learning outcomes and teaching standards, can indeed promote consistency and reduce subjectivity in judgement among evaluators. The emphasis on considering both strengths and weaknesses indicates a more balanced and holistic approach to evaluation which could foster a more constructive and supportive evaluation culture, focusing on areas for improvement rather than simply identifying deficiencies. The reliance on professional judgement and classroom cues and consistency in process suggests confidence in the judgements of teaching.

In the judgement-making process, it seems to have become a question as to what can be 'controlled' for in order to vouchsafe consistency. The input is variable, the classroom of learners is highly variable, and there is no guarantee with respect to the intended or desired outcomes of teaching given the co-dependence of teacher and learner. Consistency was deemed by participants in this study (and also by our Delphi participants; see Chapter 8) as being crucial for effective teacher preparation for a number of reasons, including fairness and reliability (though, importantly, for the senior professionals who participated in the Delphi seminar, consistency seen as a synonym for 'sameness' or even 'replicability' was universally rejected). Based on responses, consistency seems to often be conflated with replicability, in particular replicability of the process (see the World Bank's observation protocol for teacher quality; Figure 5.2). There is no conceivable way that any given observation will always produce the same results; rather, reliability describes the degree that the results can be repeated or replicated under the same conditions. In classrooms, the same

condition will never exist. It remains something of a conundrum that there is a substantial difference in the way in which consistency is considered by those making the judgements and those leading systems and thinking, with the latter group much more permissive and open to complexity of the base terms.

While many of the respondents across the various phases of this study consider consistency (to re-emphasize, not sameness) to be important, there are no obvious calibration exercises conducted with respect to individual students or between students (and supervision tutors of all stripes). The normal cross-marking exercises that striate British university life are remarkably absent in one of the most complex exercises of teacher education. Of course there may be a number of reasons for this, not least of which is likely to be resource constraints. Another may be the rhetorical call of professional autonomy and a third, the challenge of securing a sufficiently grounded and material evidence base.

Finally, participants identified that consistency in the processes of making judgements is essential for advancing the profession. Participants spoke of credibility and protection and safeguarding of the teaching profession as reasons why consistency in judging teaching effectiveness matters. Regaining professionalism can occur through gaining clarity and actionable descriptors that more clearly delineate what teachers should know and be able to do (Danielson, 2007; Darling-Hammond, 2017; Wyatt-Smith & Looney, 2016). Standardization, Danielson (2007) claimed, is the process of 'legitimization of the profession'. But from what we have seen in this study, it is vital that we evolve an altogether more sophisticated notion of consistency that is not, in fact, driven primarily by standardization. Perhaps moving from the cannon of quantitative language of consistency and reliability to that of trustworthiness and dependability is more fitting for judging the phenomenon of teaching that defies uniformity.

5.5.3 Indicators of Complexity

Results from the case study also illuminate several indicators of complexity which define the nature of shared judgement and challenge consistency and reliability. Participants resoundingly agreed that factors of complexity have an impact on judgement-making (see Table 5.13). These indicators are expressed in terms of interconnectedness, ambiguity, and cycles of cause and effect. Each are exemplified in results of this study, particularly in participants' judgement-making strategies, influences on judgements, and suggestions for ways to reduce barriers to a common understanding and partnered approach to judgement-making in ITE, particularly in school-based experiences.

Dimensions that constitute effective teaching, while presented as separate items for observation and evaluation, cannot be easily disentangled. As previously acknowledged (Council of Chief State School Officers, 2013), while assessment and evaluation processes often emphasize discrete aspect of teaching, teaching and learning are 'dynamic, integrated and reciprocal' (p. 6). We therefore observe *interconnectedness* and overlap, which must be taken into consideration as a whole to convey an accurate picture of the act of teaching. Our need for reductive measures to evaluate professional competence can sometimes leave us grappling with how to attribute the myriad of factors that impact on pupil learning (Hattie,

2023) and affect a teacher's teaching (Anderson et al., 2020. This is indicatively demonstrated in case study participants grounded judgements in one domain being based on evidence from others. For example, in explaining how a judgement was reached with respect to the domain of 'research', participants gave explanations backed by cues from multiple other domains (see Table 5.19). The dependency of rationales on others suggests that no part of quality teaching exists in isolation, and this resulted in some redundancy within responses.

Table 5.19

Examples of Complexity in Justification of Decisions

Role	Domain	Reasons	Correlated domains
Teacher educator	Research	I can see the teacher implementing *formative assessment and feedback*. However, I am uncertain if this is the most appropriate/effective approach to assessing *student learning* in the *discipline*.	Assessment Learners Content
Associate tutor		Student teacher continually extends her *questions* to *challenge pupils*, looks for *evidence cause and effect*. Her *question techniques and pupil responses* inform her of *pupil understanding* of *learning intention and task and success criteria*.	Learners Assessment Instructional strategies Planning & preparation
Mentor teacher		*Praise and encouragement* was freely given when students answered well. Lots of *reference to work done in previous lessons* and *whole class feedback* of prior learning in the form of *whole class recitation* helped to *re-embed the prior learning*.	Learners Planning & preparation Instructional strategies Learning environment Assessment

In making judgements as to student teacher competence, one is not simply drawing on the observation of the teacher's actions, discursive practices, non-verbal cues and so forth. One is also looking at the response indicators from a unique mix of pupils who are in 'continuous formation through action' (Dewey, 1916) – always changing and shifting (and presumably learning) – in order to make a judgement about the teacher, which speaks to the ecological validity of using observations for evaluation. This works under 'experimental' considerations acknowledging the complexity and dynamic nature of the learning environment as well as the realities of shifting social situations. Hence, instead of controlling or trying to eliminate it, we come to better understand the influence on judgement-making. Indeed, despite its manifest complexity, any worthy observation will include some careful consideration and calibration of pupil success. This interconnectedness gives evidence to what Hammond et al. (1977) noted as 'the zone of ambiguity'; as defined by Cooksey (1996), 'the region of entangled probabilistic relationships with which a decision maker must cope in order to reach a high

degree of achievement in the decision' (p. 142). As Opfer and Pedder (2011) noted, even the simplest teacher decisions can have multiple causal pathways.

Ambiguity is a fundamental characteristic of complex systems which arises from interconnectedness. The phrase 'context matters' is ubiquitous in education, yet often lacks clarity. It was therefore compelling that the questionnaire item that showed a lower level of agreement than all other items and the greatest degree of variation was related to this very concept – that judgements are always related to particular teachers at particular points in time and in particular situations (Q13d; see Table 5.12). The reality of a task such as effective teaching is that the constituent pieces are always in dynamic and shifting *cause and effect* relationships not only with each other but also with forces and influences that sit entirely outside the spaces of observation (e.g., the playground, the home, the local economic landscape). This is a key aspect of SJT, which emphasizes understanding the decision-making environment and conditions under which judgements are made. It is often the context of learning that creates ambiguity. We know well that circumstances surrounding learning significantly impact the learning process and outcomes. And while we advocate for teachers themselves to be able to contextualize decisions, there is a tendency in teacher preparation to not do so in the same way with student teachers as we advocate doing with learners (see Table 5.16). More clearly defining and operationalizing how context is taken into consideration when making judgements about teaching effectiveness could help bring greater consistency. However, it must be acknowledged that unwanted variability in judgements can create inequity and economic costs (e.g., related to teacher attrition).

One significant enhancement that might mitigate some of the failings of many current practices is the more considered adoption of threading, a tool of enquiry and intellectual process which considers change and continuity over time (Bermudez, 2015). Many of the infelicities seen in the variable responses noted in responses of participants in this case study might be, at least partially, mitigated by threading, which suggests maintaining longer, sustained relationships in a growth/developmental model. As Bermudez (2015) suggests:

> Threading consists of tracing the different manifestations of phenomena over time and linking them in accounts or explanations that show both the continuity of features of the past that remain in the present and the transformation of features of the present that have not always been the same. In this way, systemic thinking moves fluidly between past, present, and future. It represents processes that characterize phenomena at various points in their development and reveals different dynamics of change (e.g., progress, regression, reform, revolution, gradual change, crisis, cyclic repetition, assimilation, and marginal accommodation). (p. 112)

It is therefore a threaded approach to judgement-making which can take into account multiple perspectives and complexity with the normative requirements of teacher education. It helps us offer a more robust account of consequential validity, which in turn may protect (but not insulate) the profession of teaching and teacher education from externally imposed reductive models of evaluation driven by imperatives that derive their genesis and energy from neither educational nor student welfare imperatives.

5.6 Conclusion

In this chapter, we have synthesized the findings from a mixed methods case study that employed video analysis, questionnaires, and focus groups and interviews to examine the complexities of teaching effectiveness judgements. Our analysis, grounded in a comprehensive theoretical framework, revealed the multifaceted nature of this process, as experienced by university-based teacher educators, school experience tutors, and school-based mentor teachers. The chapter highlighted the complexity of the task and process behind what may seem to be simple ratings of teaching effectiveness. The multiple perspectives of university-based teacher educators, school experience tutors/associate tutors, and school-based mentor teachers provide insights into their judgement-making experiences. This chapter offers some important indicators and suggestions as to how we might develop teacher education and school-based experiences in ways that are more professionally robust as well as useful to the various partners in the educational space they teach. In Chapter 6, we extend this investigation through a comparative case study with the partner institution in England. This comparative analysis allows us to explore variations in judgement-making practices across different educational contexts and further refine our understanding of this critical aspect of teacher education.

6 Case Study 2: Leeds Beckett University, England

This chapter presents a case study of judgement-making in the initial teacher education (ITE) programme at the Leeds Beckett University's (LBU's) Carnegie School of Education in England. This descriptive case study includes empirical data collected through a video lesson observation task, questionnaire, and focus groups and/or interviews with university-based teacher educators, link tutors, and school-based mentor teachers. It is the second of three cases in a descriptive, multi-case approach that comprises Phase 3 of this project. The case study approach allowed for contextualization and data collection from several sources to provide a multidimensional exploration. First, the chapter presents information about provision of teacher education at the participating institution, including an explanation of school experiences and evaluation processes. Second, case-specific methodological information is detailed, and, third, there is a presentation of results. The chapter concludes with discussion of key findings.

6.1 Context of Case 2

The context of this case provides necessary background information and a description of the environment in which the research took place. It is essential within a study guided by social judgement theory (SJT; Cooksey, 1996) to consider the decision-making environment and understand the conditions under which judgements of new teachers' practices are made. This includes the educational landscape, relationships among stakeholders, programme provision, criterion measures, and types of cue information available to judges (e.g., visual and auditory cues in an observation), which can facilitate comparisons designed to highlight judgement activities. Additionally, understanding the professional teaching standards that inform judgements and the evaluation tools employed during school-based experiences, where observations of teaching occur, is valuable.

The professional standards which align with qualified teacher status (QTS; Department for Education [DfE], 2011) at the time of this research were introduced by the Conservative–Liberal Democrat coalition government formed in 2010. This makes them the longest-standing set of teaching standards since the first statutory teacher competencies were established in England in 1984. Following the general election in 2010, the outgoing Labour government's Department for Children, Schools and Families was reconfigured and renamed as the Department for Education, and Michael Gove was appointed Secretary of State for Education. His stated intention was to improve the quality of teaching, and as part of his rhetoric he claimed that the existing criteria for teachers, by which he meant the qualification standards, lacked rigour (Spendlove, 2024). The revisions to the QTS standards formed part of a catalogue of changes that impacted significantly on teacher education and were themselves part of a mosaic of changes in terms of schools policies.

The evolution of standards for the teaching profession in England began in 1984 under Conservative Prime Minister Margaret Thatcher. In 1984, the first set of statutory teacher competencies was issued in Circular 3/84, followed by amendments issued in Circular 24/89 in 1989 and subsequent updates for new secondary teachers in Circulars 9/92 and 14/93 and

for new primary teachers in 1992 and 1993, as circulars for competencies, presented as annexes 'appearing subordinate to the regulations' for teacher education (Smith, 2013, p. 430). It is interesting to note changes in terminology over this period. In the documents from the 1980s and 1992, the term 'student' is used with reference to student teachers completing their university or teaching college qualifications. This term is replaced by 'newly qualified teachers' (NQTs) in 1993. Both the change in language and the nature of the frequent updates to the competencies can be seen as 'consistent with the technical-rational approach to teacher education' (Ellis & Childs, 2023, p. 7) adopted during the 1990s, which identified specific skills and competencies required of new teachers.

The development of the competencies listed in the Circulars described above also reflects the progress towards and bringing into law of the Education Reform Act 1988, which made the National Curriculum and associated assessments mandatory for all state schools in England. Thus, the competencies written in 1992 and 1993 relate to the requirements on new teachers related to teaching and assessing pupils in line with the National Curriculum. At the same time, the regulations for teacher education were changing. Circular 24/89 directed a more school-based approach to teacher education. There was an enhanced requirement for both student teachers and their university-based lecturers to spend more time in schools, and in addition staff in schools were expected to be involved in the planning, delivery, and assessment of teacher education. Circulars 9/92 and 14/93 reinforced the statutory nature of partnerships between schools and universities, with schools receiving money for training that had previously gone to universities. From 1992, ITE was also brought into the regulatory framework through a schedule of inspections by the Office for Standards in Education (Ofsted), which brought new levels of state surveillance, scrutiny, and accountability into teacher education. The Education Act 1994 established the Teacher Training Agency, which had responsibilities for the provision and funding of teacher training in England and was charged with improving careers information about teaching and the quality of routes into the teaching profession, the aim being to support a raise in standards of teaching.

A Labour government was elected in 1997, and this change in government occurred concurrently with the transition from competencies to significantly more detailed 'standards' for NQTs. 'Although development of the first set of standards took place during the final stages of Conservative rule, they were finally published in July 1997, by which time Labour had been in power for almost two months' (Smith, 2013, p. 436).

6.1.1 Teaching and Teacher Education in England: An Overview

Prior to devolution of Scotland and Wales in 1997, policy and practice were implemented centrally by the UK government and this influenced educational practices in Scotland and Wales. Between 1997 and 2010, a swift and sweeping set of education policy initiatives were introduced by the Department for Education and Employment, which became the Department for Education and Skills and later the Department for Children, Schools and Families. The Teaching and Higher Education Act 1998 led to the establishment of the General Teaching Council for England (GTCE) in 2000 to support improvement of the quality of teaching and learning and become the regulator of teacher conduct, therefore holding responsibility for

professional standards. In the Education Act 2005, the Teacher Training Agency was relaunched as the Training and Development Agency for Schools (TDA), which was directly accountable to Parliament. In line with Labour schools policy, such as Every Child Matters, the TDA had an expanded remit with responsibility for improving the training and development of the entire school workforce. Many of these changes impacted directly on teacher education and the expectations placed on teachers by the state. New legislation, standards, and organizational infrastructure embedded the term Initial Teacher Training (ITT) rather than ITE, and there was rapid growth in what was framed as the school-led ITT sector. Two new sets of standards were introduced in this era, in 2002 and 2007. In 2002, standards were categorized into three groups:

- professional values and practice
- knowledge and understanding
- teaching

A major change in 2007 was a newly differentiated model of teachers' standards based on professional development and career stages. This meant that for the first time, standards for trainee teachers (as they were then typically known) became the foundation for a hierarchy of new descriptors for expected standards for NQTs: main scale, upper pay scale and advanced skills teachers. Despite recognizing the different career phases, this new document was more condensed than the 2002 version and was presented as a large, coloured poster showing career progression and related professional expectations. These new descriptors included references to reflective and reflexive practice, which Knight (2017) suggested were welcomed by ITE providers and teachers.

Following the 2010 election and under the Conservative–Liberal Democrat coalition, the newly designated DfE, with Michael Gove as Secretary of State for Education, undertook what it called the 'bonfire of the quangos', which led to a series of changes. In 2012 the Teaching Agency was established as an executive agency of the DfE, in place of the TDA, with some of the former GTCE roles (the GTCE was abolished). The Teaching Agency was thus responsible for ITT in England as well as the regulation of the teaching profession. It was then merged with the National College for School Leadership to become the National College for Teaching and Leadership in 2013. A consequence of these changes included 'the loss of significant teacher education policy expertise and sector intelligence' (Spendlove, 2024, p. 48).

Amid these changes, the 2011 Teachers' Standards (DfE, 2011) were established, and these remain current at the time of this research. There are eight generic standards covering teachers qualifying to teach in primary and secondary sectors and incorporating all existing teachers. While offering a simplified document and reduced set of standards (from the previous 102 separate standards), the generic nature of these is contentious. The same standards now apply to assess trainee teachers during and on completing ITT, at the end of their 1 year as NQTs, now 2 years with early career teachers (ECTs) status and throughout their time in the profession.

Although the 2011 Teachers' Standards have not been altered, there have continued to be significant changes in the sector. Despite the persistence of the QTS standards, it is noteworthy that DfE-designated academies and free schools can and do employ teachers without QTS (DfE, 2011). The majority of secondary schools (about 80%) are now academies, either stand-alone or within multi-academy trusts, as are almost 50% of primary schools, so this exclusion is not insignificant. A new Early Career Framework (DfE, 2019) became statutory in 2021 following pilot and early roll-out phases. This meant that all new teachers were classed as ECTs for 2 years (replacing the 1-year NQT status). The Early Career Framework sets out the training content which all new teachers are expected to master, and it is framed as a series of evidence statements worded as 'learn that' and practice 'learn how to' statements covering five core areas: behaviour management; pedagogy; curriculum; assessment; and professional behaviours. The framework is aligned with the Teachers' Standards, which remain the benchmark for assessment of trainee teachers and ECTs. Teachers in England can gain QTS through a wide range of ITT routes, including those offered by universities, school-based consortia, and new providers. This diverse ITT provision landscape was further consolidated following the DfE ITT accreditation process in 2022.

6.1.2 Initial Teacher Education at LBU

Carnegie School of Education is a long-established provider of ITE with an extensive school partnership. In 2023–2024 so far, we have organized block placements with 248 schools across 14 local authorities. We work closely with our school partners through our Strategic Partnership Committee, which meets once each term and is composed of members of the university team and school colleagues. We also hold regular meetings with our School Direct partners; these meetings are attended by lead mentors from schools and help develop joint practice within and across our School Direct partnerships. In addition, we hold periodic meetings with school partners and fellow higher education institutions (HEIs) to review our partnership and ensure that our practice is consistent with that of other school and HEI-led providers. In the current year, we are running the following Postgraduate Certificate in Education (PGCE) and undergraduate (BA Hons) routes into teaching:

PGCE Primary Education (3–7)	University led
PGCE Primary Education (3–7)	School Direct
PGCE Primary Education (5–11)	University led
PGCE Primary Education (5–11)	School Direct
PGCE Primary Education (5–11) – Physical Education	University led
PGCE Primary Education (5–11) – Physical Education	School Direct
PGCE Secondary Education (11–16) – Physical Education	University led
PGCE Secondary Education (11–16) – Physical Education	School Direct
PGCE Secondary Education (11–16) – English	University led
PGCE Secondary Education (11–16) – English	School Direct
PGCE Secondary Education (11–16) – Mathematics	University led

PGCE Secondary Education (11–16) – Mathematics	School Direct
PGCE Secondary Education (11–16) – Physical Education with EBACC English	University led
PGCE Secondary Education (11–16) – Physical Education with EBACC Mathematics	University led
BA (Hons) Primary Education (3–7) Final year	University led
BA (Hons) Primary Education (5–11) Final year	University led

Annually we train between 500 and 600 student teachers across all undergraduate and postgraduate courses with the aim of ensuring that we produce teachers who go on to be an essential part of the education workforce regionally, nationally, and globally.

Our shared partnership mission and aims heavily influence the intent, design, and development of all our programmes. We are constantly seeking to enhance our provision, which takes shared responsibility for improving the achievement of pupils in partnership schools (and beyond) by sharing resources and expertise within the partnership. We fully and purposefully implement the intended ITE curriculum, which is designed and aligned to the core content framework and beyond to achieve the right balance between subject content and pedagogical skills and ensures that our student teachers develop the highest level of pedagogical content knowledge. Our partners have a shared vision of our programme and are integral to all aspects of the education we provide. Our mission is to

> contribute to the transformation of the lives and outcomes of children, young people, lifelong learners and families regionally, nationally, and internationally, by utilizing our capacity for knowledge creation, enhancement and dissemination to develop exceptional, socially aware and responsible members of the children's and educational workforce.

The undergraduate programme is a 3-year course of study and the postgraduate programme is a 1-year course. Both have the following core elements:

- assessed placement and practice;
- taught sessions within the university which focus on subject knowledge, pedological development, reflective practice, and professional values;
- additional enhancement opportunities, such as forest schools for primary students;
- considering transitions and practices in phases that are not in the assessed practice – e.g., secondary PGCE students having some time in a primary classroom; and
- special educational needs and social justice as the core value of the programmes.

6.1.3 LBU Practices and Processes for Judging Teaching Effectiveness

The school-based training element of the course is integrated within the whole curriculum to ensure that there is a clear purpose to what is being taught in university and then practised on placement and, in turn, reviewed and reflected on to ensure that students are developing and

deepening key skills and knowledge. Also, the training has been designed to ensure that it is fully compliant with the ITE framework both logistically and in terms of content, and it incorporates the core content framework and beyond.

Each student is supported by a school-based mentor and a university link tutor (the titles of link tutor and associate tutor are interchangeable) while on school experience. During each school experience, students will engage regularly with their mentor and associate tutors about the progress they are making towards becoming an outstanding teacher.

On placements, students learn through a range of experiences underpinned by:

- a carefully structured programme that scaffolds their progress each term;
- a gradual build-up of teaching commitments over time;
- continued observation of experienced teachers across all key stages;
- reflective activities that cover the roles and responsibilities of a teacher and explore the specialized knowledge of a teacher;
- weekly mentor meetings;
- professional development events;
- opportunities to be involved in all aspects of school life; and
- targets set for further development.

Also, while on placement, schools support students via the school-based mentor, who supports students on a day-to-day basis, focusing on classroom practice, teaching and learning, and what it means to be a teacher and their wider responsibilities. Furthermore, secondary students have a subject mentor and a professional mentor. Our priority is to ensure that all elements of our courses are fully integrated and feed into and out of each other. To ensure that students have the opportunity to deepen their understanding, students come back to university for a day while on placement. This allows students to enhance their reflective practice skills and to share ideas in the development of new skills. This process of reflection is a key aspect of developing and assessing for progress.

On the undergraduate course, the placements are sequenced carefully throughout the year to work synergistically with the university-taught sessions, to ensure that students have opportunities to not only meet the requirements as set out in the DfE ITE framework, but also work in school and explore ideas that they are learning in the university sessions. The formal placements are also allocated at different times of the year to ensure that students, by the end of the course, will have experienced the whole year in school. The opportunities are as follows:

- Year 1 – Holistic placement, autumn term
- Year 1 – Phase 1 placement, summer term
- Year 1 – Lower Key Stage placement, summer term
- Year 2 – Phase 2 placement, spring term
- Year 2 – Higher Key Stage placement, summer term
- Year 3 – Phase 3 placement, autumn term

Within our PGCE courses, we have two phases for our placements, which are long blocks.

Within these blocks, students return to university at points in the curriculum to focus on and develop skills to evaluate their knowledge and examine their understanding.

Within both the undergraduate and PGCE courses, the key design feature of having clusters where ideas developed thus far are considered and evaluated by the students is vital to ensure that students' progress in their development is considered and mastered. Progress on placement is measured formatively against the Expected Progress Statements (EPS) and then evaluated at the midpoint and final point of the placement. The process of establishing how a student is progressing has the following stages:

- Weekly reviews are carried out by the school-based mentors, calibrated against the EPS.
- Targets are set against the EPS each week to support small-step progress.
- There is a review of progress against the targets each week to ensure that students are developing.
- Associate tutors have an overview of the development via PebblePad and communication with the school-based mentor.
- There is a midpoint evaluation with student, school-based mentor, and associate tutors. This is where the associate tutors visit the school, observe the student, and review progress and evidence on PebblePad. This meeting is also to quality assure the mentor's judgement and process for supporting the students.
- The previous stages are repeated weekly throughout the placement, and any challenges that the student may have are evaluated and they are supported to develop skills and knowledge.
- Final review against the expected progress statements is carried out with the mentor, student, and associate tutors. Targets are set for students to develop in readiness for the next placement and while at university.
- If the student is in their final phase, a summative assessment is carried out against the Teachers' Standards to ensure that they can meet the requirements for QTS. Targets are set for their ECT year to support the transition.

6.2 Case-Specific Methods

Methods applicable to the entire research project are presented in Chapter 2; this includes the theoretical framework of SJT, strategies to ensure trustworthiness of results, and the ethical approach taken. Methods which relate to all three case studies in the multi-case design are also presented in Chapter 2, Section 2.7 (the case study protocol is provided in Appendix A2.3). Therefore, this section only includes considerations specific to recruitment and data collection for this case.

A total of 24 participants completed the video task and questionnaire: 13 university teacher educators; 7 tutors; and 4 school-based mentor teachers (see Table 6.1). Participants were selected through purposeful sampling (Cohen et al., 2018). The goal was to select participants who reflected the various roles of individuals who conduct observations and evaluate teaching effectiveness during educator preparation and could best contribute to answering the research questions. These participants demonstrated a perspective within a defined context and had enough information for in-depth exploration (Merriam, 1998). There were also a few

participants who agreed to contribute to 45-minute focus groups or interviews; this included two teacher educators, two tutors, and three mentor teachers.

Table 6.1

Case Study 2 Participants

	Teacher educators	Tutors*	Mentors	Overall
	(n = 24)			
Potential participants	32	34	Approx. 250	N/A
Video task and questionnaire	13	7	4	24
Focus group/interview	2	2	3	7

Note. * Tutors include participants who indicated they were link tutors or associate tutors.

The potential participants included 32 teacher educators, all permanent members of staff at the time of the research. They were sent a summary of the research project via email and given the option to contribute to the research. Time was allocated in their workload model to support the research. This was not a mandatory requirement, and they were given the option to determine the extent to which they would like to contribute. All 32 members of the team were supervising students in school across the primary or secondary course. They were also at the time personal tutors for many of the students, but not for those they were supervising in school. Their role within the department also consisted of the following:

- marking assessments
- teaching on ITE modules
- supporting personal development of their students
- contributing to curriculum development
- preparing for inspection

Recruitment occurred during the late autumn term of 2023. A total of 13 teacher educators completed the video task and questionnaire. Of these, two agreed to contribute to a 45-minute focus group.

Recruitment of link tutors to take part in the research occurred in the late autumn term of 2023. The 34 tutors were sent a summary of the research project via email and given the option to contribute to the research. Time was allocated in their workload model to support the research. This was not a mandatory requirement, and they were given the option to determine the extent to which they would like to contribute. At the time of this study, all link tutors (or associate tutors) held part-time, non-permanent teaching roles at the equivalent academic level of lecturer. The role is flexible and can involve delivering instruction, marking, providing instructional support for students, and supervising students while in

school-based experiences (i.e., on placement). Predominately all link tutors are allocated students to supervise in school and very few teach on the course. The link tutors report to the head of ITE and are supported through a range of training, which takes place at the same time as training for the permanent teacher educators. In total, seven associate tutors completed the video task and questionnaire; two agreed to an individual interview.

Recruitment of school-based mentor teachers was facilitated by the School's Placements and Partnerships lead, whose role involves coordination of a team that organizes and administrates all the school experiences. The mentors carry out the following key functions:

- contact individual schools
- match students to schools
- support students in organizing travel to schools
- work with mentors to ensure that they are fully aware of the requirements
- allocate link tutors to students
- communicate and remind all link tutors of the process for supervision of students
- collate assessment of placements

Mentors were sent details of the project via email and invited to offer their participation. There was an initial high interest in the project; however, as more details were sent on request of individual mentors, the time requirements seemed to reduce the numbers who finally agreed to take part. Four mentor teachers completed the video task questionnaire, and individual interviews were conducted with three of them.

A number of individuals completed the informed consent and demographic questions, but when the first judgement item in the task was presented and a rationale queried, these individuals did not continue. The completion rate for each group of participants is included in Table 6.2.

Table 6.2

Case Study 2 Completion Rates

	Teacher educators	Tutors	Mentors
Began video task and questionnaire	21	10	4
Completed video task and questionnaire	13	7	4
Completion rate	62%	70%	100%

While definitive reasons for the survey dropout rate remain unknown, plausible explanations can be attributed to both survey design and participant-related factors. Some of these factors could be:

- the time commitment required for the survey;
- the time of the year when the survey was carried out (as this might have conflicted with what was happening in school and also in university; and

- the inspection taking place within ITE at LBU and the preparation needed meant that teacher educators had limited time to give to other activities, such as this research.

6.3 Video Task and Questionnaire Results

6.3.1 Participant Demographics

For the 24 participants in the video task and questionnaire, all of whom were current or former teachers, a detailed overview of participant roles, qualifications, and experience is presented in Table 6.3.

Table 6.3

Participant Demographics for the Video Task and Questionnaire

		Teacher educators (n = 13)	Tutors (n = 7)	Mentor teachers (n = 4)	Overall (n = 24)
Gender	Female	7	6	4	17
	Male	4	0	0	4
	Non-binary/third gender	0	0	0	0
	Prefer not to say	2	1	0	3
Overall experience in education	Under 25 years	7	1	3	11
	25 to 29 years	1	2	1	4
	30 to 39 years	5	3	0	8
	40 to 49 years	0	1	0	1
Year of experience in current role	Under 25 years	13	7	4	24
	25 to 29 years	0	0	0	0
	30 to 39 years	0	0	0	0
	40 to 49 years	0	0	0	0
Route into teaching	Undergraduate	3	3	3	9
	Postgraduate	10	3	1	14
	No qualifications	0	0	0	0
	Other	0	1	0	1
Teaching qualification	Nursery	0	0	0	0
	Primary	9	7	3	19
	Secondary	4	1	1	6
	Specialist	1	1	0	2
	None	0	0	0	0
	Other	2	1	0	3

Country where teaching qualification was obtained	Scotland	0	0	0	0
	England	13	7	4	24
	Wales	0	0	0	0
	Northern Ireland	0	0	0	0
	Other	0	0	0	0
Highest level of qualification	Below bachelor's degree	0	0	0	0
	Bachelor's degree	0	3	3	6
	Postgraduate	6	3	1	10
	Master's degree	7	1	0	8
	Doctorate	0	0	0	0

Most participants were female and had substantial years of experience in the field of education. Many of the participants had qualified as teachers through the postgraduate route (58.3%; *n* = 14), with some (37.5%; *n* = 9) undertaking the undergraduate programme for a teaching qualification. Nineteen (79.2%) of the participants had experience teaching primary education. All the participants obtained their teaching qualification in England. Eight participants (33.3%) had attained a master's degree and 18 (75.0%) had qualifications beyond the bachelor's level.

6.3.2 Results from the Video Observation and Judgement Task

Participants' range of responses and patterns of consensus and dissensus on observed teaching effectiveness are presented in Tables 6.4, 6.5, and 6.6. Participants were asked to watch a 15-minute video, which simulated the natural process of lesson observation used in teacher education, and then provide judgements in each of the seven dimensions of the United Nations Educational, Scientific and Cultural Organization (UNESCO) *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019; see Chapter 2) and an overall judgement of the teaching effectiveness, and indicate which dimensions were most and least difficult to judge (see expanded results in Appendix A6.1). They were also asked in open-ended prompts to explain *how* they made these judgement decisions in order to capture the cues utilized, their judgement policies, and potential influences. Results are presented according to role in the judgement-making process as well as overall.

Table 6.4

Teacher Educators' Judgements on Seven Elements of Observable Practice of UNESCO Professional Teaching Standards

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(*n* = 13)							

Q1. Learners	2	7	3	1	0	4	3.77	0.80
Q2. Content	4	8	1	0	0	4	4.23	0.58
Q3. Research	1	7	5	0	0	4	3.69	0.61
Q4. Planning & preparation	5	6	2	0	0	4	4.23	0.70
Q5. Instructional strategies	1	7	5	0	0	4	3.69	0.61
Q6. Learning environment	7	4	2	0	0	5	4.38	0.74
Q7. Assessment	3	4	6	0	0	3	3.77	0.80
Q8. Overall rating	2	8	3	0	0	4	3.92	0.62

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.

The overall judgement by teacher educators of teaching effectiveness was 3.92 out of a possible 5.0, indicating above satisfactory teaching was demonstrated. This was the highest rating among all three groups. There was overall agreement that the teaching demonstrated was above a satisfactory level. The judgements made by teacher educators varied, with some dimensions being considered highly effective to satisfactory. No rating of unsatisfactory (1) was given, and only one occurrence of nearly satisfactory (2). These judgements indicated a relatively high degree of agreement around the mode of 4. The dimension rated highest was 'learning environment' and the lowest two, which were still above satisfactory, were 'instructional strategies' and 'research'. The dimensions of highest standard deviation (SD) were in 'learners' and 'assessment', followed by 'learning environment'. Mean scores for the seven individual areas ranged from 3.69 to 4.38 (R = 0.69).

Table 6.5

Tutors' Judgements on Seven Elements of Observable Practices of UNESCO Professional Teaching Standards*

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(n = 7)							
Q1. Learners	0	3	3	1	0	3, 4	3.29	0.70
Q2. Content	1	2	4	0	0	3	3.57	0.73
Q3. Research	0	2	5	0	0	3	3.29	0.45
Q4. Planning & preparation	1	2	4	0	0	3	3.57	0.73
Q5. Instructional strategies	0	2	1	4	0	2	2.71	0.88
Q6. Learning environment	2	2	3	0	0	3	3.86	0.83

Q7. Assessment	0	2	2	3	0	2	2.86	0.83
Q8. Overall rating	1	1	5	0	0	3	3.43	0.73

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.
* Includes associate tutors and link tutors.

The overall judgement of teaching effectiveness by tutors was 3.43 out of a possible 5.0, indicating a slightly above satisfactory level of effective teaching was demonstrated. Judgements made by link tutors also varied from highly effective to nearly unsatisfactory (2). A tendency to rate towards the middle was reflected. The lowest-scoring option of unsatisfactory (1) was not given by any tutor. Three areas were rated highest: 'learning environment', 'planning & preparation' and 'content'. The lowest-rated area was 'instructional strategies', which was also the area of highest deviation. Across all areas, there was a similar degree of deviation among the ratings of tutors as the teacher educators. Mean scores for the seven individual areas ranged from 2.71 to 3.86 (R = 1.15).

Table 6.6

Mentor Teachers' Judgements on Seven Elements of Observable Practices of UNESCO Professional Teaching Standards

	Level of performance							
	5	4	3	2	1	Mode	Mean	SD
	(n = 4)							
Q1. Learners	0	2	1	1	0	4	3.25	0.83
Q2. Content	0	1	3	0	0	3	3.25	0.43
Q3. Research	0	0	2	2	0	2, 3	2.50	0.50
Q4. Planning & preparation	0	1	3	0	0	3	3.25	0.43
Q5. Instructional strategies	0	1	1	0	2	1	2.25	1.30
Q6. Learning environment	1	2	1	0	0	4	4.00	0.71
Q7. Assessment	0	0	2	2	0	2, 3	2.50	0.50
Q8. Overall rating	0	1	2	1	0	3	3.00	0.71

Note. Questionnaire: Q1–8; 5 = highly effective and 1 = unsatisfactory.

Results of the task for mentor teachers are presented in Table 6.6. The overall judgement of teaching effectiveness by mentors was 3.0 out of a possible 5.0, indicating satisfactory teaching was demonstrated; this rating was the lowest among the three groups. Judgements made by mentor teachers varied from highly effective to unsatisfactory. The area rated highest was 'learning environment' and the lowest was 'instructional strategies', followed closely by 'assessment'. Across all areas, there was more deviation among the ratings of mentor teachers than the other groups, with the most variation occurring in the area of

'instructional strategies'. The highest range of scores across the seven dimensions was demonstrated by mentor teachers, with mean scores ranging from 2.25 to 4.00 (R = 1.75).

The 'learning environment' was the highest-rated dimension by all three groups. The teacher educators did not have any dimensions that would be considered a low rating. No single area was consistently seen as a weakness in teaching observed. However, the domains of 'instructional strategies' and 'assessment' were both rated as unsatisfactory (below 3) by both the tutors and mentor teachers. These were the only items to be rated below satisfactory. The areas of highest deviation were also variable across groups. There was more variation in the ratings of the small group of mentor teachers than the ratings of tutors or teacher educators.

6.3.3 Results: Strategies and Rationales for Ratings

Along with the ordinal judgement provided for the observed video lesson, participants were asked an open-ended question for each of the seven dimensions: 'How did you decide what level of performance was demonstrated?' This occurred in order to capture cues utilized, judgement policies, and potential influences. This was asked for all seven dimensions which were rated, and qualitative responses were analysed using the constant comparative method of data analysis for each of the three groups of participants. Findings are presented in Tables 6.7, 6.8, and 6.9 according to roles, with indicative statements of participants provided. In social judgement theory (Cooksey, 1996), the ways (i.e., strategies) in which judges use available cues to make decisions is termed 'cue utilization validities'; these are judges' attempts to understand the teaching observed. If a strategy was used even once, it was recorded. Prevalence and distribution of strategies and rationales (i.e., warrants), for judgements are presented. We have included quotes from participants to illustrate and provide credibility to findings; participant codes from the analysis processes are included.

6.3.3.1 Teacher Educators

Results of our analysis suggest that the 13 university-based teacher educators used four strategies to determine an observation rating: (a) classroom cue utilization; (b) suggestions for lesson improvement; (c) internal expectation criteria; and (d) no identified strategy. As teacher educators reasoned with a given strategy, they employed a specific rationale or backing for the strategy used. Three types of justifications were evident: professional judgement; personal judgement; and indeterminate judgement. We now describe the strategies and warrants in detail, with typical examples provided. The most recurrent justification was professional judgement, with the most utilized strategy being classroom cue utilization. Many of the judgement cues used to assess the student teacher's performance were observed actions of the teacher, observed pupil actions, and context cues from the learning environment.

Table 6.7

Teacher Educators' Judgement Strategies and Rationales

Professional judgement		Personal judgement	Indeterminate judgement
Classroom cue utilization (n = 184)	Suggestions for lesson improvement (n = 54)	Using internal expectation criteria (n = 7)	No identified strategy (n = 14)
Observed teacher action (70) Observed pupil action (34) Context cues (26) Explanatory rationale (14) Multiple general examples of evidence to support rationale (14) Physical environment (11) Learning materials (9) Teacher and pupil interaction (5) Pupil learning (1)	Lesson improvement (38) Observed omissions (16)	Internal criteria (7)	Restatement of dimension (7) Need more to make judgement (6) Generalization (1)

Note. Total codes from qualitative questionnaire statements: Q1–8.

Classroom cue utilization (rationale: professional judgement). Participants utilized perceived aspects of a student teachers' observable practices and cues considered relevant from the classroom in decision-making. This strategy accounted for approximately 71% of cues coded, indicating what judges looked to most when making a decision. Their attention was directed to multiple cues, some of which were interdependent. The most common cues were from the teacher's actions, the pupils' actions, and context cues from the classroom learning environment (e.g., learning materials, board, classroom layout). Both positive and negative occurrences of these cues were noted. A few examples of observed teacher actions included:

- They [pupils] are supported by teacher questioning and encouragement to extend their thinking (Q1b. Learners)
- The teacher articulated her points very clearly and was able to answer many questions to unpick the content and guide the students (Q2b. Content)
- Teacher made clear links to prior learning journey and connections to previous content (Q3b. Research)

- She was using questioning and scaffolded group work effectively (Q5b. Instructional strategies)
- Teacher was constantly formatively assessing both individuals and groups through observation and discussion (Q7b. Assessment)

Pupils' actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. A few of these were:

- Learners who are unsure are able to ask for help (Q1b. Learners)
- There was a clear focus on the skills that were been taught and the pupils were required to use the skills throughout the lesson (Q2b. Content)
- Some children found it hard to read the teacher's writing and to find the relevant part on the board (Q2b. Content)
- Most groups had to ask questions about what to do after the initial instructions were given out (Q4b. Planning & preparation)
- The group work also enable[d] lots of social interaction and active learning (Q6b. Learning environment)

Another main strategy involved a statement of what the teacher did, but this was specifically followed by multiple examples as evidence to support the main statement or an explanatory rationale of what a particular action caused or resulted in. For example:

- Pupils were working in groups, so those students who may have been struggling with the concepts were scaffolded within the group. The visual organizers were used as a scaffolding tool. The sequence of lessons had been planned in small steps to reduce the likelihood of cognitive overload. (Q1a. Learners)
- Teacher reviews prior learning. She moves from the known to the unknown. She scaffolds their learning. She explains terminology. (Q2b. Content)
- The teacher seems to employ elements listed by Rosenshine – daily review, checking student understanding, scaffolding tasks. The clip suggests that formative assessment is built into this task. (Q3b. Research)
- Behaviour management was good and at some points it was clear she had developed positive relationships with the pupils. They were polite and not afraid to ask questions if they need some clarification of the task. I am unsure if the pupils needed to read the questions on the board. (Q6b. Learning environment)

Judges also observed the physical environment of the classroom. This included how the desks were arranged, what was on the blackboard, and the materials used for learning, such as graphic organizers. Rationales supporting identification of these cues included:

- Teacher supported group through use of textbook and graphic organizers (Q1b. Learners)
- The blackboard was cluttered and some children found it hard to read the teacher's writing and to find the relevant part on the board. This made the content slightly inaccessible. (Q2b. Content)
- The teacher had pre-written instructions on the board and had materials to support the lesson (Q4b. Planning & preparation)
- Questions on the board for students to refer back to (Q5b. Instructional strategies)

- The students were seated in groups facing each other (Q6b. Learning environment)
- The teacher had appropriate texts available (Q6b. Learning environment)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy builds from the evaluators' professional judgement and reflects their role as an individual responsible for preparing new teachers entering the profession as well as their own experiences with teachers, student teachers, and pupils in multiple classrooms and schools. Suggestions for improvement constituted 20.8% of the rationales to support judgements. From this perspective, years of experience, and prior knowledge, evaluators used professional judgement by indicating what was not observed (i.e., omissions) and suggested how the lesson might be changed to improve the quality and rating assigned. Examples from the data to support this reasoning strategy included:

- Perhaps could have supported student to break question down (Q1b. Learners)
- Due to inappropriate questioning on the board (too long, wrong structure and too many command words in one sentence), some of the content was not accessible to learners (Q2b. Content)
- Did not see evidence of pre-assessment but there was questioning throughout. Not sure enough open questioning or use of formative assessment. (Q3b. Research)
- There may have been too much to do and pace a little too fast (Q4b. Planning & preparation).
- Tended to be teacher driven rather than student inquiry led (Q5b. Instructional strategies)
- Did not offer pupils opportunities to research own texts and follow own interests/develop own opinions (Q6b. Learning environment)

Using internal expectation criteria (rationale: personal judgement). This rationale for respondents' judgements appeared to involve underlying personal constructs such as the evaluator's beliefs, value systems, expectations, or even emotions. While relatively uncommon among the strategies used (i.e., 2.7%), perceptions that come from within the judges themselves were evident. It is important to note that strategies coded as internal criteria may have developed through professional experience; the scope of the data collected did not provide any indication as to whether or not internal criteria were based on professional knowledge or personal preferences. Statements given from participants included:

- I felt the use of questions was an effective assessment (Q3b. Research)
- To be a '5', I would want the board sorted (Q4b. Planning & preparation)
- I was slightly uncomfortable about making a less confident reader read aloud in front of the whole class (Q5b. Instructional strategies)
- Style for me is too authoritarian but I think it is cultural (Q5b. Instructional strategies).

No identified strategy (rationale: indeterminate judgement). Some participants could not give the basis for their judgments or the basis was not evident (5.4%). A few teacher educators restated the description of the dimension instead of providing a rationale. Some

participants stated they were not in a position to provide a judgement or needed additional evidence. Indicative responses included:

- This was an area of strength (Q2b. Content)
- Teacher seemed to have a good grasp of the subject matter (Q2b. Content)
- It would have been useful to see the lesson end to gain more evidence of 'research' (Q3b. Research)
- I'm not completely sure whether this includes extension tasks but what I see suggests a very competent teacher (Q4b. Planning & preparation)
- Teaching activities appeared to be sound for the subject matter (Q5b. Instructional strategies)
- Yes to all the above [referencing the description] (Q6b. Learning environment)
- Note – I wanted to leave the grade blank. As it is needed a grade to move on I have given a '3' but simply for moving on! (Q7b. Assessment)
- Hard to say from this clip alone but elements of assessment for learning are there (Q7b. Assessment)

6.3.3.2 Tutors

Table 6.8 indicates the range of evidence participants drew on to judge teaching effectiveness. Participants in the role of tutors at LBU used four strategies to determine an observation rating: (a) classroom cue utilization; (b) suggestions for lesson improvement; (c) internal expectation criteria; and (d) no identified strategy. As tutors reasoned with a given strategy, they employed three rationales for backing strategies: professional judgement; personal judgement; and indeterminant judgement. We now describe raters' strategies and rationales in detail, with typical examples included.

Table 6.8

Tutors' Judgement Strategies and Rationales

Professional judgement		Personal judgement	Indeterminate judgement
Classroom cue utilization (*n* = 115)	Suggestions for lesson improvement (*n* = 71)	Using internal expectation criteria (*n* = 7)	No identified strategy (*n* = 14)
Observed teacher action (38) Observed pupil action (24) Explanatory rationale (17) Context cues (10) Multiple general examples of evidence to support rationale (8) Teacher and pupil interaction (7) Learning materials (5) Physical environment (4) Pupil learning (2)	Lesson improvement (37) Observed omission (26) Question posed (8)	Internal criteria (7)	Restatement of dimension (7) Need more to make judgement (5) No response (2)

Note. Total codes from qualitative questionnaire statements: Q1–8.

Classroom cue utilization (rationale: professional judgement). In their decision-making, participants utilized perceived aspects of a student teacher's observable practices and relevant classroom cues. This strategy accounted for approximately 55.5% of cues. The most common cues were from the teacher's actions, the pupils' actions, interactions between the teacher and pupils, and contextual examples and cues. A few examples of classroom cues from observed teacher actions were:

- Her lesson was intellectually challenging, asking pupils to come up with a common theme across several texts and reminding them to find evidence to support this (Q1b. Learners)
- She made reference to previous lessons to recap the texts read (Q2b. Content)
- She did move around the room to support the independent learning. (Q3b. Research)
- The teacher was clear about what she wanted from the lesson. (Q4b. Planning & preparation)
- She supported groups and a clear timeline [was] given (Q5b. Instructional strategies)
- She asked one group what their theme was and asked them if they were able to find evidence (Q7b. Assessment)

Pupils' actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. These included:

- Appeared some had difficulty in reading from the board – lots of words! (Q1b. Learners)
- During the discussion the learners were passive (Q2b. Content)
- Pupils were struggling with how much to write for their summary (Q3b. Research)
- Allowed pupils to work in groups to find a theme and summarize texts while linking them to the common theme (Q5b. Instructional strategies)
- The relationships had been developed with the students to provide a safe and secure environment (Q6c. Learning environment)

Another main strategy involved a statement of what the teacher did, but specifically followed with multiple examples as evidence to support the main statement, or with an explanatory rationale which clarified what was achieved through the action. For example:

- Her lesson was intellectually challenging, asking pupils to come up with a common theme across several texts and reminding them to find evidence to support this. Her movement around the room was good – this allowed her to support pupils. (Q1b. Learners)
- She seemed to have a good understanding of her topic – giving examples and supporting groups (Q2b. Content)
- Questioning was focused and probing with challenges for evidence, recapping and use of skills such as summarizing – which draws knowledge and understanding together. Finding a theme and evidencing show analytical skills. (Q3b. Research)
- Her creation of the graphic organizer to support pupils with their answers was a good example of scaffolding and preparation for the session (Q4. Planning & preparation)
- Clear focus on pace and expectations, and checking that all knew what was expected – encourages security in the task and support (Q6b. Learning environment).
- Use of questioning to assess at the start. Probing questions used in group discussions to support further learning and deepen the understanding. The tasks set would have provided evidence of this further and allowed the teacher to assess more individually against the criteria – vocab, evidence, key details. (Q7b. Assessment)

Another main strategy involved a statement related to the context cues and materials from the classroom and the lesson being taught. For example:

- The pace of the lesson was good for most (Q1b. Learners)
- Clear learning objective (Q2b. Content)
- The final summary activity would have given the teacher a good understanding of the students understanding and progress (Q3b. Research)
- There were three texts being read, a knowledge organizer, and a huge paper for writing on, plus all notepads, etc. on quite small tables (Q5b. Instructional strategies)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy, which accounted for 34.3% of the strategies coded, was a focused on lesson improvement, building from the tutors' professional judgement. This reflected their role in teacher education, which specifically involves supporting students on placement in schools,

conducting observations, and completing assessments. When using the lesson improvement strategy, ratings were justified by referencing what could have been done differently to support a different rating or clarification of what the student teacher did not do (i.e., an omission). In some cases, the suggestion for improvement was posed as a question. Examples from the data of the basis for this reasoning strategy included:

- Some learners may have needed further support to process the tasks – given at pace (Q1b. Learners)
- The teacher appeared to be knowledgeable about the subject but didn't make the subject meaningful for the students (Q2b. Content)
- There was very little attention paid to assessment for learning – all questions used were closed questions so missed opportunities to challenge the more able or support those who were struggling. (Q3b. Research)
- How were groups set? Did this support learners? (Q4b. Planning & Preparation)
- For accessibility and learning inclusion, more planning needed (Q4b. Planning & preparation)
- Too much teacher talk, which dominated the lesson. Not enough thought about what she was teaching and what she wanted the students to learn (Q5b. Instructional strategies)
- Some needed prompting and some support – was this because it was too quick for them? (Q6b. Learning environment)
- Not much active engagement from the students as whole (Q6b. Learning environment)
- Were misconceptions about the concepts addressed? (Q7b. Assessment)

Using internal expectation criteria (rationale: personal judgement). This rationale for respondents' judgements appeared to involve underlying personal constructs such as the evaluator's beliefs, value systems, expectations, or even emotions. While relatively uncommon among the strategies used (3.4%), perceptions that come from within the judges themselves were evident. It is important to note that strategies coded as internal criteria may have developed through professional experience; the scope of the data collected did not provide any indication as to whether or not internal criteria were based on professional knowledge or personal preferences. Statements given from participants included:

- I felt that some learners may have been left behind (Q1b. Learners)
- Would have preferred him to have his own copy and for the very busy chalk board to have been considerably simplified (Q2b. Content)
- Felt a little like the instructions were 'thrown' at the group (Q5b. Instructional strategies)
- Did not feel from the video that the learning environment was conducive to learning (Q6b. Learning environment)
- Did not feel that there was active engagement in the lesson throughout (Q6b. Learning environment)

No identified strategy (rationale: indeterminate judgement). Some participants were not able to give a rationale or it was not evident (6.8%). A few tutors restated the dimension

description instead of providing a rationale. Some participants stated they were not in a position to provide a judgement indicated that they needed additional evidence. One did not provide a statement, but instead simply added a full stop (.). Indicative responses included:

- I might have given a higher mark if I had seen all the video (Q1b. Learners)
- It was clear that the teacher had excellent subject knowledge (Q2b. Content)
- The teacher appeared to be knowledgeable about the subject (Q2b. Content)
- Showed some research into the subject being taught (Q3b. Research)
- Would have loved to have been able to talk to the children and look at the end result to see how much progress they had made (Q6b. Learning environment)

6.3.3.3 Mentor Teachers

Table 6.9 shows the range of evidence mentor teachers relied on to judge teaching effectiveness. School-based mentor teachers used three strategies to determine an observation rating using the evidence: (a) classroom cue utilization; (b) suggestions for lesson improvement; and (c) no identified strategy. As mentor teachers reasoned with a given strategy, they appealed to specific justifications (i.e., backing) for the strategy being used. There were two types of justifications evident in the qualitative responses: professional judgement and indeterminate judgement.

Table 6.9

Mentor Teachers' Judgement Strategies and Rationales

Professional judgement		Indeterminate judgement
Classroom cue utilization (n = 92)	Suggestions for lesson improvement (n = 59)	No identified strategy (n = 3)
Observed teacher action (34) Observed pupil action (19) Explanatory rationale (16) Multiple general examples of evidence to support rationale (8) Teacher and pupil interaction (6) Context cues (5) Learning materials (3) Physical environment (1)	Lesson improvement (32) Observed omission (23) Question posed (4)	Restatement of dimension (2) Need more to make judgement (1)

Note. Total codes from qualitative questionnaire statements: Q1–8.

Classroom cue utilization (rationale: professional judgement). Participants utilized perceived aspects of a student teacher's observable practices and relevant classroom cues in decision-making. This strategy accounted for approximately 59.7% of cues. The most common cues were from the student teacher's actions, the pupils' actions, and interactions

between the teacher and pupils. A few examples of classroom cues from observed teacher actions included:

- The level of questioning from the teacher when she went round the groups showed she had high expectations of every student in terms of their intellectual development (Q1b. Learners)
- The questions/task she was asking them to complete related to key blocks they had to consider within the literature (Q2b. Content)
- When one student was unsure what the questions that he was asked to read on the board meant, she asked other students to assist (Q3b. Research)
- Questioning was probing with high-level thinking skills of 'how?' employed alongside interpretation of questions: 'What am I really expecting for you to look at?' (Q4b. Planning & preparation)
- She read aloud learning outcome and got them to repeat it (Q5b. Instructional strategies)
- When working with smaller groups, she used this opportunity to explore individuals roles within the group (Q7b. Assessment)

Pupils' actions and interactions between the teacher and pupils were also used as cues for judging teaching effectiveness. A few of examples were:

- There was peer support in groups, with the teacher asking students what their individual roles were in the group (Q1b. Learners)
- The boy who stood to read aloud from the board appeared to find the wording challenging and required assistance (Q2b. Content).
- The class readily joined in with the choral recital of objectives and key components as to how to summarize, yet application was lacking (Q5b. Instructional strategies)
- They were all very active and engaged in their learning (Q6b. Learning environment)
- She went round the room and checked on them to see they were engaged (Q7b. Assessment)

Another strategy involved a statement of what the teacher did, specifically followed by multiple examples as evidence to support the main statement, or an explanatory rationale was given which confirmed why a particular action worked. For example:

- The teacher was upbeat, confident, and maintained secure classroom control through teacher-led interactions and stages of learning (whole class and group work; Q1a. Learners)
- When she told a group how happy their response had made her, it showed she was developing the independent understanding of the students and encouraging their ownership of understanding the content (Q2b. Content)
- There was a lot of questioning and then further extending students' understanding with probing follow-on questions (Q3b. Research)
- There was a mix of individual work, group discussions and group work. The class were really engaged and enjoying the lessons. (Q5b. Instructional strategies)
- The teacher clearly had a good relationship with the students. She used humour and obviously knew how different students would react, supporting their individual

learning styles. She used group work at the end of the lesson, encouraging students to interact with each other and discuss how they would produce the final piece. (Q6b. Learning environment)

Another main strategy involved a statement related to the context cues and materials from the classroom and the lesson being taught. For example:

- The reading of the handwriting on the blackboard was tricky probably for everyone (Q1b. Learners)
- The books that the students [were] required to refer to when presenting their finding were available on the tables (Q4b. Planning & preparation)
- Start of the lesson instructions were clear (Q5b. Instructional strategies)
- Creating the big paper sheet to show what they understand and make it meaningful for them (Q7. Assessment)

Suggestions for lesson improvement (rationale: professional judgement). This second strategy, which accounted for 38.3% of the strategies coded, was a focused on lesson improvement, building from the mentor teachers' professional judgement. This reflected their role in teacher education as the classroom teacher in a local school. When using the lesson improvement strategy, ratings were justified by mentioning what could have been done differently to support a different rating or giving clarification of what the student teacher did not do (i.e., an omission). In some cases, the suggestion for improvement was posed as a question. Examples from the data included:

- Could there be a better way to present this information to meet the needs of everyone, especially as they had workbooks in front of them? (Q1b. Learners)
- No real evidence of differentiation during the introduction to the lesson. Students were asked closed questions, not allowing them to demonstrate their understanding of the subject matter (Q1b. Learners)
- There seemed to be little emphasis on the students engaging with the texts and discussing their own understanding of the themes (Q2b. Content)
- She had a habit of saying, 'Am I correct?' to the whole class. If a learner wasn't sure, it would be difficult for them to speak up (Q3. Research)
- She should have planned and showed them an example of what she wanted them to produce and planned time to model how to work out the answer and how to put all that you have learnt onto the big piece of paper and how to work as a team (Q4b. Planning & preparation)
- Modelling the layout of the poster board on the blackboard, or using a visualizer, would have given the students more confidence to start (Q5b. Instructional strategies)
- Notably, she did not repeat back what she found to be a good response, and indeed did not address misconceptions/off-beam responses. Potential for assessment through the written task/group work; however, individual responses were not formally assessed, making progress checks hazy. (Q6b. Learning environment)
- Her teaching at the front didn't check understanding; as they were closed questions, all she got them to do was just repeat yes – this doesn't show that they understand (Q7b. Assessment)

No identified strategy (rationale: indeterminate judgement). In a few occurrences (1.9%), mentors restated the domain instead of providing a rationale. One participant expressed they needed additional evidence. Indicative responses included:

- The teacher demonstrated understanding of the subject matter (Q2b. Content)
- The teacher clearly knew her subject (Q2b. Content)
- It depends what her intended outcome of the lesson is (Q3b. Research)

6.3.3.4 Comparison of Judgement-Making Strategies

Comparative analysis was used to examine the pattern of rationales among the groups of judges. Overall, participants relied heavily on the available perceived cues to make judgements of teaching effectiveness, thus demonstrating similarity with attempts to understand teaching performance ('cue utilization validities'). Of the 617 rationales coded from qualitative data across the three groups, 391 (63.4%) reflected the strategy of classroom cue utilization. The same top five strategies occurred across all groups, reflecting little variation in the way decisions to assign a level of performance were justified. Additionally, a further 29.3% of strategies (n = 181) involved suggestions for lesson improvement. Tutors gave suggestions for lesson improvement slightly more often than teacher educators or mentor teachers did. Together with classroom cues, these strategies of judgement-making demonstrated a majority of the backings founded on professional judgement. Only a few instances of warrants based on personal judgement were identified, none of which were exhibited by mentor teachers. This was similar across teacher educators and tutors. In a small number of cases for all three groups, participants decided they needed more information than was available in the video to make a judgement, or they simply restated the main domain description or were uncertain. The exhibition of indeterminate judgement was relatively low overall (5.0%).

6.3.5 Results: Easy and Difficult Dimensions of Judgement

Participants were also asked to indicate which of the seven UNESCO dimensions they found most difficult to judge for the teaching video and which they found easiest to judge. Additionally, they were prompted to *explain why*. Nominal responses are indicated in Table 6.10 for all participants according to their role.

Table 6.10

Participant's Perspective on the Easiest and the Most Difficult Element to Judge in UNESCO Professional Teaching Standards

	Teacher educators (*n* = 13)		Tutors (*n* = 7)		Mentor teachers (*n* = 4)		Overall (*n* = 24)	
	Most difficult	Easiest	Most difficult	Easiest	Most difficult	Easiest	Most difficult	Easiest
Learners	1	4	4	1	0	0	5	5
Content	1	0	0	0	2	0	3	0
Research	6	0	0	0	1	1	7	1
Planning & preparation	1	3	1	1	0	0	2	4
Instructional strategies	0	2	0	3	0	1	0	6
Learning environment	0	2	0	2	0	2	0	6
Assessment	4	2	2	0	1	0	7	2

Note. Questionnaire: Q9–10; only one choice for most difficult and easiest could be selected.

Taking an overview of the judgement-making process, there was a high degree of variation and little consensus within and across groups as to what was most difficult and what was easiest to judge. This was demonstrated by responses from all groups with a limited degree of consensus emerging. Regarding what was easiest to judge, 25% (*n* = 6) noted 'instructional strategies' and 25% indicated 'learning environment'; these were followed by the dimensions of 'learners' (21%) and 'planning & preparation (16.6%); the only dimension not selected as the easiest was 'content'. When asked to explain why, participants indicated this was due to observable and identifiable teaching strategies being used as well as cues from the classroom environment (e.g., positive atmosphere, student behaviour, organization of the physical space). As one teacher educator indicated, 'It was clear to see learners' response and progress through their interactions with the teacher.' One participant did indicate their own area of expertise was assessment so they found that dimension easiest to rate.

When it came to what participants found most difficult to judge, there was slightly less variation, with 29% indicating 'assessment' and another 29% indicating 'research'. No participants indicated that 'instructional strategies' or 'assessment' was most difficult. The greatest agreement occurred for teacher educators, with 46% indicating that 'research' was the most difficult to judge. Reasons for difficulty in judging teaching effectiveness were found to be related to a lack of context (e.g., prior knowledge, lesson sequence, individual needs), limited cues to observe, particularly concerning student engagement and assessment outcomes, the reliance on only observation to ascertain effectiveness (e.g., determining

whether a teacher is using evidence-based strategies can be challenging without explicit information or discussion), and the short time frame that made it difficult to identify practices such as questioning and feedback. Participants also found ease or difficultly of rating to be influenced by their own experience and expertise. For example, a teacher educator stated, 'If it is not your area of expertise or an age range you are familiar with, you can judge it against your knowledge of research about teaching generally but not subject-/age-specific research'.

In the questionnaire, participants were next presented with the prompt 'When making judgements of teaching effectiveness, I …' and given three options to select from based on prior research regarding judgement-making (see Table 6.11). Participants also had the option to select 'other' and write a response. The table shows the approaches of participants for making judgements about student teachers' practices. Results regarding the starting point for making a judgement show that a majority of evaluators look for strengths first and then weigh these against identified weaknesses, reflecting on whether the positives are more important than the negatives. This was indicated by 11 of 24 participants (45.8%). The second most common rationale was to consider the teaching demonstrated against the learning outcomes based on teaching standards. This was used by 8 out of 24 participants (33.3%). No evaluators started from a point of failure and looked for instances to challenge that decision.

Table 6.11

Starting Point for Participants' Judgement-Making

	Teacher educators (n = 13)	Tutors (n = 7)	Mentor teachers (n = 4)	Overall (n = 24)
Start from a point of failure and look for instances to challenge that decision	0	0	0	0
Look for strengths first and then weigh these against identified weaknesses, reflecting on if the positives are more important than the negatives	6	4	1	11
Consider the teaching demonstrated against the learning outcomes based on teaching standards	4	3	1	8
Other	3	0	2	5

Note. Questionnaire: Q11.

Additionally, five participants selected the 'other' option; this included three university teacher educators and two mentors.

When making judgements of teaching effectiveness, the three teacher educators indicated:

- Always look for the strengths and relate these to progression areas. Think deeply about identifying targets which are measurable.
- A combination of points 2 and 3 [i.e., looking for strengths first and considering the teaching against teaching standards].

- Consider the impact of the lesson/teaching on the children's learning and evidence of the teaching standards.

These responses show a focus on the individual student teachers' own growth and development, the combined approach of looking for strengths and weaknesses according to the teaching standards, and the importance of utilizing pupil learning as a component in decision-making.

The two mentor teachers who indicated 'other' stated:

- I consider the outcome/learning intended for the pupils and examine what steps were put in place for all to achieve this. If focusing on a teaching standard, I look for positives to celebrate success and then things the student can work on or how they could have done it differently to be more effective in that standard.
- I use the teacher standards as a basis for all observations, incorporating any identified targets for that lesson and focusing on the positives first, then looking for areas to improve.

These statements reflect a combined approach, considering the teaching against teaching standards and considering pupil learning along with a growth-centred approach based on the student teacher's areas for improvement.

6.3.6 Results: Views on Judgement-Making

The second part of data collection included a questionnaire regarding aspects of judgement-making and influencing factors derived from prior research (see Appendix A2.1). Participants were asked to rate their level of agreement or disagreement with the statements about judging teaching effectiveness. These items were rated on a 7-point scale from strongly agree (7) to strongly disagree (1), with a neutral option (4). The responses to the Likert scale items are summarized in Table 6.12. The analysis examines the perceptions of participants regarding various aspects of judging teaching effectiveness. Specifically, it focuses on the importance of accuracy, consistency, consensus, evidence-based judgement, professional judgement, multiple evaluators, addressing evaluator error, teacher understanding, the contextual nature of judgement, and stakeholder fairness.

Table 6.12

Participant's Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Teacher educators (n = 13)		Tutors (n = 7)		Mentors (n = 4)		Overall (n = 24)	
Statement	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Q12a. It is important that judgements of teaching effectiveness are accurate.	6.62	0.49	6.57	0.49	6.75	0.43	6.63	0.48

Q12b. It is important that judgements of teaching effectiveness are consistent.	6.69	0.46	6.57	0.49	6.75	0.43	6.67	0.47
Q12c. It is important that different evaluators reach consensus.	6.00	0.78	6.00	0.76	5.00	1.00	5.83	0.90
Q12d. It is important that evaluators use evidence to make judgements.	6.54	0.75	6.71	0.45	7.00	0.00	6.67	0.62
Q12e. It is important that professional judgement is used when judging teaching effectiveness.	6.31	0.72	6.43	0.49	7.00	0.00	6.46	0.64
Q13a. It is important that judgements about teaching effectiveness are made by more than one evaluator.	5.46	1.22	5.86	0.99	6.75	0.43	5.79	1.15
Q13b. It is important that potential sources of evaluator error are addressed.	6.15	0.53	6.00	0.76	6.75	0.43	6.21	0.64
Q13c. It is important for the teacher to understand how judgements about their teaching effectiveness are made.	6.62	0.49	6.71	0.45	6.75	0.43	6.67	0.47
Q13d. Judgements are always related to particular teachers at particular points in time and in particular situations.	5.85	1.23	5.86	0.83	4.25	1.79	5.58	1.38
Q13e. It is important that judgements about teaching effectiveness are considered fair by stakeholders.	6.38	0.84	6.29	1.03	6.75	0.43	6.42	0.86

Note. Questionnaire: Q12–13; 7 = strongly agree and 1 = strongly disagree.

Participants agreed on the importance of accuracy, consistency, and evidence-based judgement in evaluating teaching effectiveness, underscoring the need for high-quality

assessment methods. There were no areas in which participants noted disagreement. This is indicated by high mean scores (close to 7) and low standard deviations, in particular for questions Q12b, Q12d, and Q13c across all groups; these three statements had the highest agreement rating. While there was general agreement on the importance of consensus among evaluators (Q12c), the level of agreement was slightly lower compared to other items. The role of professional judgement (Q12e) seemed to be highly valued across all groups. There was also general agreement on the importance of having multiple evaluators (Q13a) and addressing potential sources of evaluator error (Q13b), although there was more variation in opinions among both teacher educators and mentor teachers. There was strong agreement that it is important that the individual being evaluated understands the evaluation process (Q13c), and also that the judgements made are considered fair (Q13e). The item with lowest agreement was for the statement on judgments being about particular points in time and in particular situations (Q13d); this item was rated between the neutral to somewhat agree level by mentor teachers. The view that judgements are specific to context (Q13d) seemed to be more strongly held by tutors and teacher educators. Tutors tended to have slightly higher agreement scores on most items compared to the other groups; in fact, the only occurrence of perfect agreement occurred for tutors in relation to the importance of using evidence to make judgements (Q12d) and utilizing professional judgements (Q12e). Mentor teachers and teacher educators showed a wider range of opinions on some items (indicated by higher standard deviations), particularly regarding the contextual nature of judgements and judgements being made by more than one evaluator.

A high degree of agreement that emerged when comparing scores across groups. Associate tutors tended to have the highest agreement scores across most items, indicating a strong emphasis on the importance of evaluation criteria and evidence. While participants generally supported the involvement of multiple evaluators, there was a notable degree of variability in their responses. This suggests that while multiple perspectives are valued, there may be differing opinions on the optimal number of evaluators or the specific roles they should play. Data also highlighted the importance of addressing potential sources of evaluator error. Participants recognized the need for measures to mitigate bias and ensure that judgements are objective and fair. Analysis also revealed a consensus among participants regarding the contextual nature of judging teaching and the importance of stakeholder fairness in judging teaching effectiveness. Participants emphasized the need for evaluations to be perceived as equitable and unbiased by all relevant parties.

6.3.7 Questionnaire Results: Agreement on Influencing Factors

Participants were further asked to rate their level of agreement or disagreement regarding factors which may influence how evaluators judge. These items were rated on a 7-point scale from strongly agree (7) to strongly disagree (1), with a neutral option (4). The responses to the Likert scale items are summarized in Table 6.13.

Table 6.13

Participant's Level of Agreement With Statements Related to Factors Influencing Judgement

	Teacher educators (n = 13)		Tutors (n = 7)		Mentor teachers (n = 4)		Overall (n = 24)	
Judgement-making is influenced by …	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Q14a. Clarity of the judgement criteria	6.08	1.00	6.43	0.49	6.75	0.43	6.29	0.84
Q14b. Tension of using judgements for both professional growth and accountability	5.54	1.08	5.14	1.55	5.50	0.87	5.42	1.22
Q14c. Clarity of procedures for making judgements	6.15	0.53	6.29	0.45	6.00	0.00	6.17	0.47
Q14d. Individual understanding of effective teaching	6.23	0.70	6.43	0.49	6.50	0.87	6.33	0.69
Q14e. Contested nature of what defines effective teaching	5.77	0.80	6.43	0.73	4.75	1.48	5.79	1.08
Q14f. Professional teaching standards	5.92	0.73	6.57	0.49	6.75	0.43	6.25	0.72
Q14g. Power relationships between universities and schools in teacher education	5.15	1.46	4.43	1.59	6.25	0.83	5.13	1.54
Q14h. Personal intuition about what happens in a classroom	5.46	1.22	5.29	1.28	6.00	0.71	5.50	1.19
Q14i. Perceived levels of importance of different dimensions of teaching	5.69	1.26	5.00	1.07	4.75	0.83	5.33	1.21
Q14j. Complexity of the classroom environment in which judgements are made	5.85	1.17	5.43	1.05	6.25	0.83	5.79	1.12
Q15a. Evaluator tendencies toward leniency or severity	5.00	1.71	5.43	0.90	5.75	0.43	5.25	1.39

Q15b. Personal biases and beliefs of the evaluator	5.08	1.82	4.86	0.99	5.50	0.50	5.08	1.47
Q15c. Experiences of the evaluator from observing other teachers	5.77	1.58	4.86	1.12	6.00	0.71	5.54	1.41
Q15d. Prior interactions between the teacher and the evaluator	4.77	1.67	4.86	1.46	6.00	1.22	5.00	1.61
Q15e. Holding a pre-observation discussion	5.23	1.76	5.57	0.90	5.75	1.30	5.42	1.50
Q15f. Level of involvement of the individual being evaluated in the judgement process	5.38	1.27	5.43	1.05	5.25	0.83	5.38	1.15
Q15g. Training of evaluators to use observation criteria for making judgements	5.62	1.55	6.14	0.35	5.75	1.09	5.79	1.26
Q15h. Observation skills of the evaluator	5.92	1.54	6.14	0.64	5.75	1.30	5.96	1.31
Q15i. Perceptual information (cues) available to the evaluator	5.62	1.50	5.71	0.45	5.25	1.30	5.58	1.26
Q15j. Policies regarding evaluation of teaching effectiveness	4.85	1.56	5.43	1.18	5.00	0.71	5.04	1.37
Q15k. Quality of the reasoning strategies used to make decisions	5.00	1.62	5.71	0.45	5.00	1.00	5.21	1.32

Note. Questionnaire: Q14–15; 7 = strongly agree and 1 = strongly disagree.

There was general agreement across all groups that the factors in Table 6.13 influence judgements of teaching effectiveness. This is reflected in the relatively high mean scores for individual items when considered according to each group and overall. There was strong agreement on the importance of clear judgement criteria (Q14a) and procedures (Q14c), the significance of professional teaching standards (Q14f), and the importance of evaluator training (Q15g) and observation skills (Q15h). The item with the highest degree of agreement and lowest standard deviation related to clarity of procedures for making judgements (Q14c).

There were also variations in responses, indicating different perspectives and priorities among the participant groups. Both teacher educators and mentor teachers exhibited a higher

level of variation than tutors. Items related to personal intuition (Q14h), personal biases (Q15b), and the complexity of the classroom environment (Q14j) showed more variation in agreement, suggesting that these factors are perceived differently by different groups. The influence of power relationships between universities and schools (Q14g) was more important for mentor teachers than it was for teacher educators and tutors. For evaluator tendencies towards leniency or severity (Q15a) and prior interactions between teacher and evaluator (Q15d), there were significant differences between groups.

While there was consensus across all three groups on the importance of several factors influencing judgement of teaching effectiveness, there were also variations. Teacher educators expressed higher agreement on items related to clarity, structure, and professional standards. Tutors showed more variation in responses, particularly on items related to power dynamics, personal biases, and evaluator behaviour. Mentors placed a higher emphasis on the contextual factors influencing judgement, such as personal intuition and the complexity of the classroom environment. Mentors were also more likely to agree on the importance of considering power relationships between universities and schools. This could be attributed to their direct experience in school settings and their understanding of the potential influence of these relationships on teaching practices and evaluation. Additionally, mentors showed more variation in their responses to items related to evaluator behaviour, such as leniency or severity and prior interactions. The differences among the three groups highlights the complexity of the evaluation process and the need to consider multiple perspectives when developing and implementing evaluation systems.

6.3.8 Questionnaire Results: Why Consistent and Reliable Judgements Matter

Finally, participants were presented with an open-ended question related to the overall research aim. They were directly asked: 'Why does it matter that judgements of teaching effectiveness are consistent and reliable?' Responses were analysed using the constant comparative method (Glaser & Strauss, 1967) to develop themes, and findings are presented in Table 6.14. The question was purposely presented after the video task, which engaged participants in a judgement-making exercise and questionnaire that considered influencing factors on judgements.

Table 6.14

Participants' Reasons for Why Consistent and Reliable Judgements Matter

Teacher educators (n = 13)	Tutors (n = 7)	Mentors (n = 4)	Overall themes (n = 24)
Fair (6)	Fairness (1)	Fairness (3)	Fairness (12) & Equity (8)
Equity (6)	High expectations (1)	Equity (1)	Credibility and Standard of the Profession (11)
Quality of teachers entering the profession (4)	Target setting for improvement (1)	Reflect the research base (1)	
Accurate judgements (2)	Ensure pupils have a quality education (1)	Take into account the unpredictability of school (1)	

Consistency (1) Understand own performance and progress (1) Parity (1) integrity of the profession (1) Ensure pupils have a quality education (1) To provide useful feedback for growth (1)	Equality (1) Credibility of the profession (1) High standards to be maintained (1) To improve the quality of teaching (1) Implication of results (1) Clarity about expectations for those making judgements (1)	Support of new teachers (1) Setting a minimum level of competency into the profession (1) Target setting for improvement Implications of results (1)	Teacher Development and Support (5) Implication of results (4) Professional Responsibility (3)

Note. Questionnaire: Q16.

Based on responses from the teacher educators, tutors, and mentors, several overarching themes emerge regarding why consistent and reliable judgements of teaching effectiveness matter. These included: fairness and equity; credibility and the standard of the profession; professional responsibility; and teacher development and support.

6.3.8.1 Fairness and Equity

The most frequent theme was fairness, followed closely by equity; these terms arose in responses from all three groups as to why consistent and reliable judgements matter. This theme emphasizes the importance of ensuring that judgements are unbiased and impartial, and treating all student teachers equally. This is closely related to equity, which highlights the need for judgements to consider individual differences and ensure that all teachers have equal opportunities to demonstrate their skills and abilities. As one teacher educator indicated, it is 'so that everybody is given a fair chance and measuring stick by which to showcase their development and progression'. Another stated: 'Fairness and consistency is key to ensuring students leave our course at the same level and [are] prepared to teach.' A third teacher educator noted it is important 'to ensure all pupils/students receive quality education and learning. To provide equity for trainees/teachers.' Related concepts such as parity, equality, accuracy, and consistency were used by participants to describe the need for fairness. The tutors also responded with an emphasis on fairness. As one tutor stated, 'consistency is incredibly important as this allows all trainees to be evaluated equally'. Another noted that 'everyone should be judged against the same criteria to ensure fairness for all'. Another added the term equality as a foundational concept. Mentor teachers likewise brought forward points in this area, with one stating that 'it is important that all student teachers are treated equally and fairly' and another indicating that 'the notion of fairness during observations is really important'.

6.3.8.2 The Profession

Consistent and reliable judgements of teaching effectiveness were also seen as important for maintaining standards and credibility of the profession, in particular because there are implications for learners. This could be found in responses regarding setting a minimum level of competency and a high standard of quality for new teachers entering the profession. Consistent and reliable judgements were seen as helping to ensure only qualified and competent individuals enter the teaching profession, thus maintaining the reputation of the profession. One teacher educator noted that judgements matter for the 'integrity of the profession'; another mentioned this was 'to ensure a high quality of teachers entering the profession'.

This was confirmed by tutors. One acknowledged: 'We need a profession where teachers are aware of the high standards and expectations in the classroom.' Another added: 'Teaching is a profession and as such, consistent, reliable quality is expected.' In consensus, another mentor noted: 'Unreliable and inconsistent judgements could mean that students are unable to pass their practice fairly.' Accurate and fair judgements based on clear and consistent standards were, thus, seen as important so that all teachers can be evaluated against the same criteria. Evaluation should take into consideration the evidence about what is good teaching, especially amid the unpredictability of the school environment. An essential component of professional responsibility includes accountability for learners' educational outcomes, and consistent and reliable judgements of teaching effectiveness also contribute to ensuring pupils receive high-quality instruction. A teacher educator noted the need 'to ensure all pupils/students receive quality education and learning'. This sentiment was confirmed by tutors, with one commenting: 'All children have a right to a quality education which again can only be achieved if this quality of teaching is monitored and judged against the teaching standards – making it consistent and fair.' These views were further supported by mentors, with one noting that beyond the individual learner, 'the class dynamic and the expected learning' are at stake. By setting and maintaining high standards for the profession, fair and consistent judgements can be made and this was seen as helping to ensure teachers are held accountable for providing quality instruction.

6.3.8.3 Teacher Development and Support

The third major consideration brought forward by participants was the importance of accurate and consistent judgements to help new teachers understand their own performance and progress over time, as well as to provide valuable feedback to help improve practice. In this study, teacher educators considered judgements as significant in order to, 'make accurate assessments of where student teachers are at, and so the feedback they receive is useful development'. Another stated that assessments allow student teachers 'to showcase their development and progression'. A tutors also noted: 'We need a profession where teachers are aware of the high standards and expectations in the classroom which can only be achieved through a cycle of target setting and practice to improve.' By identifying areas for improvement, judgements can help teachers set specific goals and targets for their professional development. As one mentor teacher added:

> The notion of fairness during observations is really important. Being observed is very personal and for a teacher to be given a judgement when they have not been able to discuss their intentions for the lesson and to share their planning thought processes can lead to professional disagreements and unhappy staff. Open and honest conversations where both the observer and teacher understand the context of the class dynamic and the expected learning will support the consistency and reliability of observation judgements.

New teachers are developing their professional knowledge, skills, and dispositions during their school-based experiences; the evaluation process in ITE is a time to reflect on teaching practices and make adjustments as necessary. This includes staying current with educational research, learning new teaching strategies, and engaging in ongoing professional development. This third theme considers the continuous professional learning necessary beyond initial preparation necessary for teachers to be effective in their practice.

6.3.8.4 Comparison Among Evaluators

Overall, the data suggests that there is a general agreement among all three groups about the importance of consistent and reliable judgements of teaching effectiveness. All three groups emphasized the importance of fairness and equity in judgements of teaching effectiveness and their role maintaining the credibility and standards of the teaching profession. All groups acknowledged the importance of consistent and reliable judgements for teacher development and support. Teacher educators and tutors tended to place a slightly greater emphasis on equity, potentially reflecting their closer involvement in ITE. Teacher educators and tutors were more likely to focus on the role of consistent judgements in supporting teacher development and growth. Mentor teachers were more likely to emphasize the professional and moral obligation to ensure consistent and reliable judgements, likely due to their direct experience in mentoring and supporting new teachers in their own classrooms. These findings suggest that a consensus exists across these groups regarding the significance of these judgements for the overall quality of the teaching profession. Findings highlight the crucial role of judgements during initial teacher preparation in supporting teacher development, maintaining high standards, and ultimately ensuring all learners receive a quality education.

6.4 Focus Group Results

Focus groups and individual interviews were carried out to facilitate discussion concerning results of the video observation task and to corroborate judgement strategies and rationales identified through initial analysis as presented in this chapter. Detailed methods are provided in Chapter 2. Data included responses from a small group of teacher educators (n = 2), tutors (n = 2), and mentor teachers (n = 3), all actively involved in preparing future teachers as per their various roles at the time of the study. Results from the constant comparative method of analysis according to each of the focus group questions are presented in this section.

6.4.1 Reasons for Consistency or Inconsistency

Participants were first presented with the question: ‘What could be reasons for consistencies (or inconsistencies) between raters?’ Responses from each group are presented in Table 6.15.

Participants provided many possible reasons for the inconsistency among raters captured by the questionnaire. The two teacher educators who participate provided 20 potential reasons for consistency or inconsistency in judging the quality of a student teachers' teaching. These focused on the evaluator and processes associated with making judgements of teaching practice. Teacher educators who will have had similar teaching experiences and have been through similar training to conduct observations were considered more likely to have consistent views on effective teaching; this appeared to help develop a common understanding of teaching and learning. Furthermore, teacher educators are expected to know and model effective teaching practices, and thus they are more likely to recognize and appreciate others' practices. There was also a recognition that while context is important, there are certain fundamental principles of teaching and learning which are universal. As one teacher educator articulated: 'There are really core elements of effective teaching and learning. And if people have been trained correctly and they model it in their classrooms, we'd call it our bread and butter in this country' (F3I1).

Participants also gave insights regarding inconsistencies. Despite shared experiences, individual teacher educators also have unique perspectives based on their personal experiences; teaching effectiveness can be subjective, and different individuals may have slightly different criteria for what constitutes outstanding teaching. Participants thought that the diversity of perspectives among teacher educators can be beneficial, as it can prevent a 'one size fits all' approach to teaching and learning. One participant stated:

> we'll all have our own subtle differences based on our own experiences as teachers…. And that's what makes teaching so great. It's unique, it's individual, there isn't a one size fits all. My worry is if there was a one size fits all and we all did come with the exact same opinions and judgements every time. (F10I1)

A willingness to consider different viewpoints through dialogue was seen as a way to help mitigate inconsistencies in judgements. Additionally, regular contact, meetings, and opportunities for sharing and support among teacher educators were put forward as contributing factors to consistency of judgements.

Table 6.15

Reasons for Consistency and Inconsistency Between Raters

Teacher educators (*n* = 2)	Tutors (*n* = 2)	Mentor teachers (*n* = 3)
Evaluator centred: • Classroom teaching experience • Own teacher training • Not all teacher educators are experienced teachers • Raters expectations of what to see in a good lesson • Personal judgement based on internal criteria • Modelling of good teaching for students • Understanding context • Evaluator's knowledge of the subject/content • Evaluator's understanding of processes and procedures • Self-efficacy in judgement-making • Openness and willingness to share and to support each other • Different paths taken before being a teacher educator • Years of experience • Different perspectives	**Evaluator centred:** • Variability in engagement of mentor • Impact of attitudes and commitment • Role of rater's experience • Need for training and support • Need for specific observational training • Challenges in school settings • Importance of setting clear and achievable targets • Differences in training and perspectives • Recognition of fundamentals **Student teacher centred:** • Recording and providing feedback on progress • Importance of consistency in mentoring • Consistency and universality of effective teaching • Accessibility • Inconsistencies and contextual factors • Consideration of cultural differences	**Evaluator centred:** • Individuality and interpretation • Consistency in evaluation standards • Honesty and authentic feedback • Concerns about inconsistency in evaluations • Subjectivity in evaluation **Student teacher centred:** • Performance evaluation and anxiety • Perceived expectations and pressure to conform • Focus on student-centred assessment • Balancing pedagogy with student engagement • Value of constructive feedback and diverse perspectives. **Aspects of the teaching:** • Positive reinforcement in teaching • Critique of overbearing teaching styles

• It's subjective as to what is viewed as outstanding and quality teaching • All evaluators have own subtle differences • Teaching is not one size fits all **Processes:** • Consistency in views of teaching and education among the [university] staff • Regular contact and meetings • certain agreed things that need to be in every lesson	• Recognition of universal teaching standards **Aspects of the teaching observation:** • Teacher effectiveness • Planning and preparation **Processes:** • Areas for improvement • Consistency in teaching evaluation • Role of guidelines and standards • Teacher standards and expectations • Differences in training and standards	• Complexity of teaching and evaluation **Processes:** • Contextual influence on evaluation • Subjectivity and focus in evaluation criteria • Importance of simplified and focused observations • Consistency in evaluation judgements • Overcoming bias and pressure

Note. Focus group: Q1.

The two tutor participants noted 23 potential reasons why there could be inconsistency in judging the quality of student teachers' practice. These were organized into four main categories to explain the inconsistencies among those who judge student teachers' effectiveness through classroom observation. These focus on the tutor as the evaluator, the student teacher, aspects of the teaching lesson being observed, and processes associated with making judgements of teaching practice. Tutors noted that evaluators' own teaching experience in primary or secondary education and their expertise in the specific subject being taught may have influenced perspectives; those with experience in the content area reflected in the video may have a more nuanced understanding of the student teacher's performance. Additionally, participants noted that individual raters may have personal biases that influence their judgements, but also that the presence of the observer can impact the teacher's performance and, consequently, the rater's evaluation.

Participants also noted the inconsistencies could be due to variations in rater training and experience, or the lack of standardized guidelines, or raters may use different criteria or interpretations. As one tutor indicated, 'there should be some conformity across the board if they are using the guidelines that the university set' (F7I1). Further, the tutors also noted that understanding the specific context of the lesson, including the age group, subject, and learning objectives, can influence rater judgements. Clear communication and shared understanding among raters can help improve consistency as well as collaborative discussions and peer review. As one tutor noted, 'I guess the fundamental things that make a teacher effective are always the same' (F11I1).

According to the three mentor teachers, there were several potential reasons why consistency or inconsistency in judging the quality of student teachers' teaching would occur. These were organized into four main categories to explain the inconsistencies found among mentor teachers. These focus on the evaluator, the student teacher, aspects of the teaching lesson being observed, and processes associated with making judgements of teaching practice. The focus was on the variability of the mentor teacher as an evaluator, the student teacher themselves, and their teaching, and the process by which a valid judgement is made. Mentors expressed that the anxiety and pressure that comes with being assessed or evaluated closely, particularly in situations where one's responses are being scrutinized, is a real challenge. This theme reflects concerns about meeting expectations and saying the 'right' things during evaluations. The importance of emphasizing positive aspects and giving constructive feedback to student teachers was highlighted. As one mentor teacher explained:

> you're always kind of like three stars and a wish kind of scenario. When you're looking for the positive, you're trying to give positive feedback, and then giving one small element that needs work on, to be worked on, rather than thinking, 'oh, this needs doing, this needs doing, this needs doing'. It's trying to be positive with the student and making them feel confident. So, you're trying to pull out the positive strands. (F5I1)

The concept of 'three stars and a wish' represents a method of focusing on strengths while also addressing areas for improvement, aiming to boost student teacher confidence and

motivation. The recognition that individuals have unique perspectives and interpretations is acknowledged. This theme touches on the potential for variability or inconsistency in data interpretation but underscores the need for alignment and consistency in assessments between different evaluators (university versus mentor). Ensuring consistency in evaluation standards and aligning expectations between universities and mentors to ensure fair and accurate assessments of teaching performance were key drivers for mentor judgement-making.

There was a concern about the difficulty of providing honest feedback in evaluation scenarios and mentors expressed hesitancy about giving negative opinions or critiques, suggesting a fear of not being allowed to be truthful about their assessments. One mentor teacher explained this tension:

> It was kind of a bit worried about actually being honest about how we felt about the teaching. So, I think that potentially somebody might sit there and watch that video and think, 'maybe I'm supposed to think this is good'. So it might be that people are watching it thinking that they want to give the answer they think you want to hear because of the situation, rather than actually being more natural. (F5I2)

This highlights a concern that individuals might feel pressured to provide responses they believe are expected rather than expressing their genuine opinions. This pressure could lead to evaluations that do not accurately reflect true assessments. The discussion raises the point that the context in which evaluations occur can influence judgements. The simulation of the judgement-making task in the study (i.e., watching a video for evaluation purposes versus observing in a classroom setting) elicited different reactions and assessments, suggesting the importance of considering context in evaluations. The subjectivity inherent in evaluation processes underscores the need for clear and consistent evaluation criteria. A strong concern was expressed about significant inconsistencies in evaluation outcomes. The potential for teaching to be described as 'really good' and 'really bad' at the same time raises doubts about the reliability and fairness of the evaluation process despite clear assessment criteria. Overall, the mentors highlighted the complexities and challenges involved in conducting evaluations, including the importance of honesty, the influence of perceived expectations, the contextual nuances of assessments, and the need for consistency and clarity in evaluation criteria to ensure fair and meaningful feedback.

The mentors emphasized that teacher evaluations can be highly subjective and influenced by personal perspectives. Different mentors may focus on aspects such as the teacher's personality, confidence, or relationships with student teachers, leading to varied interpretations of teaching effectiveness. One mentor suggested that mentors should assess teaching effectiveness from the perspective of student teachers; they should consider what student teachers have learned and how they have absorbed the material, rather than focusing solely on the teacher's performance.

There was criticism of teaching styles that may be perceived as overbearing or demanding. The mentors mentioned concerns about excessive scaffolding and lack of student teacher autonomy, suggesting that effective teaching should empower student teachers to think critically and work independently. Successful classroom observations should focus on

specific elements of teaching rather than overwhelming observers with multiple considerations. Simplifying the evaluation process could lead to more meaningful assessments and targeted feedback. The mentors critiqued teaching methods that prioritize pedagogical terminology and presentation over student teacher comprehension and engagement. Effective teaching was seen as a balance between clear teaching and fostering meaningful student teacher learning experiences. Furthermore, mentors suggested that diverse feedback approaches can offer valuable insights and help teachers improve their teaching practices. While acknowledging the subjectivity of evaluations, they also stressed the importance of consistency in judgement.

Overall, these ideas underscored the complexities and challenges of evaluating teaching effectiveness, emphasizing the need for student-centred assessments, constructive feedback, and a balanced approach to judging teaching practices. The participants also advocated for simplified and focused observation techniques that support meaningful professional development for educators.

6.4.2 Possible Ways to Gain Consistency

Next, participants were presented with the question: 'What would make judgement among evaluators more consistent?' Participants provided several suggestions as to how greater consistency could be gained. Responses for each group are presented in Table 6.16.

Table 6.16

Strategies to Gain Consistency in Judging Teaching Effectiveness

Teacher educators (n = 2)	Tutors (n = 2)	Mentor teachers (n = 3)
Preparation: • Training for observers • Understanding nuanced points **What is being judged:** • Collectively agreed what you are looking for • Clarity of criteria **Making the judgement:** • Using criteria of teaching standards • a tick box which says exactly what you need to look for • very rigid checklist of criteria **Not attainable:**	**Preparation:** • Observational techniques • Creating a safe learning environment • Understanding student teacher needs • Contextual understanding **What is being judged:** • Comprehensive evaluation **Communicating judgements:** • Feedback and assessment methods • Mentorship and feedback • Reflection and learning	**Preparation:** • Importance of mentor training • Collaborative learning and discussion • Challenges with professional development **What is being judged:** • Diverse teaching methods • Shift from lecturing to teaching • Flexibility and adaptability in teaching **Making the judgement:** • Structured evaluation criteria

• There will always be disparity • There are different views about what good teaching is about • Years of experience will be relied on • Teaching experience paired with human qualities (dispositions) • Still different opinions	• Consistency through moderation • Adaptation to remote learning • Consistency and collaboration **After the judgement:** • Need for mentorship and coaching • Importance of reflection and professional development

Note. Focus group: Q2.

Teacher educators suggested several ways that judgements might be made more consistent. The first was training for observers to ensure they have a shared understanding and clear, consistent criteria for evaluating teaching effectiveness. Participants also noted that open discussions about the nature of effective teaching to address philosophical differences would be helpful, as this could help recognize the value of professional judgement while still ensuring judgements are made against clear criteria. There was an emphasis on acknowledging the subjectivity of judgement-making and recognizing the limitations of consistency. It was noted that valuing the diversity of perspectives among teacher educators and recognizing different viewpoints can contribute to a more comprehensive understanding of effective teaching. As one teacher educator stated, 'consistency, I think it can be aspired to if that's what your aspiration is. But I think there are some, there will always be philosophically different perspectives about what good teaching and learning looks like' (F3I1). Another articulated: 'I'm not sure you'll ever … get it consistent because as people, we all have different thought processes at different times, and we've all got vastly different experiences' (F10I1). Participants did suggest that by providing clear guidelines, fostering open dialogue, and leveraging the expertise of teacher educators, it is possible to develop more consistency in judgements.

The tutors contributed three potential ways to make improvements in preparation, processes and, in particular, provide feedback for growth. These areas reflected suggestions made by the other groups but were much more focused overall on helping the student teacher to develop their teaching through consistent feedback. As on mentor teacher noted:

> So, the mentor has got to be in a position where if the student is doing something, give them a chance to reflect well and then give them good strategies to make sure that they can put those things in place. Because if a student's not doing it, it's because they don't know. And I think that if that kind of thing was happening across the board, I think there would be more consistency. (F7I1)

The themes suggest a nuanced understanding of the complexities involved in teacher training and evaluation, while also recognizing the importance of core standards and expectations in

effective teaching. The themes emphasized the need for a holistic approach beyond observation to evaluating student teachers, incorporating context, feedback, and multiple assessment methods to ensure consistency and accuracy in assessments. As one mentor teacher noted, 'it very much comes down to that understanding of the trainee and speaking with the mentor as well' (F11I1).

Mentor teachers also brought up consideration of what occurs after the judgement is communicated. Mentors suggested breaking down evaluation elements into a checklist for specific criteria rather than relying solely on subjective judgements. The idea was that this approach would introduce more objectivity by assessing whether specific teaching elements are present in a lesson. Furthermore, having moderation involves bringing mentor teachers together to watch the same lesson and discuss their observations, which allows for the standardization of evaluation criteria and ensures that mentors are aligned in their understanding of what to look for in teaching assessments. As one mentor teacher suggested, 'with the mentor training, doing some sort of moderation, where you actually bring a group of mentors together in a similar situation to we were. And like watch the same lesson, and actually discuss it' (F5I1).

The conversation highlights the significance of mentor training programmes in providing guidance and standardization of evaluation, aiming to clarify expectations and align assessment practices. Also, the value of collaborative learning among mentors through watching and discussing teaching videos together was brought forward. This approach would foster a shared understanding of assessment criteria and help mentors develop a clearer sense of what the university expects from teaching students. Training needs to go beyond logistical aspects and focus on enhanced consistency and clarity in mentor teacher evaluations through structured criteria, collaborative learning, and standardized approaches. By implementing clear evaluation guidelines and promoting shared understanding among mentors, universities can better support the development of effective teaching practices among student teachers. The discussion touched on the variability in teaching methods across schools. One participant explored this tension:

> There's a recognition that different schools approach teaching differently. I think every school teaches very differently. And I think with the Rosenshine principles and things like that, you know, every school's looking for something different. Maybe, doing it with somebody – if you've got somebody else there with you, that might mean that it's more consistent. (F5I2)

This suggests a theme of diversity in educational approaches and the challenge of standardizing methods across different institutions. This feeds into the idea that involving others or working with colleagues could enhance consistency in teaching practices. Collaborative efforts might help align different approaches and provide support in implementing new methods or standards. There are challenges in engaging with professional development resources, such as teaching videos, due to time constraints and workload. This reflects a broader theme of the practical challenges teachers face in integrating professional development into their busy schedules and the need for more accessible and integrated

approaches within school time. This would be difficult in the context of the complexities and practicalities of educational reform and professional development within a diverse school environment, where aligning practices and integrating resources effectively can be challenging.

6.4.3 Perceptions of Inconsistencies From Video Task Results

The third question posed to participants related to initial findings from the video task and queried perspectives regarding the dimension of teaching which yielded the greatest degree of variation of ratings. The question was: 'What are your thoughts on the finding that [name of dimension] had the most inconsistent/consistent rating?'

Teacher educators noted the following views on the finding that 'learners' had the most inconsistent rating yet was considered the easiest dimension to rate:

- instinct plays a role
- observation cannot capture everything
- cultural context
- seeing the learners
- observation is a multi-sensory experience
- care taken in making the judgement
- some elements of making judgements about lessons which are more individual
- behaviours are difficult to make a judgement on
- personal and strong views of the evaluator
- instruction easier to judge than learner engagement (teacher centred)
- could see what was going on in the class, but not at a deep level
- evident through interactions
- easier to see than something such as planning and preparation
- observable cues made it easier to judge
- other dimensions are more unseen (e.g., research)
- some aspects of teaching occur over a period of time
- some dimensions of teaching require additional sources of evidence (e.g., lesson plan).

While the 'learners' dimension might seem straightforward to assess, participants recognized the factors involved in making accurate judgements, such as subjectivity, limitations of observation, and the nature of the dimension itself, which contributed to the variation in ratings. One teacher educator explained:

> some elements of making judgements about lessons are more individual … what does good learning look like from the learner's perspective? What should be considered more in the judgement-making … you know, learners that sit in a classroom stare out of the window, fidget, draw doesn't mean they're not learning. But those behaviours are much more difficult for the observer to make a judgement on. You know, how do you know that somebody is actively learning? It's very difficult to make a judgement. (F3I1)

For the tutors who completed the video task, there was consistency in rating the dimension of 'instructional strategies' as the easiest to rate, while 'learners' was consistently selected as most difficult element to judge in the video. Participants articulated that this consistency could be due to knowledge and information on assessment and understanding of context. The tutors shared the following points regarding consistency of responses:

- ease of assessment
- challenges in assessing learners' needs
- assessment challenges and methods
- importance of context
- observability of curricular knowledge
- observability of behaviour management
- observability of teaching practice
- role of questions and communication

Overall, the reasons reflected agreement with certain ideas, recognition of routine mentoring tasks of a tutor, and a desire for clarification on specific terms or concepts. The themes revolved around the dynamics of mentorship, including the potential influence of mentor experience on their effectiveness in supporting student teachers. Furthermore, one of the themes was the need to highlight the role of the tutor in quality assurance and the process of observation and feedback, and the importance of discussion and review for enhancing teaching effectiveness. As one tutor noted from their experience:

> My role as link tutor when I go in is quality assurance. So, I make a visit into the schools with each of the students once the placement [starts] and I do a paired observation with the mentor and I will speak to the student. I will speak to the mentor. And most importantly, I listen to the feedback that the mentor is giving to the student. Then we will discuss that feedback. (F7I1)

There was some consideration of consensus and agreement, the routine aspects of mentoring, understanding of learning and development, the impact of mentor experience, the role of the link tutor in quality assurance, and variation in mentor approaches. Also, the challenges involved in assessing teaching effectiveness were expressed, particularly regarding the availability of context and the observable aspects of instructional strategies versus the more nuanced understanding required for assessing learners' needs. There was also some consideration as to the importance of observation and communication in gathering evidence for assessment and the observability of teaching practices during sessions.

There was no clear pattern in the dimensions mentor teachers found the easiest and most difficult to rate. This variability was shared with the mentor teachers who were interviewed, and they voiced the following thoughts possible reasons for this:

- visibility of planning
- nature of learning environment
- assessment challenges
- impact of video format

- contextual understanding
- experience and variables
- context and relationship
- challenges of judgement
- subjectivity in evaluation
- comparison of real life versus video observations
- measurability and evidence-based judgements
- difficulty in judging transitory elements
- subjectivity and feasibility of feedback
- variation in mentor evaluations.
- role of experience and expectations
- personal factors in assessment

Mentor teachers noted the difficulty of assessing the planning aspect of the lesson since it was not explicitly shown in the video. The absence of visible planning made it challenging to judge the overall coherence and effectiveness of the lesson sequence. There was acknowledgment of the constraints of observing a lesson solely through video, and they emphasized that being physically present in the classroom might provide a more comprehensive understanding of the learning environment and student engagement. Also it was noted that having more context, such as on the student teacher's work progression and the overall assessment criteria, would add to the depth of assessment. As one mentor noted, 'the things that are harder to judge, I think, are you can't always see everything that's going on in the classroom' (F6I1). Furthermore, the video format might have affected the mentors' ability to evaluate certain teaching elements, such as the learning environment and assessment practices, suggesting that a live observation might yield different insights.

There was emphasis on the importance of context and personal relationships in understanding and assessing student teachers. One mentor teacher mentioned that building a relationship with student teachers in a classroom setting allows for a deeper understanding of their behaviour, mood, and overall performance. As another mentor shared: 'There wasn't a context around it. When you have a student in your classroom, you build a relationship with them. And, you know, if they're having a good day, a bad day, you know that person better' (F5I1). This contrasts with the use of video observations to make judgements, as such context may be lacking. These factors could lead to difficulties in making accurate judgements about student teachers in daily classroom interactions. The mentor suggested that understanding student teachers' behaviour and progress on a day-to-day basis can be challenging, especially when trying to assess various aspects of their performance and development. Subjectivity of feedback was also mentioned. The mentor acknowledged that providing timely and meaningful feedback requires the ability to make sound judgements, which can be challenging when certain aspects are less objective. Overall, mentors recognized the complex nature of student teacher assessment and evaluation, highlighting the importance of context, personal relationships, and the subjective nature of judgement in educational settings.

6.4.4 Professional Judgement and Professional Standards

The fourth focus group/interview question asked participants: 'What are your views about using professional judgement and professional standards to judge teaching effectiveness?' Responses from each group are presented in Table 6.17.

According to the teacher educators, there is a need to use both professional standards and professional judgement when assessing student teachers' practices. As one participant articulated: 'If there are no standards, then I think we're maybe discrediting or de-professionalizing a vocation of teaching, which is really, you know, it's a special career and a job to have' (F10I1). The responses suggest that both professional judgement and standards are essential for assessing teaching effectiveness. Standards provide a framework for consistency and evaluation, while professional judgement allows for flexibility, context-specific considerations, and personalized feedback. A balanced approach that incorporates both standards and professional judgement can lead to more comprehensive and meaningful evaluations of teaching practice.

The tutors also brought forward the complex interplay between professional judgement and formal criteria when evaluating teaching effectiveness. Their views revolved around the importance of clear standards, achieving progress, maintaining consistency, and effective communication between student teachers and mentors in teacher training programmes. One tutor described this standards-based approach:

> Leeds have broken those down very, very nicely into extremely understandable chunks and what they expect the students to achieve over three different placements. And then when they get part way through the middle of the final placement, then we work more towards the teacher standards. So, there are very, very clear expectations on those students. The mentors know what they have to achieve, and the students know exactly what they have to achieve. (F7I1)

The interconnectedness of professional judgement and teaching standards in assessing student teachers was highlighted. It was emphasized that each is essential and that they should be used together to ensure fairness and robustness in judgement-making. As one tutor articulated:

> I very much do go back to using our standards to support my judgements. Just to make sure that it's fair and robust, and making sure that we're making those fair professional judgements about where the trainee is. I do think you can't wholly just use your own kind of professionalism to observe and grade a trainee. I think it's really important that you do use those standards as well. (F11I1)

Collaboration with mentors and school leaders was also viewed as crucial for informing professional judgements. There was an emphasis on the importance of using standards to ensure fairness and consistency in assessments, with tutors taking the role of moderator, seeing that judgements across different schools are consistent. Tutors described their role as overseeing assessments and ensuring that judgements align with standards and are fair to the student teachers.

Table 6.17

Participants' Views on Professional Judgement and Professional Standards

	Teacher educators (n = 2)	Tutors (n = 2)	Mentor teachers (n = 3)
Standards	• Bring a level of consistency • Used as criteria by a range of professionals • Important for the receiver (of the judgement) to know areas for improvement • Form a base for what it is being looked for • Set a minimum requirement • Bring credibility to the profession • Help inform the dialogue	• Provide a framework for consistent evaluation • Ensure judgements are based on shared criteria • Serve criteria against which teaching can be measured • Help identify areas for improvement and constructive feedback • Establish a baseline or minimum expectation for effective teaching • Contribute to credibility of the teaching profession • Inform discussions about teaching practices and professional development	• Effectiveness of teaching standards • Usefulness of teaching standards for support • Role of assessment tools • Development of practical resources • Granular approach to feedback and teaching. • Impact of professional standards • Role of teaching standards
Professional judgement	• Demonstration of standards is different across a teacher's career – growth • Effective teaching is an iterative process – never a final product • Important for giving feedback • Strengths-based approach • Can limit the number of targets • Using criteria is helpful • Brings context into play • Prevents a one-size-fits-all approach	• Teaching is a dynamic process • Effective teaching varies across teachers' careers and experiences • A continuous learning journey – judgements should reflect this ongoing development • Essential for providing targeted and meaningful feedback • Allows for a focus on strengths and areas for growth • Considers the context of the teaching situation	• Subjectivity and reliability of judgements • Impact of grading and evaluation stress • Professional standards and grading • Differentiation of student progress • Professional judgements based on experience • Differentiation based on student needs

	• A necessity due to the nuance of the judgement-making task • Requires strong relationships • Requires being able to see what occurs in a classroom • Considers experiences of the evaluator • Prevents depersonalizing the process • Facilitates open, honest dialogue	• Prevents a one-size-fits-all approach to evaluation • Is necessary to capture complexities and nuances of teaching • Relationships between evaluators and teachers are crucial • Direct observation of classroom practice is essential for informed judgement • Evaluator's own experiences and expertise can inform judgements • Helps avoid a depersonalized evaluation process • Facilitates open communication	
Need for both			• Challenges of mid-placement reviews • Role of practising teachers in refinement • Importance of clarity and real-life scenarios • Practical guidance and support • All for teacher-driven refinement • Proposal for comprehensive resources

Note. Statements from focus group: Q4.

Mentor teachers demonstrated a similar overall consideration of professional standards and professional judgements. However, mentors considered to a greater extent the stress and pressure associated with aiming for specific grades or performance levels, similar to the stress experienced with evaluation systems like Ofsted. There was recognition that while evaluation systems can be motivating, they can also induce stress and anxiety among student teachers and educators. The mentors expressed scepticism about the objectivity and reliability of judgements in educational evaluations. They highlighted concerns about the validity of feedback and the variability of assessors' opinions, emphasizing the need for more constructive feedback models to balance positives and areas for improvement. As one mentor teacher articulated:

> I think there needs to be some guidance because, without it, it's too arbitrary. However, I think there needs to be more support perhaps put in place for mentor teachers to help make the correct judgement. It's okay having a list of standards, but I think what falls short is that middle ground where you are giving advice on how to meet the standards and it's the ability to feed back. (F6I1)

Another mentor teacher also shared from their experience:

> I find when I've got a really good student, I can make a much easier professional judgement about where I expect them to be at the end of their teaching practice, compared to if I have someone who needs a bit more support, I find the teaching standards help me then really to break down what they need to work on next. So, actually sort of moving up that ladder. And even like with the Leeds Beckett things, where you've got their sort of not on track, or they're working towards or they've already met it, or they've exceeded it. (F5I1)

This pulls on discussion about the role of professional standards and grading in guiding professional development, suggesting that while standards are helpful in identifying areas for improvement, the grading system can be challenging and may not adequately differentiate progress across different academic levels. The challenge in differentiating student progress based on standardized evaluation criteria, especially when comparing student teachers across different academic years or levels, lies in applying uniform evaluation standards to student teachers at varying stages of their academic journey. In contrast, there was also recognition of the benefits of teaching standards in providing guidance for professional development and as a valuable tool for identifying specific areas of teaching that require improvement. Overall, these themes underscore the complexities of evaluation systems in education, emphasizing the need for more nuanced and constructive feedback models that support growth and development while addressing the challenges associated with standardized grading and assessment. Mentors recognized the importance of teachers' professional judgements based on their experience and understanding of student teachers' abilities and progress. Teachers used their knowledge to assess student teachers' potential and needs even before formally applying teaching standards. The mentors discussed how teaching standards helped differentiate teaching and support based on individual student teacher needs. For student teachers who require more support, teaching standards provide a framework to identify areas

for improvement and set achievable goals. The teaching standards were viewed as helpful tools for providing targeted support and boosting student teacher confidence. However, it was also noted that there needs to be more support and guidance for teachers to apply them effectively in practice. There was a call for development of comprehensive resources for the Ofsted process, such as booklets or guides that break down standards into actionable steps and provide clear examples of teaching strategies. The goal is to empower mentors and observers with practical tools to help them support student teachers in meeting professional standards.

Overall, these themes underscore the need for comprehensive support systems that go beyond merely listing standards, providing practical guidance, resources, and tools for teachers and mentors. The emphasis is on developing accessible and actionable strategies that empower educators to apply standards effectively in diverse classroom settings, ultimately enhancing the quality of teaching and learning experiences.

6.4.5 Universities and Schools Working Together

The fifth question posed to participants was: 'How might schools and universities work together to gain greater reliability in evaluation of teaching effectiveness?' Teacher educators outlined suggestions to prioritize areas for improvement and to develop strategies for enhancing processes. Together, the two teacher educators noted the following considerations.

Systems changes:

- better student-to-staff ratio
- funding
- reduce workload
- intervene in significant school issues (e.g., mental health)
- address the pressing issues so working together for teacher education can then occur
- new government

Practices:

- leverage talent and knowledge of colleagues
- clear expectations of roles
- training to work with schools
- training for teacher educator position
- do context-specific, local calibration exercises
- standards help as a core level
- context is key and every school is so vastly different
- relationship between tutor and schools
- conversations with mentor teachers

Participants shared they found the current educational situation very difficult. As one teacher educator noted, 'colleagues in schools and universities in England are on their knees because there's too much expected of them' (F3I1). Another acknowledged the complexity of identifying quality teaching:

> Whilst there may be some commonality between what good teaching is, what good teaching is and works for one class in one school with Year 3 might not work quite as well with another class in Year 3 that's just a half a mile down the road … I don't think we could ever narrow down to a specific list of ways of looking at it … I think if we did, the danger could always be are we taking the autonomy and creativity away from teachers to be themselves and explore things in their own way? (F10I1)

The tutors who were interviewed outlined the following ways that schools and universities can work together to gain greater reliability in evaluation teaching effectiveness:

- partnership between university and schools
- consistency and fairness in assessment
- experience as a mentor and class teacher
- training development and encouragement for mentors
- the importance of relationships
- university–school partnership
- consistency and fairness in assessment
- personal experience as an educator
- training development and encouragement

The tutors emphasized the crucial role of relationships, particularly between educators and schools. They expressed that building strong relationships is essential for effective collaboration and support, with trust highlighted as a cornerstone of these relationships. Collaboration between educators and schools was viewed as vital for success. As one tutor shared, 'relationships are absolutely crucial … if you can build up a relationship with the school and they know that they can trust you, then they will work hand in hand with you' (F7I1). Tutors discussed the support they offer to both successful and struggling student teachers; they serve as a point of contact for mentors and schools, providing assistance, guidance, and even support plans when needed. There was indication of the tutors' commitment to their role as a link tutor, demonstrating professionalism and a sense of responsibility towards student teachers, mentors, and schools. Overall, the significance of nurturing relationships, fostering trust, and providing support within the educational context were emphasized. Also, the link tutors highlighted the partnership between the university and local schools. They noted the development of training for mentors, indicating that it has become more robust and that there is active encouragement for mentors to participate. They highlighted the importance of 'the same message being passed on' (F11I1). However, they also acknowledged the challenge of finding time for training amid the busy schedule as a class teacher. There was emphasis on consistency and fairness in assessing student teachers. They suggested that the investment in training mentors and the fact they are assessing student teachers year after year helps ensure that judgements are fair and consistent.

Collectively, participants saw the value of professional standards in providing a clear framework for judging teaching and serving as a benchmark to gauge competence and promote consistency. The standards also help student teachers understand expectations and areas for improvement. The role of professional judgement and the need to take a holistic

view and often make quick, intuitive assessments was brought forward too. Professional judgement was viewed in light of the value of teaching experience. Professional judgement and standards appear to be seen as complementary, with both essential when judging teaching effectiveness.

6.4.6 Barriers and Assets for Working Together

The sixth focus group/interview question was: 'Is there any barrier or asset you would like to raise attention to that would impact working together?' The barriers and assets that impact schools and universities working together for reliable judgements are provided in Table 6.18.

Table 6.18

Barriers and Assets in Collaboration

	Teacher educators (n = 2)	Tutors (n = 2)	Mentor teachers (n = 3)
Barriers	• Must be a benefit to the school • Some subjects are marginalized • Competing priorities • Teachers are exhausted	• Challenges with problematic student teachers • Impact of university support or lack of support • Relationship between university and school	• Support from school leadership • Time management and availability
Assets			• Professional growth through mentorship • Professional growth through observations and feedback • Value of collaboration and support • Reflective practice
Both			• Responsibility and preparation • Partnerships between universities and schools • Cross analysis of core ideas

Note. Focus group: Q6.

Only one teacher educator provided a response to this question, and they expressed barriers but did not mention any assets. Regarding the barriers, the circumstances around time and workload constraints were highlighted. The teacher educator explained: 'It isn't because the

teachers don't want to do it. It's because they have so many competing priorities and they're exhausted' (F3I1).

Additionally, one tutor provided a response to this question, raising the potential challenge of problematic student teachers on school placement. They suggested that while the problems might be related to academic performance, they could also encompass other issues. They highlighted the importance for the school of support from university staff in addressing any issues and suggested that if the school does not receive adequate support from the university in dealing with such situations, it can affect the willingness to host student teachers in the future. The tutor shared:

> I've had this conversation, it makes them very, very wary about having students again. And that's extremely sad, because it may be that's a school that is very, very good with students and, does a good job in helping them along the way on their career path. (F7I1)

Mentor teachers further provided a viewpoint on barriers, yet also contributed potential assists to collaboration. Becoming a mentor was seen to contribute to personal growth as a teacher. As one participant shared: 'I think actually becoming a mentor has made me a better teacher because you do, you delve down into the nitty gritty of everyday life … It helps you to look at your own practice and evaluate yourself' (F5I1). The role requires mentor teachers to pay attention to details and be more organized in their own teaching practice, which ultimately benefits their overall effectiveness as both teacher and mentor. There was an emphasis on the importance of responsibility and preparation in teaching. The mentors highlighted the necessity of having lesson plans and materials ready in advance, which not only facilitates smooth teaching but also supports continuity in case of unforeseen circumstances like illness. By engaging in mentorship, teachers are prompted to evaluate their own teaching methods and practices more critically, leading to continuous improvement and refinement of their professional skills.

Some challenges were raised, such as balancing time, especially during busy school hours, and the need to use available time effectively by arranging interviews or engaging in professional development activities like research outside of regular teaching hours. As one mentor noted:

> getting in school time and getting the time from and having support of headteachers … how that will impact us, I think I'm a better teacher because I'm watching people teach and giving them pointers and things to improve on. (F5I2)

There was recognition of the support received from headteachers in allowing mentors to participate in activities that enhance their professional development. The mentors suggested that universities should collaborate more with headteachers to demonstrate the value and impact of university initiatives on classroom teaching.

6.4.7 Additional Insights

Finally, participants were asked: 'Is there anything you would like to add about reliability and consistency or inconsistency in judging teaching effectiveness from your perspective?'

In response, one teacher educator shared several final thoughts regarding consistency and reliability in judging teaching effectiveness: 'I don't think you're ever going to get a consistency across everything all the time. But I think that as long as it isn't miles off, I think that's a good thing' (F10I1). They found that sometimes this inconsistency could 'generate professional dialogue' about how judgements were made and help acknowledge different perspectives. They further stated:

> I think as long as there's some similarities and it's broadly along the right lines, it's a really good thing in terms of having differing opinions. And I think we've a very boring place [in] the world if we have the same ideas and same perspectives all the time. (F10I1)

One tutor provided further insight, mentioning the university's involvement in conducting training sessions, specifically on setting targets. They expressed a desire to catch up on the session, indicating the importance of continuous professional development for educators, even when scheduling conflicts arise. They stated there was a 'meeting yesterday on exactly this kind of thing, setting targets, and of course, I couldn't go to the meeting because I had an appointment with Mr Ofsted' (F7I1). There was an emphasis on the importance of consistency in the messages conveyed to educators going into schools. The tutor highlighted the benefit of having all participants receive the same information, suggesting that it helps them understand what to look out for during their interactions in schools.

One mentor teacher also provided a few additional comments. They referred the use of the teaching video, suggesting that people might be more critical or harsh in their feedback because there is no face-to-face interaction. It was suggested the lack of direct contact could potentially influence the nature of feedback, with remote communication encouraging less filtered or thoughtful responses compared to face-to-face contact. Furthermore, the mentor teacher reflected on how the student teacher might perceive or react to feedback that is delivered remotely and anonymously. They further highlighted the importance of granting more autonomy to teachers within the education system:

> there are so many sort of judgements made that are outside of the control of the teacher that the more autonomy that you could give to the teacher, it would be better for the mentors, it'd be better for the trainees in the classroom (F5I2)

They expressed concern about the extensive influence of government advisors and external perspectives on teaching practices, suggesting that increased autonomy would benefit teacher mentors and student teachers. Autonomy was viewed as essential for fostering a sense of ownership and authenticity in teaching.

6.5 Discussion

In exploring the nature of judgement-making processes regarding ITE students' teaching effectiveness, this case study has illuminated the inherent complexities of evaluating teaching quality as, evidenced by the findings from the video task, questionnaire, focus groups, and interviews. Our analysis has underscored critical considerations related to evaluators' roles and responsibilities, the intricacies of assessing student teachers during their preparation, and the influence of the multifaceted nature of consistency in the collaborative judgement-making process. These insights are instrumental in addressing the research questions posed in this project and developing informed recommendations.

6.5.1 Refining Judgement-Making Practices

The findings from the case study revealed a general congruity between the respondents with respect to their judgement of teaching effectiveness and their approaches. The data highlights the importance of fair, consistent, and evidence-based judgements as a shared value among teacher educators, tutors, and mentor teachers. The role of professional judgement in teacher evaluation and how it is leveraged to complement teaching standards was also emphasized across participant groups. There was an overall consensus on the importance of student teacher understanding of the evaluation process and the need to ensure fair judgements are made by evaluators. While there was a strong foundation of shared beliefs about the importance of effective teacher evaluation, nuances in group responses provide valuable insights into the complexities of the judgement-making process and the need to consider different perspectives and the value of multiple voices.

Interestingly, there was a degree of variability among the rating of the seven dimensions of teaching from the video task (see Section 6.3.2), but not a substantial difference in the final overall rating beyond associate tutors giving a higher overall rating (mean ratings were: teacher educators = 3.92; tutors = 3.43; mentor teachers = 3.00). Some interesting questions emerge when examining ways in which the groups of evaluators distributed scores. Teacher educators showed the greatest degree of consensus (see Table 6.4) followed by tutors (Table 6.5); consensus is demonstrated through a low range of scores and low standard deviations. Although only four mentor teachers participated, this group showed the highest degree of variability across responses to the video task, with mean scores ranging from 2.25 to 4.00 (R = 1.75). The most variation occurred in the dimension of 'instructional strategies'.

In providing justification for their ratings, the mentor teachers drew greatly on their own practices as educators, calling on tacit knowledge and lived classroom experience in different schools; this came through as a potential source of variability among the groups. Participants' responses reflected their role in terms of *mentoring*, which maintains a focus on practices such as supporting students through ongoing professional learning and helping new teachers to continually interrogate and refine their teaching practice. In this role, it is many times important to actually *reserve* judgement.

In the rationales, there was a real emphasis around omissions – what student teachers did not do – and ways lessons could be improved. In ITE, processes often dictate that mentors take

responsibility on behalf of the teacher education programme for making assessments of a student teacher's progress and attainment of professional standards (Lofthouse, 2018). Given the significance of the mentor teachers' responsibilities, it would be good to reconsider this dual role of judge and provider of student support. As Papay (2012) proposed, decoupling rating of performance from support and feedback for skill devolvement can reduce the likelihood of differences in scores attributed to the evaluators themselves and support the use of context-specific knowledge, resulting in more precise and actionable insights. Furthermore, Haigh and Ell (2014) suggested that including additional professionals, such as the headteacher or other teaching team members, could strengthen the judgement-making process, give space for productive discussions around consensus and dissensus of evaluations as well as clarity in expectations. As Haigh et al. (2013) noted, 'several informed perspectives seems critical to balanced judgement of readiness to teach' (p. 10). This suggestion to strengthen processes through multiple raters has been confirmed in studies by Chaplin et al. (2014) and Saltis et al. (2020).

Given some of the variability that was evidenced, it remains important to explore ways to achieve greater consistency in evaluating teaching effectiveness. Participants' responses on how they determined their ratings helped us understand more fully the judgement processes in evaluating teaching. Participants relied heavily on the available perceived cues to make their judgements (see Section 6.3.3), and they demonstrated similarity in the ways they understood teaching performance and in the use of professional judgement (i.e., a synthesis of knowledge, experience, and practical wisdom) as a rationale for decisions. They had a common starting point for making judgements about student teachers' practices, which was to consider the teaching that was observed in relation to the pupil learning outcomes based on professional teaching standards (see Table 6.11). This suggests a strong emphasis on student teachers' ability to meet the expectations and criteria of the teaching standards. The second most common rationale was to look for strengths first and then weigh these against identified weaknesses, reflecting on whether the positives are more important than the negatives. This suggests there is also a focus on student teachers' strengths and that these are considered important in making judgements about overall performance. A variety of other rationales were used by participants to make judgements about student teachers' practices, suggesting there is no single 'right' way to make these judgements. This finding is supported by prior research by Bell et al. (2018), which concluded that due to the complexity of the act of observing teaching and learning (i.e., it occurs in real time and is interactional between teachers, pupils, and the content), evaluators need to simplify information about the lesson being observed in ways that, while varied, lead to accurate evaluations (p. 242). There are many applied heuristics developed from professional experience that can help address the highly complex cognitive task of judgement-making.

The results from analysing judgement-making strategies and warrants indicated a small number of cases of teacher educators and mentor teachers (5.0%) stating that they needed more information than was provided in the video to make a judgement or that they could not explain a rating or were simply uncertain in their decision (see Section 6.3.3). Instrumentation and piloting of the video task used in this study to capture judgements and

policies was carefully conducted and included selecting dimensions of teaching which could reasonably be observed through perceptual information (cues) in a teaching video (see Section 2.7 and Table 2.2). However, participants did query some of the 'invisible' dimensions of teaching and made inferences about the cause and effect of observed practices (i.e., explanatory rationales). Some participants noted other sources of evidence that could help provide a more accurate judgement, such as lesson plans, contextual information about the student teacher's prior evaluations, background on the pupils' learning and individual needs, and being able to speak to the pupils and the student teacher.

Interestingly, within judgement-making rationales, pupil learning did arise as a classroom cue, however this was fairly infrequent (see Tables 6.7 and 6.8), noted once by a teacher educator and twice by associate tutors. Additionally, one mentor teacher noted their starting point for making a judgement was to consider the impact of the teaching on the children's learning along with the evidence in relation to the teaching standards. This finding suggests consideration of pupil learning and their interactions as a component in decision-making. This emphasis on multiple sources of evidence aligns with prior research regarding valid and reliable assessment of teaching (see Goldhaber et al., 2017; Hylton et al., 2022; Parkes & Powell, 2015; Sandoval et al., 2020; Tanguay, 2020). And as Boguslav and Cohen (2024) have stated, we must contemplate the trade-offs with measurement decisions, understand the affordances and constraints of these, and ideally shape a set of measures with distinct strengths for distinct purposes.

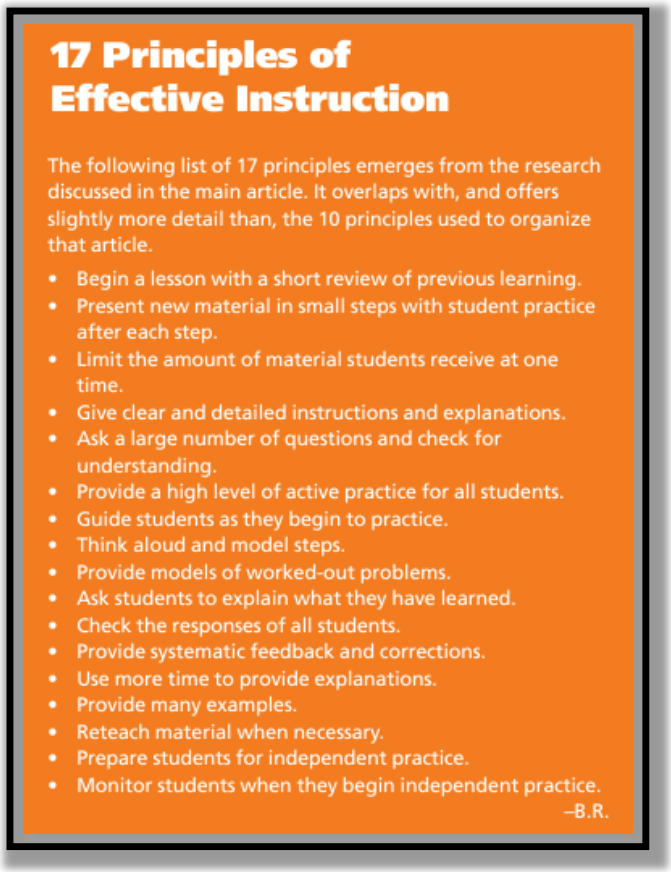
To incorporate multiple sources of evidence, the Teacher and Administrator Evaluation Framework, put forward by Linda Darling-Hammond (2013), a leading expert in the field of education from the Massachusetts Teachers Association, could be followed. This presents a triangulated approach to judgement-making comprised of observation of practice and artifacts, measurement of pupil learning outcomes, and consideration of evidence of professional contribution (p. 51). In this framework, the purpose of the multifaceted approach is to validate judgements about practice and the practitioner, a potential pathway to gaining the consistency desired without diminishing the importance of context. According to Darling-Hammond (2013), 'because student learning is the primary goal of teaching, it appears straightforward that it ought to be considered in determining a teacher's competence. Yet how to do so is not so simple' (p. 70). The consideration of how evidence of student learning can be used appropriately during initial teacher preparation to inform development thus arises. Teaching is, after all, a profession that is ultimately for the learners.

The findings further bring forward the need to consider better alignment between the form of evidence gathering (i.e., observation) and what may actually be observable from perceptible cues in a lesson observation. Findings add focus to deliberations of construct validity and the need to ensure the formats of judgement-making in ITE and the tools used speak to what they are intended to measure. Prior research confirms this need; an intensive exploration of assorted domains and dimensions for judgement and 11 authentic evaluation tools is provided in the systematic literature review in Chapter 3 (see Table 3.19).

Interestingly within this phase of the study, a teacher educator, when providing supporting rationales for judgements, and a mentor teacher in a focus group both brought forward Rosenshine's (2012) elements of instruction, as this had informed what they were looking for in the video observation task. Rosenshine's 17 principles, provided in Figure 6.1, were founded on research in cognitive science, master teachers, and cognitive supports (p. 12) and put forward as a 'valid and research-based understanding of the art of teaching' (p. 39). It is intriguing to see an explicit list of actionable descriptors outwith the professional teaching standards used as a resource for identifying cues by participants, in particular as the dimension of 'instructional strategies' was determined one of the easiest overall to judge (see Table 6.10). These principles and instructional suggestions utilized by participants, and other frameworks that likewise centre the core work of teachers' instructional practice (e.g., Australian Council for Educational Research, 2014; Marzano et al., 2011; Matsumoto-Royo & Ramírez-Montoya, 2021) could be explored as a potential way to focus the attention of judges on observable cues in teaching practice. As (Boguslav & Cohen, 2024) noted, a modified observation proforma with indicators focused squarely on the student teacher's practices could bring greater consistency in judgements of practice. Furthermore, what is included in an observation protocol can facilitate a shared understanding and common language in the judgement-making process.

Figure 6.1

Rosenshine's (2012) Principles of Effective Instruction



Note. From Rosenshine (2012, p. 19).

Findings from this case study emphasize the importance of considering the various factors that influence teaching effectiveness and the evaluation process. In participants' responses to considerations of influences on judgement-making (see Table 6.13), there appears to be a connection between the perceived value of multiple evaluators and sound processes and the importance of addressing evaluator error, suggesting that participants may view multiple evaluators as a means to mitigate bias and enhance the reliability of judgements. There was also indication from results of the questionnaire that those who valued consensus may be more likely to think that judgements should be consistent and objective, regardless of the specific situation, while those who value context-dependent judgements may be less concerned about complete agreement among evaluators, demonstrating a believe that judgements should be flexible and adaptable to different circumstances. This highlights the need to ensure clarity in what conclusions can be drawn and decisions made from the outcomes of observational judgements (Boguslav & Cohen, 2024).

It therefore may be of interest to explore an alternative approach to discourse aimed at consensus-seeking, as Moss and Schutz (2001) proposed in the format of hermeneutic conversation, an exploration of understanding where participants work together through a process of questioning, listening, and reflecting to uncover meanings and significance. The authors offered a structure that shifts from focusing on achieving agreement to understanding and learning from different perspectives (p. 58). It is in the exploration of dissensus through dialogue, they contended, that false assurances can be avoided and a fairer approach enacted. Moss and Schutz (2001) asked us to consider the imperative question: what level of agreement is reasonable to expect? This points to the value of the diversity of perspectives among stakeholders in the judgement-making process, recognizing that different viewpoints can contribute to a more comprehensive understanding of effective teaching. This was noted by participants in this study in a similar way to prior research. As Haigh et al. (2013) stated, 'the dissension between those charged with assessing readiness to teach is not necessarily negative and can be framed as potentially opening opportunities for professional growth if collaborative approaches to evaluation are taken' (p. 19). In fact, the emphasis on agreement, in data and across stakeholder voices, may actually reduce the way teaching is represented, prevent multiple perspectives, and even lead to an avoidance of matters related to values (Kornfeld et al., 2007).

The results of this case point towards confirmation that capturing observable skills of what student teachers do continues to be a challenge to implement in reliable ways, yet ensuring a fair and inclusive dialogue and exploration of dissensus may help to protect us from the false assurance of an articulated consensus that may misrepresent or exclude. This also highlights the need to ensure clarity in what conclusions can be drawn and decisions made from the outcomes of observational judgements (Boguslav & Cohen, 2024). Colón et al. (2024) acknowledged the tensions that exist between desiring agreement and high ratings and the irony that variability in the data lends itself to authentic and robust improvement efforts. These considerations underscore the lack of a definitive best practice for student teacher observation and the use of such judgements for consequential decisions in the profession. However, they suggested that we may have a clearer understanding of more effective

approaches, and more consistent evaluation of teaching effectiveness could be achieved through these methods.

6.5.2 Fairness

Emergent themes from the analysis of questionnaire responses revealed an emphasis on fairness in judgement-making processes – for the student teacher being evaluated, for the mentor teacher in the classroom, and for the individual(s) who is ultimately responsible for making and communicating the evaluation decision. This is a vital point considering the consequences of judgements for entry into the profession. Participants put forward several terms related to a sense of fairness (see Table 6.14), including fair, equity/equitable, parity, and equality. Participants' suggestions for strategies to gain consistency and reliability align with recognized principles of fairness in educational assessment (i.e., fairness in treatment during assessment, fairness as reducing bias, fairness as access to the construct being measured, and fairness as an opportunity to learn; American Educational Research Association et al., 2014). However, it was interesting to observe how these were used as seemingly interchangeable terms yet reflect conceptually different ideas that all contribute to understanding, and potentially influencing, consistency and reliability of judgements. Fairness implies treating everyone equally, without bias, favouritism, or prejudice; it is about applying the same rules and standards to everyone in a situation and as a product of moral judgement (Rasooli et al., 2023). Considering judgements of teaching effectiveness, we see the necessity for evaluators to understand and bracket potential bias, follow an established process, and apply the agreed professional standards for teaching. This would reflect a fair approach to judgement-making and a non-binary construct of what fairness looks like when enacted.

Equity, however, recognizes that student teachers potentially have different starting points and may require different levels of support to achieve equal outcomes. Equity is therefore about ensuring there is an opportunity to reach the desired outcome. During ITE, equity might mean providing additional resources or support to students who are struggling, which does require identifying and responding to individual needs and a close look at specific facets of instruction (Bastian et al., 2022) instead of a holistic overview. The key difference between fairness and equity lies in the approach to equality. Fairness focuses on sameness, while equity focuses on equal opportunity. It might also require a movement towards collecting multiple judgements over time to look for growth instead of a one-point-in-time judgement. This more equitable approach, however, does require more resources. A third and related term was also put forward by a teacher educator in this study: parity. Parity refers to the state of being equal or 'at par'. In the context of these terms of about fairness and equity, parity would mean that everyone has achieved an equal level of opportunity or outcome.

We see, therefore, that fairness is about equal *treatment*, equity is about equal *opportunity*, and parity is about an equal *outcome*. While fairness and equity are often discussed as foundational principles, achieving parity can be more challenging and, in fact, potentially unattainable in incredibly complex and dynamic circumstances. It requires addressing systemic inequalities and providing targeted support to ensure that everyone has the same

opportunities to succeed. The word that encapsulates all three concepts of fairness, equity, and parity is *justice.* Justice implies a state where all individuals are treated fairly, have equal opportunities, and ultimately achieve equal outcomes. Participants in this study clearly brought forward concepts reflecting the key point that future teachers are treated justly during their preparation evaluations. This resounds with social psychological conceptualizations of fairness, as Rasooli et al. (2023) put forward in terms of distributive, procedural, and interactional justice (p. 262) when studying teachers' concepts of fairness. Distributive justice centres on equitable allocation of outcomes, such as ratings of effectiveness; procedural justice considers the fairness of the processes used to determine these outcomes, considering factors such as accuracy, transparency, consistency, impartiality, correctability, participation, and reasonableness; and interactional justice takes into account fairness of interpersonal interactions, emphasizing respect, care, politeness, and the effective communication of information (Rasooli et al., 2023, p. 262). Ultimately, responses from participants in this case study reflect and reconfirm the challenge, and the necessity, of judging student teachers' practices in fair and reliable ways.

6.6 Conclusion

In this chapter, we synthesized the findings from a mixed methods case study carried out at LBU that employed video analysis, questionnaires, and focus groups to examine the complexities of teaching effectiveness judgements. Our analysis, grounded in a comprehensive theoretical framework, revealed the multifaceted nature of this process, as experienced by university-based teacher educators, link tutors and associate tutors, and school-based mentor teachers. The chapter provides insights into their judgement-making experiences and sharpens the consideration of dissensus, fairness, and dialogue in judgement-making processes. This chapter offers some important indicators and suggestions as to how we might develop observation and evaluation practices in school-based experiences in ways that are more equitable as well as useful to the various partners in the educational space they teach. In Chapter 7, we extended this investigation to include a third case study with a partner institution in Wales to further strengthen understandings about the nature of judging teaching effectiveness and the potential power dynamics among stakeholders that impact our collective understanding of professional competence.

7 Case Study 3: Aberystwyth University, Wales

This chapter presents context of the initial teacher education (ITE) programme at Aberystwyth University in Wales. It is one of three cases in the multi-case approach that comprises Phase 3 of this project. The chapter presents information about provision of teacher education at the participating institution with an explanation of school experiences and evaluation processes. Importantly, it sets out the present context and reasons why data collected is not included in this report.

7.1 Context

At the time of the project's inception, planning, and data collection, the Aberystwyth ITE Partnership was one of seven providers in Wales whose ITE provision was accredited by the Education Workforce Council, the accreditation and registration body for the education workforce in Wales. The Aberystwyth ITE Partnership's provision was accredited for a period of 5 years between September 2019 and August 2024. During this time, the Partnership was led by Aberystwyth University and six Lead Partner Schools and covered a large rural area in mid-Wales, encompassing Ceredigion, Powys, and North Pembrokeshire. Aberystwyth University acted as the awarding body for the Postgraduate Certificate in Education (PGCE) and qualified teacher status (QTS), and it exercised responsibility for partnership governance, course oversight, and delivery of the academic programme and held ultimate responsibility for quality assurance. Each of the Partnership's six Lead Partner Schools has been selected on the basis of a number of key quality benchmarks related to teaching, learning, professional learning, and leadership. Lead mentors from each of the Lead Partner Schools chaired or sat on all of the Partnership's key governance boards, including the Quality Assurance and Enhancement Committee, the Operational and Management Committee, and the Strategic Accountability Board, the Partnership's highest-level governance committee. Lead Partner Schools were responsible for coordinating student teachers' school-based experience and overseeing the input of a number of affiliated Partner Schools, which provided placements in their geographical areas.

7.2 Initial Teacher Education in Wales and Student Teacher Assessment

Following a series of reviews of ITE provision in Wales (Furlong, 2015; Furlong et al., 2006; Tabberer, 2013), in 2017 the Welsh Government published a set of accreditation criteria for ITE programmes in Wales. These were designed to form a framework within which provision would be delivered from September 2019, when the new round of accreditation commenced. Taking their cue from the second Furlong review (Furlong, 2015), these criteria outlined a key set of principles which accredited courses should embody and according to which partnerships between schools and universities should be conducted. These were:

- An increased role for schools;
- A clearer role for universities;
- Joint ownership of the ITE programme;
- Structured opportunities to link school and university learning;
- The centrality of research (Welsh Government, 2017a, p. 2).

The Aberystwyth ITE Partnership's PGCE course was designed to be an integrated programme, affording all students an 'all-though experience' by offering them opportunities to gain school experience in both secondary and primary settings, regardless of phase or subject specialism. This was achieved via the provision of two 'specialist' periods of school experience where student teachers would spend the majority of their course in their specialist phase (be it secondary or primary). In between these two specialist school experience periods was a 4-week 'enrichment' placement, during which secondary specialist student teachers would spend time in primary settings gaining insight and experience in primary pedagogies and assessment (and vice versa for primary specialist student teachers). Student teachers were therefore provided with three school experience placements during their 1-year PGCE course. The programme was highly innovative in this respect, and the rationale for its design was to provide an integrated experience (Thomas et al., 2020) which would enrich student teachers':

- understanding of progression at key points of transition;
- flexibility and range of pedagogical understanding; and
- employability as qualified teachers, especially so given the growth of the all-through school model in Wales (Harris et al., 2022).

Student teachers in Wales are assessed against the QTS statements of competence in the *Professional Standards for Teaching and Leadership* (PSTL; Welsh Government, 2019). The PSTL are made up of a series of descriptors, organized under five 'standards', based on key domains of practice:

- Pedagogy
- Collaboration
- Professional learning
- Innovation
- Leadership

Each 'standard', or domain of practice, includes graduated competence descriptors which are appropriate to the following career stages: QTS, induction; sustained highly effective practice; effective formal leadership; sustained highly effective formal leadership.

7.3 Aberystwyth's Initial Teacher Education Partnership: Practices and Processes for Judging Teaching Effectiveness

The professional standard descriptors for QTS were embedded in the PGCE course delivered by the Aberystwyth ITE Partnership. Scheme-level learning outcomes and assessment criteria for academic work were mapped directly to the QTS standards and, depending on the module, also aligned with Levels 6 and 7 in the *Credit and Qualifications Framework for Wales* (Welsh Government, 2021). The QTS standards were used throughout the academic year to assess student progression on a continuous basis while they were on school experience placement. Individual QTS descriptors from the PSTL were listed in the mentor observation booklet (commonly referred to within the Partnership as the 'blue book'). Mentors were provided with training on how to assess student teachers' progress against the criteria, using the PSTL QTS descriptors in the 'blue book' to note their strengths, areas for

development, targets, and progress since their last observation. The short duration of the Enrichment placement (4 weeks at the start of the programme, reduced to 3) meant that we couldn't guarantee a full visit for each student, though we did try to at least give each student a wellbeing check in person if possible.

Each student teacher's teaching effectiveness on practice was assessed as follows:

- via fortnightly formative reviews of progress while on placement, uploaded to a central portal by the mentor and visible to student teacher and link tutor;
- via three end of placement reports; and
- via one end of course final evaluation where all evidence was assessed holistically against the QTS PSTL.

7.3.1 Situation

ITE in Wales is accredited by the Education Workforce Council, which also monitors partnerships annually for compliance against the accreditation criteria (Welsh Government, 2017). In addition, Estyn, the education and training inspectorate for Wales, inspect all accredited ITE providers every 5 years.

The Education Workforce Council monitor and assess ITE Partnerships' compliance with the accreditation criteria (Welsh Government, 2017a), which outline the key provisions and guarantors of quality that Partnerships must have in place in order to be granted and maintain accreditation. The criteria themselves are extensive, and the 2017 version, under which programmes ran between 2019 and 2024, includes detailed provisions relating to schools, higher education institutions (HEIs), and partnership provisions, including:

- the selection of schools
- the need to develop a 'whole school' approach to teacher education under the leadership of senior teachers
- school staffing and responsibilities for supporting student teachers' learning including mentoring and the provision of structured opportunities for students to reflect on their practice
- staff development opportunities
- school facilities
- schools' involvement in the joint management of the programme

In relation to HEIs …:

- required staffing levels, staff qualifications and requirements for staff to be 'research active'
- the responsibilities of HEIs for student teachers including the support they must provide to develop their skills in literacy, numeracy, digital competence and the Welsh language to ensure that they are well prepared for the teaching context that they are entering
- staff development opportunities
- HEI facilities and student welfare

[For programmes:]

- the course's conceptual framework
- course aims
- course design and areas of study
- entry requirements and selection procedures
- core studies
- professional and pedagogical studies
- subject studies
- well being
- school experience
- the Equality Act 2010
- the assessment of student teachers. (Welsh Government, 2017a, pp. 3–4)

Estyn's role is to assess the quality of programmes' provision. Estyn undertake this role in line with their revised inspection framework for ITE in Wales, which is outlined in *Guidance for Inspectors: What we inspect. Initial Teacher Education (ITE) from September 2022* (Estyn, 2022), and further elaborated on in *Guidance for Inspectors: How we inspect. Initial Teacher Education (ITE) From October 202*3 (Estyn, 2023). The *What We Inspect* document (Estyn, 2022) outlines five broad inspection areas (IAs) and specific components of these, which Estyn examine to reach their judgements:

IA1 Learning

- 1.1 Standards and progress overall

IA2 Well-being and attitudes to learning

- 2.1 Well-being
- 2.2 Attitudes to learning

IA3 Teaching and learning experiences

- 3.1 The breadth, balance and appropriateness of the curriculum
- 3.2 Quality of teaching and mentoring

IA4 Care, support and guidance

- 4.1 Personal and professional development, and the provision of learning support
- 4.2 Safeguarding

IA5 Leadership and management

- 5.1 Quality and effectiveness of leaders and managers
- 5.2 Self-evaluation processes and improvement planning
- 5.3 Professional learning (Estyn, 2022, p. 1)

The Aberystwyth ITE Partnership was inspected by Estyn from January 2023 to June 2023, and an inspection report was published on 29 September 2023, which was critical of a number of aspects of the Partnership's provision, including:

- student teacher progression against the standards;
- quality and consistency of mentoring across the Partnership;
- communication across the Partnership; and
- coherence between the University and school-based aspects of the Partnership's programme.

All ITE partnerships in Wales which were accredited in 2019 for 5 years were required to apply for re-accreditation for their programmes by 8 January 2024. Aberystwyth ITE Partnership was unsuccessful in its bid for re-accreditation by the Education Workforce Council and so has withdrawn its PGCE course from September 2024.

7.3.2 Data Collected

We currently hold data from questionnaires and focus groups collected in February–March 2024.

7.4 Conclusion

Given the situation outlined above, the PI made the decision jointly with colleagues at Aberystwyth and those formerly of Aberystwyth and now at Swansea to not include the data collected from six mentor teachers and one university staff member. Following conclusion of the project and when more is known regarding employment, the team does intend to follow up to analyse the data in the context of it being collected in the midst of the re-accreditation decision. We find this to be an even more imperative time to bring forward the voice of mentor teachers in a study examining collaborative processes in teacher education and power dynamics.

8 Delphi Panel

In Phase 4 of the research project, a Delphi panel was convened to take up convergent findings across Phases 1–3 in rounds of discussion and consensus building. The goal of the panel was to generate a reliable opinion on the topic of judgement from a group of nine educational experts through an iterative process of questions and feedback. The Delphi technique is a communication structure aimed at producing critical examination and discussion (Green, 2014); it was appropriate for this project as it was developed to engage expert opinions on issues for which the is no clear answer and potential for dissensus. Presented in this chapter are details regarding preparation, recruitment, procedures, analysis, and findings of the Delphi panel. The results of the panel as part of the larger project serve to broaden knowledge regarding judgement-making on teaching effectiveness.

8.1 Methods

The primary purpose of the Delphi technique is to generate a reliable consensus opinion of a group of experts through an iterative process of questionnaire interspersed with controlled feedback (Beiderbeck et al., 2021). Originally developed by the RAND Corporation for technological assessment and forecasting future trends (Brown, 1968), this technique was named after the oracle at Delphi, as described in Homer's *The Odyssey* (Hasson et al., 2000). The Delphi process is intended to assist in clarifying central strategic questions at stake in a given practice where the outcomes may, at first impression, appear to be uncertain. The technique also serves to build collective understandings of research with participants (Cohen, 2018, p. 434) and has real value when practices are considered undetermined and contested (Baumfield et al., 2012. The method recognizes that a feature of expertise can be its diversity of perspectives (Hassan et al., 2000). We followed steps of the Delphi process as outlined by Stewart and Shamdasani (1980):

1. Develop the initial Delphi probe or question
2. Select the expert panel
3. Distribute the first-round questionnaire
4. Collect and analyse Round 1 responses
5. Provide feedback from Round 1 responses, formulate the second questionnaire based on Round 1 responses and distribute
6. Repeat Steps 4 and 5 to form the questionnaire for Round 3
7. Analyse final results
8. Distribute results to panellists

Beiderbeck et al. (2021) noted the Delphi technique has been used frequently in various disciplines, including education, which is evidenced in an increasing number of studies involving educational professionals (Baumfield et al., 2013; Borremans & Split, 2023; Oxley et al., 2024). The technique involves conducting a series of conversations that have, at their conclusion, a distilled account of important insights on consensus and dissensus evident in the 'idiosyncratic approach' (Haigh & Ell, 2014, p. 19) university and classroom-based mentor teachers use to reach decisions about teaching effectiveness.

8.1.1 Recruitment

A particular appeal of the Delphi method for this project was that it was developed to generate insights and lead to convergence (or divergence) of opinions while also recognizing the heterogeneity of expertise. This honoured the role of professional judgement inherent across the project and leveraged collective tacit knowledge of the group. We followed Beiderbeck et al.'s (2021) guidance that a more condensed set of experts was appropriate for a specialized topic and that five to eight experts would be a sufficient panel size (p. 7). We aimed to recruit 10 panel members (to account for potential withdrawal) with different roles within teacher education to obtain a more comprehensive view of judgement-making. The identification strategy was based on the individual's areas of expertise and their familiarity with teacher education systems and practices in the UK, and it sought to include teacher educators and researchers from beyond the UK and stakeholders in schools. We were also careful to consider the interest level of potential participants, as we know time and attention of experts is highly valued and a personal interest can increase the overall quality of engagement (p. 11).

Experts were nominated by the research team members based on the above criteria and their perceived expertise to contribute on the topic. A list of 20 potential panellists was generated. The Principal Investigator (PI) and Co-Investigator (Co-I) contacted individuals via email; the recruitment message and brief are included in Appendix A8.1. Three individuals took up the invitation for a dialogue prior to deciding to participate; the PI held these conversations and all three consented to participate. Once 10 individuals had agreed, recruitment ended. Unfortunately, one member, an executive dean of an Institute of Education, had to withdraw at the last minute due to unforeseen circumstances, and thus the final panel included nine experts. Collectively the panellists expressed that participation in the Delphi technique was a unique and rewarding opportunity for meaningful discussion with colleagues.

8.1.2 The Panel Participants

The convened panel included experts from the teacher education community who have been active figures in teacher education and development. Expert participants were drawn from within the UK (three from Scotland and one from England) and beyond (Australia, Norway, Switzerland, the US); eight of the nine participants had direct primary or secondary school teaching experience. Anonymity in the Delphi process is important both to limit bias and to allow for freedom to engage openly and express opinions and criticisms. This Delphi technique reflects 'quasi-anonymity', as participants were known to the researchers and to one another; however, their responses and opinions remain strictly anonymous (McKenna, 1994). The panel included the following experts:

1. Professor and faculty of education dean
2. Professor and former university dean of education
3. Executive leader of local authority
4. Head of secondary school
5. Primary teacher
6. Professor and leader of a teacher education professional organization

7. University director of school partnerships
8. Professor in education
9. Professor and leader in European education research

Due to the intensity of a 1-day, real-time panel discussion, we arranged for the panel members to meet one another in a social setting the evening before. Six of the participants were in attendance. This allowed for introductions to be made and a degree of familiarity and comfort among the panel to be established prior to the first round of discussion.

8.1.3 Procedures

Preparation is essential to ensuring the validity and accuracy of a Delphi study (Schmalz et al., 2021). The preparation stage involved defining a clear goal for the panel, agreeing the Delphi format, defining statements, and selecting questions to propose to the experts. We began preparations with a review of convergent findings emerging from the systematic literature review (see Chapter 3), review of professional teaching standards (see Chapter 4), and the case studies carried out in Scotland, England, and Wales (see Chapters 5–7). We refined the goal of the panel as gaining a practical contribution to decisions about judging teaching effectiveness based on the research questions. Statements and questions in the initial brief (see Appendix A8.2) were tested and refined by the PI, Co-I, and Research Associate (RA).

We selected a real-time format of three rounds of discussion, two involving the expert panel in discussion together with a facilitator and a final plenary session with the members of the research team (see Appendix A8.3). Two members of the research team were chosen to facilitate: Professor Jim Conroy, who is well versed in data collection and knowledgeable about the Delphi protocol, having utilized it in prior research (Baumfield et al., 2013); and Professor Rachel Lofthouse, who was involved in the project since its inception and guided design decisions yet has a level of neutrality having not engaged in data collection or analysis in the prior phases. The PI, RA, and project team member from Leeds Beckett University observed and took copious notes but did not engage in facilitation. Through careful distillation during the first round, they brought forward questions for the facilitators for Round 2 and repeated this for the plenary, with a consensus summary drafted. They did contribute in a limited manner in the plenary.

8.1.4 Analysis

To facilitate mapping key themes in the professional discourse and synopsis of the first round, the sessions were audio-recorded. After the panel, the audio recording was transcribed. The transcripts were analysed using the constant comparative method (Glaser & Strauss, 1967) to construct inductive codes, categories, subcategories, or themes. Guidelines of thematic analysis were used to ensure reliability, and data were explored through the six steps of qualitative thematic analysis (Braun & Clarke, 2006): familiarization; initial coding; generating themes; validating themes; defining themes; and interpreting and reporting. The analysis was conducted by an initial evaluator to determine emerging patterns of core ideas.

An independent audit was conducted by another member of the research team to determine a consensus of findings.

8.1.5 Credible Interpretations

To ensure credible interpretations of the findings are produced, we applied criteria for qualitative studies from Lincoln and Guba (1985). The criteria considered credibility (truthfulness), fittingness (applicability), auditability (consistency), and confirmability. The technique is based on the assumption that several people, in particular experts, are less likely to arrive at a wrong decision than a single person. Decisions of the group are then strengthened by reasoned discussions and consideration of perspectives, some of which may challenge assumptions, thus helping to enhance validity. The heterogeneity of the participants was preserved to assure validity of the results. Findings from the panel are considered reliable, as they reflect distilled expert knowledge from individuals who work across a range of professional interests and have direct involvement in teacher education; the inclusion of participants who have knowledge and an interest in the topic may help to increase the content validity in a Delphi study (Goodman, 1987). Additionally, the use of successive rounds helps to increase concurrent validity. Once prepared, the summary consensus statement was sent to participants in a member checking exercise to assure accuracy and resonance with individuals' experiences. Of course, there is always a possibility of some bias and the consequent emergence of groupthink. This was somewhat mitigated by three key features of the exercise: first, the geopolitical and occupational diversity of the participants; second, the deployment of the key findings from the wider project in the stimulus questions; and, third, the use of joint chairs who are embedded in different intellectual traditions, legislative contexts, and sociocultural practices in education.

8.1.6 Ethics

Ethical approval was granted as part of the full research project (see Section 2.6). Delphi panellists were provided with a participant information sheet and gave written informed consent to participate. The decision to record the sessions was a departure from typical Delphi processes. However, the transcripts enabled analysis of the discussion in a more detailed way that allowed us to answer the research questions. The recordings and transcripts were stored in the University of Glasgow's protected OneDrive system and were only accessible by the PI and the RA. As noted earlier, this Delphi technique reflects 'quasi-anonymity' since the participants were known to the researchers and to one another, but their contributions remain anonymous (McKenna, 1994).

8.2 The Iterative Delphi Process

In this section, we consider the nature and structure of the professional conversation that has fidelity to both the method and import of the Delphi philosophy. To this end the following sections reflect the gradual refinement of the nature and challenge of the competencies approach to judgement. As observed above, in this process, what begins as a set of questions shaped by the literature review and the analysis of empirical evidence concerning the exercise of judgement gets shaped by the respondents into questions about the origins, import, and

nature of the competencies – as the argument goes, only by garnering a sufficient understanding of *das Ding an sich* (the thing-in-itself) can one grasp its instruments of assessment.

8.2.1 Delphi Panel: Round 1

Prior to meeting, the panel members were invited to offer preliminary reflections, based on a series of questions (Table 8.1), on the current state of professional judgement with respect to student/early career teachers' competence relating to their practicum/classroom practice. As noted above, the questions were carefully drawn in response to the findings of Phases 1–3 of the project. These questions were sent to the expert panel members 10 days prior to the in-person Round 1 session (see Appendix A8.2 for the full brief).

Table 8.1

Questions Sent to the Expert Panel Members Prior to Round 1

1a.	In your judgement, what are the advantages and disadvantages of having a wide range of providers drawing upon varied and various schemas for assessing effectiveness of beginning teachers?
1b.	Do you consider that university teacher educators, associate/link tutors, and school-based teacher educators (i.e., mentor teachers) draw upon the same criteria when making judgements? Please explain your response.
1c.	Please explain your response.
1d.	In what ways, if any, does it matter in theory and in practice if there is disagreement in observations of teacher effectiveness?
1e.	How important is it that we encourage consistency? (Please explain your response.)
1f.	How important is it that we allow for breadth of opinion? (Please explain your response.)
1g.	How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?
2a.	What kinds /sources of evidence do you consider to be most important in coming to a judgement of teacher effectiveness? Note up to four.
2b.	To what extent do you consider judgements based on tacit knowledge to be important in assessing student teacher quality?
2c.	What do you consider the relationship between this tacit knowledge and centrally determined competencies is?
2d.	What do you consider the relationship between tacit knowledge and such centrally determined competencies should be?
2e.	How might the roles of university-based and school-based teacher educators in judging teaching effectiveness in initial teacher education be shaped by power dynamics?
3a.	What role is AI likely to play in teacher assessment?
3b.	What role should AI play in assessing early career teachers?
3c.	What might be the advantages and disadvantages of using AI in making judgements of early career professionals?

The individual responses were returned via email to the PI, who anonymized and consolidated the responses and provided these to the team member facilitating the panel; the latter then created a synopsis of key points, as provided in Table 8.2 (see Appendix A8.4 for a full summary). At the beginning of Round 1, panellists were presented with the synopsis along with the second set of questions, provided in Table 8.3, and they were invited to respond during the session in light of the new information.

Table 8.2

Example Synopsis: Anonymized Expert Panel's Responses to Questions 1a and 1b

1a. In your judgement, what are the advantages and disadvantages of having a wide range of providers drawing upon varied and various schemas for assessing effectiveness of beginning teachers?

Advantages:

- breadth avoids parochialism
- allows for comparison
- allows for breadth of interpretation about the import of certain teaching/educational outcomes
- broader levels of discernment
- heterogeneity offers some reflection of the complexity of the demands
- challenges consensus

Disadvantages:

- fit for particular circumstances/not the profession
- too frequently default to personal preferences
- too loose and the teacher student struggles to understand the expectations
- unreliability of judgement
- too many opinions to offer much discernment and with too little experience
- third parties are an unnecessary burden on the system
- waste/inefficiencies/redundant competition

1b. Do you consider that university teacher educators, associate/link tutors, and school-based teacher educators (i.e., mentor teachers) draw upon the same criteria when making judgements? Please explain your response.

- it requires maintenance, servicing, vigilance iterativity and collaboration
- the exigencies of the local/pressing determine evaluation
- college based – more general
- even where there are generic frameworks, local practice/interpretation differ much
- assessment tools have produced greater consistency
- different stakeholders have diff[erent] emphases – culture/exam/differentiation (more political)
- people bring their diff[erent] experiences so make diff[different] judgements
- (collaborations produce more consistency)
- diff experiences bring different judgements
- university – ideal; school – practical

Table 8.3

Questions Sent to the Expert Panel Members at the Beginning of Round 1

1a–g:

- Many of you argued in favour of individual judgement guided by adherence to agreed competence frameworks. How specific should we be? Qualitative judgements may be too positionally freighted to be of much good!
- What, if anything, is actually lost in imposing instruments/practices of consistency?
- There seemed to be some ambivalence about the role of tacit knowledge – as a local phenomenon worthy but too opinion laden. Why should we favour different experiences over consistency?
- Where are the limits of our flexibility? How are we to determine them?
- Can we arrive at a definitive 'command' list of competencies?
- Would teacher education be improved if we were able to draw on an OECD-mandated (or similar) competence framework that was internationally consistent?

2a–e:

- Should we pay more attention to the personality of the teacher candidate?
- Is tacit knowledge necessary in making judgements about early career teachers' progress/achievement?
- Can we justify judgements made on the basis of tacit knowledge?

3a–c:

- Generally negative comments on the use of AI given that judgement is a human activity – to what extent might this be denial?
- What happens if/when a voice-activated AI video monitoring system can feed an AI 'brain' and make consistent judgements that, in all respects, mirror precisely human form?
- In what ways should/might we vouchsafe the distinctly human lexicon of life?

Note. Questions in italics are the queries taken up by the panel members.

8.2.2 Consensus From Round 1

In the first round of discussion, expert panellists emphasized the importance of delineating the specific skills, qualities, and knowledge required for effective teaching. While acknowledging the interrelatedness of competencies and dispositions, they differentiated between the former's amenability to shared assessment and the latter's more subjective nature. Despite the challenges in assessing dispositions, panellists concurred on their significance in evaluating a teacher's potential. They cautioned against overly granular criteria, arguing that a more holistic professional judgement is essential. To enhance teacher development, panellists advocated for a stronger emphasis on school-based mentorship and training. Recognizing the contextual nuances of teacher evaluation, they stressed the importance of tailoring judgements to specific student teacher settings.

8.2.3 Delphi Panel: Round 2

The research team utilized the Round 1 responses to formulate the second questionnaire, utilized in the Round 2 panel discussion (see Table 8.4). These questions resulted in further deliberations, out of which a number of broad agreements emerged.

Table 8.4

Questions Presented at the Beginning of Round 2

- If we acknowledge there is a set of competencies that is situated in context, how actually are judgements being made about new teachers?
- How do we know new teachers are effectively prepared to be teachers?
- Who has a role and responsibility in how we gather, capture, and talk about evidence of effective teaching?
- How do the interrelated roles of those working in teacher education (e.g., mentors, those with gatekeeping responsibilities, student teachers themselves) play out together in relation to how we make judgements?
- How are judgements actually being made about new teachers and their capacity to enter the profession?
- To what extent do you think these judgements are valid, reliable, consistent and/or inconsistent?
- What comes together to form a qualitative and quantitative account of how good that teacher is?
- What is the purpose of the judgements that we are making, and to what extent does that purpose influence the way we judge?
- Do those new to the role of judgement making during teacher preparation have the same insight into making judgements?
- Do we want to say anything else about the nature of evidence used to make these judgements?
- Can we spend some time thinking about the mentor teacher? How important are mentor teachers in the ability to make decent judgements about student teachers?
- Are there any other models of preparation that we could put on the table?

8.2.4 Consensus From Round 2

In the second round of discussion, panellists highlighted the multifaceted nature of teacher effectiveness judgements, acknowledging the diverse roles and perspectives of those involved. They emphasized the importance of theoretical frameworks in guiding these judgements. While school-based experiences provide valuable insights, panellists cautioned against using them as sole determinants of professional entry. Instead, they advocated for a cyclical approach that supports ongoing learning and development. The process of judgement was described as a collaborative endeavour involving student teachers, mentor teachers, and teacher educators.

Panellists called for a reimagined teacher preparation system that aligns with the complex and often intangible competencies required for effective teaching. They stressed the need for extended mentorship experiences within collaborative teams to foster the human aspects of teaching. This approach shifts away from traditional individualistic models and emphasizes the collection of diverse evidence over time to assess progress. Recognizing the variability inherent in teaching, panellists advocated for a more nuanced understanding of the less observable yet critical aspects of the profession. They emphasized the importance of maintaining evaluative distance between mentors and evaluators to ensure objectivity.

Concerns were raised about the potential for datafication to oversimplify complex judgements. While uniformity in outcomes may not be desirable, panellists stressed the need for equitable procedures and processes. They emphasized the ongoing nature of teacher evaluation, advocating for a cyclical approach that incorporates feedback and monitoring of progress. Differences in understanding the purposes of judgements were identified as a factor influencing consistency. Panellists clarified that consistency does not equate to sameness, recognizing the diverse ways in which teachers can demonstrate their competencies.

8.2.5 Delphi Panel: Round 3

This final stage of the Delphi process focused on refining consensus and ensuring that the outcomes were relevant, practical, and representative of the group's collective wisdom. The PI offered an overview of the project and explained how the Delphi symposium fitted into that landscape, together with a summary of the results from the previous Delphi rounds, including the convergence of opinions and any areas of disagreement (Table 8.5). Participants were encouraged to share their thoughts, ask questions, and provide feedback on the findings. This open dialogue helped to identify any overlooked issues or areas that required further clarification. The research team facilitated a discussion on the key recommendations and conclusions.

Table 8.5

Summary of Key Points and Questions Presented at the Beginning of Round 3

Key points from prior rounds:

- Judgements about teaching effectiveness should be grounded in theoretical frameworks.
- Judgements during school experiences should contribute to the ongoing development of the new teacher's discernment.
- Judgements involve collaboration between the student teacher, mentor teacher, and university teacher educator.
- Judgements are influenced by the different roles and experiences of those involved, but combining these perspectives can lead to valid and fair assessments.
- The relationship between university tutors and mentors is crucial for effective judgements.
- The most important aspects of teaching, often difficult to observe, should be prioritized in teacher preparation.
- Those making judgements should maintain distance from the student teacher's mentor to ensure objectivity.
- Overreliance on data can lead to reductionist judgements.
- While uniformity of outcomes may not be necessary, equitable judgement-making processes are essential.

Questions asked during the plenary:

1. What comments, questions, or follow-up to the discussions do you have?
2. Why does all of this matter?
3. What do we make these judgements about readiness for? What is the purpose of the judgement?

4. What areas about teacher preparation and our judgement-making processes do we need to rethink?
5. Do you foresee other ways of making judgements in a different way?

8.2.6 Consensus From Round 3

Expert panellists emphasized the need for a more nuanced understanding of teacher effectiveness, acknowledging the inherent variability and complexity of the profession. They called for a redesign of teacher preparation programmes to align with the actualized purpose of teaching and the less tangible yet highly desired competencies. The panellists advocated for an ongoing cycle of feedback and progress evaluation, emphasizing the insufficiency of isolated observations. They identified differences in understanding the purposes of judgements as a factor influencing consistency.

To make accurate judgements, panellists stressed the importance of multiple sources of evidence evaluated over an extended period to demonstrate progression. This necessitates enhanced mentor development and support. They highlighted the urgent need for a re-evaluation of the mentor teacher's role, accompanied by a reconsideration of mentor cultivation, education, and training.

8.3 Emerging Themes

Based on the discussions and consensus reached, the research team finalized conclusions and recommendations from the Delphi panel. The review of recordings, analysis of the transcript, and researcher memos from the panel sessions and the final plenary revealed significant areas of consensus and foregrounded several emerging themes.

8.3.1 Types of Competencies

First, the panel discussion readily resolved into a consensus that current methods of judging student teachers' readiness to teach (i.e., assessment) during preparation are too focused on easily observable behaviours and insufficiently focused on the more complex, and expressly essential, aspects of teaching. In the delineation of knowledge, skills (performances), and dispositions of effective teaching and the relationship among them, emphasis was placed on the need to shift from declarative and procedural knowledge to critical dispositions necessary for effective practice – the habits of professional action and moral commitments that underlie how teachers act in practice. The group discussed the importance of finding a way to assess these dispositional competencies within the complexity of a teacher's practice.

The panel considered that rather than thinking of micro-competencies as the unit of assessment, the altogether more synoptic notion of the 'repertoire' might better capture what is professionally required. This notion of the repertoire offers a more dynamic account of teacher competence, one which can be considered not at a moment in time but 'over time'. The task of early career education was to afford teachers an opportunity, or opportunities, to acquire a range of capacities and abilities, to build up a repertoire of capabilities and insights or, as a colleague put it, 'a reservoir' of professional aptitudes. None of the experts considered that early career teachers should be evaluated as if they were a seasoned

professional rather than someone embarking on a professional career, yet the competencies were often deployed as if they were the sum of professionalism. The judgements of teacher educators should be seen in this light – as a moment on a journey rather than as a destination.

Our judgement as to competence should be considered not only as developmental but also as broad based and focused on teaching and learning. A number of colleagues suggest that a critical component in any judgement about the competence of a student teacher (or indeed, any teacher) was having a sense of their impact on their pupils: Are they engaged? Are they understanding the subject matter? and so forth. One participant, with substantial experience of inspecting schools and evaluating teachers, recounted an evaluation where the teacher removed six students from the classroom for misbehaving. He went on to comment that his judgements were based on broad concepts about effective teaching and learning (for him and a number of others); perhaps *the* central component of a judgement was the application of 'the broad concepts about what effective learning and teaching was'. In a mathematics lesson he observed the principal teacher remove six kids from the class because of their behaviour and went on to comment that he asked the teacher, 'did you think it was a good idea to throw them out of the class when I was actually in with you?' After all, he continued:

> I always imagined teachers this is their Sunday best. So what are they doing when I'm not there, but it's based on a broad, broad concepts of what effective language and I think it's the same for students. Your judgement is made on these broad concepts that are largely shared.

There was much discussion about and agreement on what one some participants referred to as the mysterious 'it' factor that makes a good teacher. This factor, they considered, was difficult to define or assess, but they believed it to be essential for teaching success. The group discussed whether the 'it' factor can be developed or if it is something that teachers are born with. It was recognized as related to the character and disposition of the teacher and that while personality might indeed shape the form and shape of a given teacher's discursive, behavioural, and interrelational practices, there were myriad ways in which these character and dispositional traits might successfully work themselves out in the everyday life of the early career teacher. This notion of seeing competencies as flexible and broad-based descriptors was a recurring theme in the conversation and one not amenable to a simplistic technocratic solution.

8.3.2 Desired but Unrealized Collaboration

Expert panellists identified the underdevelopment of the mentor role and the need for increased collaboration among teacher education programmes, schools, and mentors. They expressed concerns about the limited attention paid to the professional development of university teacher educators, who often lacked clear justification for their authority as 'judges'. While acknowledging the importance of a truly tripartite conversation, panellists recognized the significant challenges and costs in establishing such a collaboration.

Resource constraints and dispositional obstacles were identified as key impediments. Panellists proposed a shift towards a more social and collective approach to teacher

education, drawing inspiration from the Northern European tradition, where a shared curriculum fosters a greater sense of agreement. They argued that a collaborative approach could enhance teaching effectiveness through shared language, common understanding of assessment, and a tripartite partnership. The panel emphasized the need for both standardization and individuality, acknowledging the importance of consistency while considering context and individual teaching styles. They suggested that pupil learning should be a key indicator of teaching effectiveness, alongside other factors.

8.3.3 Generated Consensus

The consensus opinion generated by the expert panellists emphasized the importance of both competencies and dispositions in assessing teacher effectiveness. They highlighted the need to explore the tacit knowledge of teacher educators to identify the 'it' factor that contributes to exceptional teaching. Current assessment methods were critiqued for their focus on observable behaviours and neglect of the more complex aspects of teaching. Despite challenges in judging teaching effectiveness, panellists advocated for maintaining a focus on expected competencies, particularly in the context of teacher shortages. They acknowledged the potential limitations of current judgement processes in identifying the best teacher candidates. A more collaborative approach to assessing new teachers' readiness was proposed to enhance consistency and reliability.

Panellists emphasized the importance of shared language and a common understanding of assessment for effective collaboration. They cautioned against making observations performative and advocated for a holistic approach that avoids laundry lists of competencies. The panel stressed the need to avoid homogenization of the teaching profession, recognizing that teacher preparation should not result in a standardized product. Desirable qualitative and functional competencies were acknowledged as requiring extended time in schools. The panellists highlighted the importance of defining the 'it' factor to legitimize the teaching profession. They discussed the need for flexibility and the acceptable degree of deviation from the ideal, particularly in areas with teacher shortages. The development of a personal language for working with pupils was considered crucial, and discussions considered the desired levels of consistency and coherence.

8.4 Discussion of Delphi Panel Findings

This part of our report draws on the raw material of the conversations and reflections of our Delphi participants and attempts to shape them around a number of overlapping themes. Moreover, it attempts to connect these reflections to wider considerations in the research literature and connect with several of the themes that emerged from the empirical parts of this study. In focusing on this small number of themes, we are conscious that we have left much out, but we consider that we have captured the timbre of the conversation as well as the substantive agreement among participants. Equally, we are mindful of the limitations of the Delphi protocols as they are historically practised, wherein the impulse and imperative is to resolve the complexity into a singularity – into what is most important or urgent. But we believe that the themes adumbrated here cannot be so resolved. Rather, they are a set of

reciprocal relations where success depends on producing some harmony out of the sum of the parts.

8.4.1 Democratization: Stakeholder Involvement

One of the themes that emerged early in the discussion was the belief that everyone wants their say and that this tended towards the creation of what one participant described as the 'laundry list' of competencies. The conceit of the 'laundry list' has its roots in competing imperatives, one coming from an ethical and socially motivated position to facilitate 'full participation' by interested parties and the other from a more neoliberal impulse. With respect to the first of these, the 8th World Congress of Education International, held in Bangkok in July 2019, passed a resolution supporting the implementation of the joint Education International/United Nations Educational, Scientific and Cultural Organization (UNESCO) *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019); it was made crystal clear that teachers and their unions must be at the centre of the process, working with governments and other education stakeholders. This is in line with the spirit of the Education 2030 Framework for Action, which calls for the '*full participation* of teachers and their representative organizations in the development, implementation, monitoring and evaluation of education policy' (UNESCO, 2016 p. 4, emphasis ours).

The key principles of such democratic evaluation – inclusion, dialogue, and deliberation – are captured by House and Howe (1999, 2000), who argued that all of those with 'legitimate, relevant interests' (House & Howe, 2000, p. 5) in an evaluation should be included in decisions that affect those interests and that there should be a 'rough balance and equality of power' (p. 6) among the various parties. Applying these ideas to teacher education and integrating them with the idea of intelligent accountability means that accountability mechanisms, processes, and content – i.e., what teacher education is actually accountable for – are jointly determined through dialogue and deliberation among teacher educators centrally involved in teacher education programmes and institutions, school-based educators at local schools and communities that partner with teacher education institutions, and representatives from other relevant professional organizations, such as teacher unions. This means that the content of accountability cannot be completely predetermined, because it emerges from deliberations about local commitments and goals as well as larger professional and national values. This also means that school-based leaders and teachers as well as relevant community members function alongside university-based personnel as co-equal teacher educators and not simply as the co-occupants of the spaces used to prepare teachers. We acknowledge, as Faul & Savage (2023) also pointed out, that every stakeholder within an education system will have their own goals, and powerful incentives, interests, that shape what emerges. As the authors stated, 'If we deny this complexity, unintended consequences multiply. It is far better to acknowledge the complexity of the systems in front of us and work with them' (p. 2).

The second driver is altogether more complicated. The impulse to expand lists of competencies, which, interestingly, gets taken up repeatedly by our Delphi participants, does have an alternative explanatory strand to its genealogy. In a representative democracy,

licensed professionals are nominated to go about particular kinds of work on behalf of the people. And yet the 'laundry list' approach to stipulating competencies would, at the very least, call into question that compact between the people, the government, and the professionals. Of course, this is not peculiar to teacher education (Milbourne & Cushman, 2013); nonetheless, it has the effect, as a number of participants observed, of turning comprehensive judgements into a catalogue of micro-competencies that don't offer any especial advantages. This is summed up in the following contribution:

> just as an aside, we have a bit of a cog sci kind of mania in England and … we're now seeing … things broken down into the most micro pieces, which some people would argue [represents] the teacher's repertoire, and actually all they've managed to do is get from that side of the room to that side but they did it in 1,010 micro-stages, so even they could tick each one off.

This of course is of a piece with and an explanation of the evolution of the long list so avidly discussed in Round 2. Hence it has become rhetorically important not to delineate some particular competence as an overarching description (e.g., the teacher was able to ensure that all the children were engaged in the lesson …), but to explicitly attend to all the sub-competencies along the way that might contribute to the realization of any such competence.

The advent of the Education Reform Act 1988 for the first time mandated a curriculum and its processes. The particular Articles in the Act that will fuel the move away from professional autonomy appear modest but will have far-reaching effects. For the first time the Secretary of State for Education will have responsibility and authority to determine and mandate:

> (a) the knowledge, skills and understanding which pupils of different abilities and maturities are expected to have by the end of each key stage …
> (b) the matters, skills and processes which are required to be taught to pupils of different abilities and maturities during each key stage. (Article 2(1))

What follows from this is an increasingly centrally determined approach not only to the curriculum and pedagogical entailments but also to increasingly stipulative entailments of teacher practice and preparation, because teachers and teacher educators could not be relied upon to 'deliver' on them. And so, the laundry list is a result, and because there are no clear or easy means of discriminating between the myriad claims on teacher education, the members of the symposium thought, we include them all. General competencies get further detailed and increasingly specific and we lose sense of the synoptic (a general view of a whole, characterized by comprehensiveness or breadth of view). Subtraction as solution to the complexity is thus considered. The panel discussion brought forward consideration of holistic competencies, and to avoid a laundry list – a broad framework of competencies – we have been conflating a complex range of measures with minimum requirements. The laundry list can potentially homogenize the teaching workforce and squeeze out the very diversity we aim to support, attract, and retain.

As a counterpoint to the more constraining character of competencies in teacher education, it is worth attending to the application of these 'technologies' in other domains. In some other professional fields, such as medicine, even where the broad outcomes are specified by a range of stakeholders, the professional actor is generally left to deliver on them. An interesting example of this difference can be seen in the competence framework of Stanford University's student medics, which includes the expectation that they will '[a]dvocate for high quality, optimal, and safe patient care systems' (Stanford Medicine, 2024, 8.3). The point here and elsewhere in the Stanford example is that the language of professional action and obligation is framed in a much broader professionally empowering discourse of responsibility and control. The history of the competence framework is altogether more politically and culturally fraught, arising, as it does, from the more liminal world of the child as they move towards adulthood. Thus, even a medic who works as a paediatrician exercises their practice with respect to the child, whereas the teacher teaches a child. The relationship is altogether more direct. While everything may not be capable of easy codification, this matters because the experiences the students have, the failures to specify (and with a young population of teachers who don't know otherwise), become the intergenerational learning and they then normalize these forms of practice as if they were indisputable.

8.4.2 Translation – Let There Be No Gap

The panel discussion surfaced the ongoing debate around the perceived theory to practice gap in teacher education. Whereas some participants held that theory is important for providing a foundation for teaching practice; others believed that theory can be too abstract and be readily seen as irrelevant to the pressing exigencies of the classroom. It is certainly the case that many of the scholarly obsessions of educational theorists with respect to theories of power, whether they be postcolonial, Foucauldian, Bernsteinian, Bourdieuian, seem to have little traction in the classroom. Teachers tend to be more concerned with the pressing issues of classroom dynamics, keeping children on task, 'delivering' a prescribed curriculum and so forth. Interestingly, analysis of the discussion transcripts revealed a set of potential competencies that are actually required to address/bridge this perceived gap. We have considered the difference between early career teacher competencies and those adumbrated in the Stanford Medical Education programme where, among other concerns was an expectation that medical students not only actively keep abreast of medical education but also involve themselves in it as agents. There is an interesting question as to the extent that teachers need to be engaged in or abreast of current educational research and whether this has practical professional implications for them. Our Delphi respondents did appear to underwrite the need for teachers to have an active engagement in such matters on the basis that without it teachers risk losing a belief in what they are doing. So it is that the arcane discussions of educational theorists can have a profound effect on the discursive practices of teachers and these in turn shape the actual and material conditions of the classroom. One interesting example that may cast some light on this difficult and often obscured question of the import of the theoretical was foregrounded by a senior, internationally based academic in trying to elucidate and account for what she termed 'pedagogies of love'. This is particularly interesting because it underscores some of the discussion in Chapters 5 and 6 around the import of the relational. In

order to consider what such a pedagogy might entail and how a teacher educator/tutor/mentor might assess it, it is arguably necessary to have an 'active' account of the term. Teachers' practices are shaped by the theoretically grounded discursive practices of the field as a whole. One participant thought these matters vital precisely because institutional generational memory is determined by such conversations and the theoretical justifications on which they are built. Hence, she observed that

> the expectations that then [are] established around how they are going to be judged, how they feel about that, that that then becomes almost the intergenerational learning that they take with them … not because, not just because, we need to be able to send students, teachers out into the profession who we can trust to do a reasonable job given the right continuing support, but because it becomes the institutional memory, it feeds into institutional memories and that then completely alters the generational experiences that we're having.

The difficulty with all *institutional memory* is that it can be resistant to self-critique and become 'the way we do things around here'. In important respects, we saw this emerge in the findings in Chapter 5, where respondents justified their judgements, sometimes quite unconvincingly, on the basis of tacit knowledge. One of the most salient reasons for supporting a competence in theory and research is that it can act as a counterweight to faddism. As one of the respondents observed, it can stop teachers from falling for the latest pseudo-pedagogical nostrums from, as he called them, 'snake oil salesmen'.

Perhaps this leads to a yet more urgent reason for closing the gap between theory and practice given the rise of AI, which formed a significant part of the collective reflections: one member of the panel offered an example of an educational adviser's account of the wonders of AI in education. The adviser was recounting to our participant some really exciting work they saw being done in a school recently, where these young people were exploring the causes of the First World War, so they got AI to produce the first draft of their essay, and then they polished it up.

Our participant recorded his shock at the naivety and intellectual misconstruction of what it is to produce a piece of academic writing or indeed to engage in the epistemic act of knowing something about the nature of history and causality. As he put it:

> could you get a more egregious misconception of how arguments are constructed and how causality … [it] is just such a misunderstanding of learning and a misunderstanding of the very concepts at the heart of drafting something. But talking about AI … crystallized for me why you need teacher knowledge and why you need universities involved in the education of teachers … and if that's what AI is going to do, then, well, that's the end of the whole project.

In this he was pointing out that a pedagogic strategy of outsourcing responsibility for gathering evidence, putting together an argument, and evaluating (in other words, coming to know something) was not, in fact, epistemically or intellectually defensible. And, if this was to be the new modus operandi of education, then it was somehow an entirely different

enterprise. There was much agreement on this from around the table with parallel examples from completely different jurisdictions. As one UK participant observed: 'I'm very glad to see [xxx] saw the same point from the other side of the globe.'

8.4.3 A Shared Exercise

The question of the import of teachers being aware of research segues into the notion of a shared enterprise/partnership between school and university. The only classroom teacher in the group (an experienced mentor) was clear that there was insufficient mapping between the particular needs of the school and those of the student. Often teacher education students were placed in contexts or with people that were unlikely to secure their flourishing or indeed that of the pupils they would be teaching. One senior local authority leader recognized that when placing very large numbers of students, such mismatches would inevitably arise.

> And I know that some of our schools would be more supportive than others, but given we have 350 to place, you know, like they're placed where [we can put them]. And there's also that bit of how the school has to [have certain needs] to get a probationer and all the rest of it. So you could have somebody that if they go to the school [x] would flourish, but if they go to the school [y], you know they end up with Miss Trunchbull and it doesn't quite work out and we need to move them.

This question of the importance of context and culture emerges repeatedly in the course of this study. All the participants acknowledged that different contexts and sociodemographic conditions meant that any judgement as to a particular student's achievement were or should be shaped by these contingencies. And yet they were equally reluctant to abandon the application of a generally applied set of competence benchmarks. Ultimately, participants shared a belief that it was possible to apply the same competence benchmarks to all so long as the descriptors were not drawn too tightly and allowed for myriad forms of evidence to count. This echoes the reasoning of a number of those teacher educator/school mentor participants who were keen to consider context but, at the same time, wanted to make comparative judgements, which we try to capture in the duplexity model. They argue, yes, context matters, but so too does consistency and fairness; indeed the desire for consistency runs through the discussion from beginning to end. But consistency is not the application of a rule-based regime without regard to the particularities of context; it is rather the application of principles of judgement. Much of the second formal session of the Delphi seminar was absorbed by colleagues wrestling with consistency and finally resolving on something like fidelity to the observed experience. One participant noted that Ofsted claimed that 'as long as the assessment is not consistent [but] objective, that it is valid and it's reliable'. Delphi panellists generally resiled from such an application of the notion of objectivity given that it is merely a displacement term. Perhaps a more helpful notion than either consistency or objectivity was one subscribed to by participants – reliability. For them, reliability consisted not in having an immutable set of competencies that might be overlaid on any situation, but in having an asset of criteria that one would draw on to make a judgement.

8.4.4 Collaboration and Consistency

While university leaders in teacher education express a strong desire to collaborate and to envisage early career progression as a continuum and shared responsibility, this did not always translate into a seamless process. For all of the desire that participants had that the development of early career teachers be considered just such a shared endeavour, there was a sense of the limitations of this. One respondent (a senior university director of partnerships) lamented the failure of efforts to provide resource, intellectual, and practical continuity between university and school:

> I have found it quite sad because we spend the final week of our course preparing this documentation with the trainees to assure and to give confidence to schools that we feel that they are now ready [and] moving through. And they sit down and they set out the four targets which … they feel they're very, very strong on and you know this is something which they hope that can be developed and four areas which they still feel need to be developed as they move through. And our whole push is [that] as early career teachers … they take this into their very first teaching job. And, for three years we actually then went through [this process] … under the disguise of bringing as many of these trainees back in the autumn term with a speaker. And it was part of their own development as well too, because we felt that responsibility for them as newly qualified teachers. And, of course, we were using it as a guide as well too to find out how useful this documentation was for schools, and probably 5% of schools were using what we sent out. So a whole week of a course spent as part of this exit plan to get that out to schools was not being used at all.

This is particularly concerning given the recent House of Commons Education Committee Report (2024) to the UK Parliament, which among other issues highlighted a concern about duplication of courses/provision for early career teachers where material covered in initial teacher education was replicated during the probationer/early career phase. Alongside this was the observation that significant gaps were evident, most especially with respect to subject knowledge for secondary teachers. At the same time one director of education observed that with some 300 plus probationers entering his authority's employment on an annual basis, he neither sought nor expected guarantees that every new entrant would indeed be successful. He was, however, aware of the dropout rates during probation from each providing institution. Moreover, a lot of the reassurance he sought came from having a close and functioning relationship between schools, local government/trust education administrators, and teacher education providers. While of course this is complicated in some jurisdictions where there are a variety of pathways into teaching, the participants nonetheless considered that the guiding principles should be the same.

8.4.5 Relationality

The necessity of having a close functioning relationship leads to another node of agreement in the concluding section (as it was throughout the discussion) and one that surfaces repeatedly in this report – relationality. In the process of understanding how to judge a student teacher's performance, all the participants were keen and committed to placing the

notion of the competence in a broader frame of judgement that included dispositions, attitudes, and commitments that could not be easily captured by a competence descriptor. One participant (a senior international leader) suggested that it might be described as a commitment to a 'pedagogy of love', mentioned earlier. But such love is not, obviously, of the private (romantic or other) kind, but something altogether more public – more redolent of Hannah Arendt's deployment of Augustine's notion of *amor mundi*. Such love may be considered a duty to love the world as it presents itself to you. Assessing such a radical commitment in the early career teacher is, as the panel noted repeatedly, an extraordinarily complex and challenging, but necessary, task. Hence, another participant observed that such love is expressed in the desire that students would flourish and that in the committed professional domain, love and flourishing were intimately related terms. Moreover, the pedagogy of love was also considered by participants to be related to trust and trustworthiness. As she went on to observe:

> it's in the embodiment, it's related to trust and trustworthiness, which is something that you can have and you can develop, and it shows when you are a trustworthy teacher that the kids know that, okay, math, I can trust him … he's, he told me [off] or he scolded me but at the same time I know that he loves me. And that kind of trustworthiness is something that you build up in relations and … you have to know that you have to develop that as a teacher. There are a very, very few that are not able to do it … but I think this is some of the system that would put them through today.

In a response to this description, another participant observed that this impulse and imperative too often got lost in the contingencies and exigencies of the highly performative character of early teacher education. Despite this performative character of early career teacher development, participants collectively acknowledged that much of the expression of this 'love' was to be found not in the formal curriculum but in the 'in-between' spaces that lie at the heart of the encounter between students and teachers, between students and students (Conroy, 2004). One senior secondary school leader put it this way:

> I would absolutely agree … but I think if we are saying that education is this much richer thing – it's the art and the science, isn't it, of teaching; it's not something that you can see then – but then [subsequently] with the instruments we use [to] in some sense measure it, calibrate it, [we] have to recognize that … the risk that we face is that we are reductive in the instruments that we are applying.

8.4.6 An Intergenerational Conversation

A modest but significant and somewhat related theme that found much agreement between participants was interestingly framed by one senior academic who asked why this kind of conversation matters so much. The answer was perhaps a little more oblique than might have been expected, but for all that still interesting. We can confidently state that all the participants, from very different perspectives, shared a robust mistrust of the priority often afforded a reductively performative characterization of competencies. If, the argument went, the only thing successive generations of teachers were exposed to was some pseudo-technical account of the affordances of a teacher, then this would delimit the importance of a public

education as a humanizing (loving) endeavour. If we suppose that a generation of teachers were to be immersed only in a singular account of educational purpose and teacher effectiveness, shaped only by the language of performativity, then, it was suggested, that is likely to diminish the collective wisdom that supports the development of the repertoire considered above.

In an echo of Oakeshott's (1971) claim that education is a conversation between generations, one participant observed that the discussion mattered

> because the experiences that our student teachers have in those few years or months that they're with us and the relationship that they have with their mentor, the expectations that then [are] established around how they are going to be judged, how they feel about that, that that then becomes almost the intergenerational learning that they take with them.

This seemingly innocuous observation, widely agreed by the panel, harbours something rather more potent. It is that teacher education, in common with all education, is a process delimited by the discursive resources available to its participants, both teachers and students, and if the language of extrinsic motivation with its performative character prevails to the exclusion of other linguistic resources, then the issue must inevitably be delimiting – and that has significant consequences for all of us. As Oakeshott (1971) argued, education is a process where the senior generation introduces newcomers to a shared heritage of human knowledge and beliefs. Unlike passing down physical objects, which would be a simple task, this inheritance consists of intangible elements like activities, aspirations, and ways of thinking. These aspects can only be truly grasped through a process of learning and understanding, making education much more than a mere transfer of material things.

In important ways, Oakeshott (1971) captures the heart of the conversation between the Delphi participants; none – not administrators, nor teachers, nor mentors, nor yet academics – deviated from a strongly held belief that neither the accountability that motivates the creation for competencies nor the competencies themselves could ever substitute for the development of the teacher as a profoundly human entailment driven by the imperative of humans to constantly realize themselves both as individuals and members of community. But, in all of this, participants were not seeking to conserve some reductive account of transmission, where previous generations would simply pass on their wisdom, a set of practices and schedules that were somehow immutable. Rather, they were recognizing that the language and practice of education is a human conversation. There was universal agreement among participants that AI could not substitute for the human agency involved in the all-too-human activity of education. AI may offer resources that 'might' enhance the evidence that we draw on in making our judgements, but it is unlikely to substitute for the interstitial, the liminal, the eruptions and responses that constitute the human.

8.5 Delphi Panel Deductions

It is difficult to answer the, arguably, very obvious question, what does this mean for changing the way we educate future teachers? Before tentatively offering some suggestions

that seem to resonate with the other parts of this study, it is worth making a few general observations. As we have recounted, the Delphi symposium threw up myriad themes, none of which are new to those of us who have spent much of our career deliberating on such matters. The conversations did, however, as we have noted, offer novel and interesting ways of framing these. Perhaps key to the whole conversation was the inability, over more than a day of deliberation, to refine the question sufficiently to end up with any kind of singular focus. Yes, the conversation gradually focused on a small number of issues, but all the other concerns raised in the earlier iterations kept erupting into the spaces of the later discussion. On reflection, there is a very good and, we would suggest, salient reason for this. Time and again, colleagues would start to focus on the question of competencies; time and again they would add another competence or source of evidence to the list. And, on every occasion that we would try to focus the question of reliability and provability, of security of judgement, participants would hedge their bets, arguing for the reinstatement of the 'it' – the elusive, the liminal, the personal. As we noted above, senior authority administrators had no difficulty in rejecting the conceit of the guarantee. And there is a very good reason for this!

Competencies are manifest in observable actions, in the action of the teacher and the reaction of the pupil: Did she use the resources well? Were all the pupils engaged? Did they understand the task? Was their comprehension manifest in their work? These are all questions that make themselves present in the time and space of the classroom or sports field or playground. But the questions as to how the student teacher 'stands' in those spaces, whether or not she loves the world and the children as beings of the world, are of a different kind – they are not observable in time and space but in the connective tissue and the interstitial byways of the classroom. And that is why, after all the competencies are enumerated and assessed, there is always a remainder, a surplus that is not measured but sensed.

The struggle experienced by the panellists to 'nail down' the competencies emerges from the challenge this project has faced from the outset and which we struggle to resolve by reflecting on the application of competence assessment in practice, by literature searches, by crosswalking varied lists of competencies. All of this is done in the hope of a kind of release from the messiness of our everyday humanity and our arrival on some platonic plateau where we will have scientifically established the pure form of the competence. But as Cochran-Smith (2021), among others, has pointed out, competencies are neither good nor bad; they are tools and they can be put to good uses and poor uses; they can be helpful, and they can be an obstacle. This recognition may be particularly helpful in understanding how we can better deploy our judgement. What actually emerges, then, from the difficulty of the panellists to arrive at the definitive account is something so obvious that it is often overlooked – competencies are cues for and clues about where to look in order to make a judgement; they are not a replacement for it. Judgement is a quite different application of our mental processes. We return to this below.

First, let's us turn to another issue with important implications for how we go forward. Competencies, as we have noted, emerge not only from a sense of public obligation but also from the mistrust of professionals. And, both of these carry important messages for teacher educators and their capacity to judge. It is not unreasonable that those who spend significant

sums of tax payers money should not do so without some sense of accountability (Institute of Chartered Accountants in England and Wales, 2018), and education can surely not be immune from that (Conroy & Smith, 2017). The question of professional mistrust usually surfaces when it appears that vested interests (be they unions, local government, central government, or universities) are less concerned with improving the particular culture/opportunities/service than they are with securing their own position (Beck, 2016). But, as Cochran-Smith (2021) has pointed out:

> More accountability is not necessarily better than less accountability; less accountability is not necessarily worse than more. Rather, the virtue or vice of any accountability scheme, initiative, or system depends on the larger policy and political agendas to which it is attached, how it is used, the goals, values, and purposes it serves, and the assumptions it makes about who should be held accountable for what, to whom, under what conditions, with what consequences, and brokered through what power relationships. (p. 9)

What matters here, and what mattered to our participants, is how we arrive at and draw on our theories and substantive competencies. While the discussion above with respect to the laundry list might be construed as a rather frustrating exercise in ensuring all the potential or actual parties to early career development get the opportunity to secure their particular favourite competence, it is in reality rather more than that. Arriving at place where we have useful competencies is a collective exercise precisely because teacher education is a collegial exercise. Knowledge about what might be a useful marker of an early career teacher's efficacy is not the prerogative of any particular group or individual. It is, rather, the responsibility of all involved in teaching and school education. In this important respect, judgement is always social. The creation of competence frameworks are not merely arbitrary but rather issues from a practice and that practice is social.

This inevitably connects to another issue that ran through the whole Delphi process from the preliminary reflections to the summative session: the importance of some version of a school-based or clinical model. As one participant put it:

> So we developed the clinical teaching cycle in collaboration with colleagues across the faculty in the school, and school stakeholders, if you like, partners. So the clinical teaching cycle, which you know in truth probably looks similar to a curriculum cycle in other places, but it was shared. So we had [xxx] and his work coming into the programme and so all of our people who are observing our pre-service teachers in that clinical model had a very basic sheet. That said: What is the student doing, making, saying, or writing? What is the teacher doing, making, saying, or writing? … And then they observed and then the judgement was working with the pre-service teacher to say this is what I observed in this classroom. So actually, what the judgement part was was not in that initial moment, it was taking down the data. And this is what we were speaking about before and what I think AI could do, you know, to actually give you what's happening.

Other colleagues concurred and one observed that a collegial approach underpinned by theoretical considerations was constitutive of judgement. Here, he suggested the logic of differentiation of task and responsibility was key to understanding such an intimately collaborative approach.

> That process is having a theory of judgement. … We used MacIntyre in Theories of Practice. So that, now this; it wasn't alchemical. It didn't work a transformation x, but it meant that there was a co-constructed, usable version of judgement-making shared by all the participants, and the process, rather than just a kind of fuzzy logic where we kind of know what a judgement means … we know how to judge a good football player, we know how to judge a good cake. We're faced with, you know, popularizations of judgement on our television, television sets all, all the time. This was something that was intrinsically educational because you're using Bernstein and McIntyre, you're using people who are educators and who formulated that judgement as the exercise of a practice.

He went on to elaborate, suggesting that there were legitimate differences between the ways in which a hard-pressed classroom-based mentor might make a judgement and the ways he, as an academic who hadn't spent significant time in a classroom teaching for many years, might. But what he brought with his knowledge of theorists, theories and effects with respect to judgement, was an important element in the creation of a shared language and it's import. Another colleague observed that 'it's a multiparty process, [which] means that there's a few people involved and that the source or the type of information is not direct measurement.'

There was strong agreement among the participants that while it might take a variety of clinical forms, effective professional judgement requires deeply structured collaborations of a kind that are rarely, if at all, seen in teacher education in the UK and beyond. And yet, in places like Melbourne and Glasgow, which had pioneered embedded clinical models of development, supervision, and assessment (McLean Davies et al., 2015), these models had struggled to maintain their original guiding impulses for two different reasons. First, there has been some hostility from teacher educators based in universities who (unlike medics) have struggled with the idea of returning to schools for a significant proportion of their professional lives. Second, and more explicitly discussed in the context of this forum, is the cost of a clinical model. All were agreed that the optimal model of teacher education, rooted in the collegial exercise of judgement could not be secured without significant investment. But to date such investment has not been in evidence. Hence, in Scotland the fee for an education student in 2011 was c. £8,400 and the institutional contribution in the University of Glasgow was 37% of that total, leaving a net per capita income of c. £5,292. In 2024 the gross per capita income is c. £,6,800, of which some 50% represents the institutional contribution, leaving c. £3,400. After applying a cumulative inflation multiplier of –32%, the actual net sum available is a somewhat risible £2,312 at 2011 prices. It is not actually plausible that we can have a world-class teacher education programme on this level of income. Of course, this is only one example of inadequate funding, and education programmes elsewhere in the UK may fare somewhat better. However, the scale of the

problem is apparent and comes from decisions to treat education as if it was conducted on the same basis as social sciences more generally.

Here we would like to insert a couple of important cautions. First, it is not at all self-evident that teacher educators themselves receive any particularly sophisticated education in the art of judgement. Second, as Dylan Wiliam (2023) has suggested, it may be that even senior teachers in schools show no particularly advanced capacity to identify teaching excellence. In a recent paper, surveying a number of studies, he observed that senior teachers were, at best, no better than 50% accurate in their observation-based judgements. However, rather that thinking of this as a rebuttal of the importance of observation, it may be considered as reinforcing a strong theme among participants – the need to do better! Central to the conversation both in the preliminary responses and in the day-long Delphi conversation was a desire to do better, to, as a number of participants put it, 'take back control' and to valorise the effective practice of teaching. Surprisingly little of the conversation was concerned with the curriculum per se, though there was a strong thread on the need for integrated learning restructured around a project-based learning approach (not just traditional courses with lectures), these carried out during sustained clinical experiences with carefully selected school sites and strong school-based and university-based teacher educators.

Strongly rooted in this Delphi conversation, we would argue for a robust focus on nurturing mentors, on seeing those mentors as part of a team, where the whole team is dedicated to continuous renewal, which should not be considered a synonym for endless process but an invitation to exchange, reflect, and 'play' with teaching. As we face the next technological advance, we should not see these as fixes for a problem that can be so fixed. Rather, we should view the advent of AI in much the same way that the panellists recommend that we view competencies – not as a substitute for judgement, but a resource. AI may indeed offer a useful resource to facilitate the professional development of teachers, old and new. This may well help in creating flexible and creative simulations where innovations in theory may be practised in a safe and engaging environment. It might also involve student teachers, during their residency/practicum, in teacher-led research being conducted by their mentor alongside academic researchers. These suggestions are at one with the myriad conversation about clinical, cohort, and other collegial models.

While the observations here emerge directly from the considerations of our experts, we were also struck by how much of the conversation reflected those of the teacher educators in our study and indeed so many of the findings in our literature review. In these regards, the Delphi technique helped to further clarify and validate findings from the other phases of this project. While Wiliam's (2023) scepticism as to the efficacy of observation as a tool in the improvement of the material quality of education experienced by pupils has to be addressed, it is not done so by running away from the task of judgement but, rather, by standing squarely to its limitations and imperfections and in so doing improving it. All the evidence in this exercise would suggest that what we require is a different kind of teacher education than that traditionally conducted. As teachers and teacher educators, we are not insensible to the accountability we have to our early career colleagues, to our communities, to our societies,

and to our politics. But this must be, as Onora O'Neill (2013) would have it, intelligent accountability.

8.6 Conclusion

In this chapter, we presented findings from the Delphi panel carried out with international experts in education. The panellists considered the convergent results derived from the review of literature, comparative analysis of professional teaching standards, and empirical case studies (Phases 1–3 of this study). Analysis of the Delphi panel discussions revealed a complex interplay of themes surrounding teacher education, highlighting the tension between standardization and individualization, theory and practice, and democratization and accountability. While participants acknowledged the importance of competency-based frameworks, they also emphasized the need for a more holistic approach that considers relationality, context, and the human element of teaching. The findings underscore the necessity of ongoing dialogue, collaboration, and a shared commitment to fostering a supportive and intellectually stimulating environment for early career teachers. Ultimately, the success of teacher education lies not solely in the mastery of technical skills but in the development of a reflective, adaptable, and compassionate teaching profession.

Taking forward the findings from the Delphi panel presented in this chapter, in Chapter 9 we provide a synthesis of findings from Phases 1–4 (as presented in Chapters 3–7), with the findings of the convergent analysis presented to answer the research questions. Chapter 10 puts forward an emerging model, based on our findings, to inform judgement-making, and in Chapter 11, conclusions and recommendations are offered.

9 Convergent Cross-Case and Cross-Phase Analysis

This chapter includes results of both the cross-case and cross-phase analyses. First is the cross-case analysis. After considering the findings from each of the descriptive cases of teacher education programmes aiming to prepare high-quality teachers (in Chapters 5 and 6), we proceed to examine relationships across the three cases. The holistic cross-case approach was carried out with the aim of retaining the integrity of each case and noting any patterns and connections across the cases. The cross-case analysis was carried out using Morse's (1994) four-stage framework – comprehending, synthesizing (decontextualizing), theorizing, and recontextualizing – paired with the coding analysis strategies of Miles and Huberman (1994) – data reduction, data display, and drawing conclusions. The process thus reflects a case-based rather than a variable-based approach (Yin, 2018) to distil key findings. We acknowledge the setting aside of data from the case study in Wales as itself noteworthy for further investigation in due course.

9.1 Findings of Cross-Case Analysis

A cross-case analysis was conducted to build a general explanation that fits the multiple cases, giving consideration to the details specific to each case (Merriam & Tisdell, 2016) and considering the sufficiency of comparability between them to warrant any presumed common findings (Yin, 2018). Thus, the cross-case analysis involved only two of the research sites, Scotland and England, as per the outcome described in Chapter 7 of the third case, situated in Wales. Through the analysis, we sought to identify relationships, common findings, and contradictions in order to draw conclusions from the cases. To that end, we bring forward convergent findings and replicative occurrences across the cases. These include an emphasis on both deconstruction and affirmation of the complex, collaborative work required to ensure new teachers enter classrooms ready to teach.

9.1.1 Convergence of Findings: A Fair Process

Although situated in different, complex environments and unequivocally unique, the cases in Scotland and England show no marked difference in overall results. The cases therefore confirm one another and identify challenges associated with making judgements about teacher candidates' practice and the quality of teaching. Findings from both cases emphasize critical considerations related to roles and perspectives of the three groups of evaluators and their interconnected responsibilities contributing to the whole of the school-based experience within initial teacher preparation. Both cases also demonstrate the intricacies, tensions, and challenges of assessing student teachers during their preparation amid such a vast array of influences.

Comparative analyses were used to determine patterns of consensus and dissensus among the judges. Taking into account the relatively small sample sizes, findings from both case studies revealed a high degree of congruity between the respondents with respect to their judgement of teaching effectiveness, as seen in their satisfaction ratings, their rationales for decisions, and their approaches to judgement-making. Participants in both Scotland and England considered the dimension of 'learning environment' as the easiest to judge. Additionally, both

cases demonstrate that while there was a degree of variability in the rating of the seven dimensions of teaching from the video task (see Sections 5.3.2 and 6.3.2), there was less variation in holistic overall ratings. There was also a common backing of judgements using observed classroom cues, though we are not able to establish whether or not these cues and rationales were deployed in the same way. The findings show that participants situated within these providers of initial teacher education (ITE) shared similar views on the importance of accurate, consistent, and evidence-based judgements and the value of professional judgement. There were also clear indicators from both groups of participants that classroom observation protocols should be explicitly focused on observable teaching practices, and that these results from multiple observations over time should be included among a suite of measures in order to constitute a fair and accurate judgement of teaching and readiness for the profession. Also reflected in both case studies was the need to acknowledge that teaching is not a solo act; it is highly collaborative and whole-school factors have an effect. It was clear from questionnaire responses that context matters (Q13d), but there was less clarity about how and in what way. It was therefore acknowledged by participants that it was risky to make consequential decisions about a student teacher's classroom readiness without triangulated approaches. There was also encouragement in both settings for dialogue and conversation to be an integral part of the gathering of 'cues'.

The desire and need for fair and just practices in judging teaching practices of student teachers was emphasized in both cases. The ways in which to do so were less clear for all. The cross-case analysis reconfirmed that observation measures which involve human judgement tend to be less consistent than evaluations that have binary outcome (i.e., a single correct answer (Bell et al., 2019; Hill et al., 2012), yet participants noted that consistency in processes and commitment to consistent messaging about purpose were both possible and imperative. Several suggestions were provided about how consistency could be gained, including through selection of evaluators and mentor teachers, training and preparation for evaluations, and careful selection/creation of the observation tool itself, to name a few. As Bell et al. (2015) noted, it is very difficult to ensure raters assign scores in the same way; however, standardizing some conditions in processes could result in greater, appropriate, consistency (Boguslav & Cohen, 2024).

A theme that transcended commonalities and agreement between the cases was the concept of justice in ITE. This concept evokes a broader applicability that is not dealt with in prior research, yet we contend it is an essential characteristic of high-quality teacher education. Although teacher preparation for social justice in their practice has been widely studied, research regarding the just practices of preparing future teachers is difficult to find. Participants in both case studies identified strategies to gain consistency and reliability that invoke the principles of distributive, procedural, and interactional justice (Rasooli et al., 2023) through their language and descriptions of fairness. We see these as incredibly vital to gaining confidence in judgement-making processes and outcomes. Distributive justice involves ensuring that student teachers have equitable access to resources, support, and opportunities, such as access to high-quality school-based experiences with strong mentoring, access to necessary technology and materials, and understanding of the standardized

instructional activities set to be observed during clinical placements (Boguslav & Cohen, 2024). If we contend that mentors matter, then our processes, procedures, and investment in school-based experiences and evaluations conducted in those settings should reflect this.

In ITE, procedural justice is a focus on ensuring processes used to evaluate and assess teacher candidates are fair and transparent. Using clear and consistent criteria for assignments, assessments, and observations, and providing opportunities for candidates to receive feedback and make revisions, are essential. Consideration must be given to factors such as accuracy, impartiality, correctability, participation, and reasonableness. Interactional justice is centred on treating student teachers with respect, dignity, and fairness throughout their preparation, both in university and in schools on placement. This involves a supportive and inclusive learning environment, encouraging open communication and collaboration between student teachers and teacher educators, tutors, and mentor teachers, and providing timely and constructive feedback (Rasooli et al., 2023, p. 262). By incorporating these concepts into judgement-making in ITE, teacher educators can gain consistency as well as help to prepare new teachers to understand and promote fairness and equity in their own future classrooms.

9.1.2 Nuances and Complexity

While the holistic features of each case are confirmational, the cross-case analysis revealed a few distinctions that warrant conceptual consideration. It is important to note, however, that the small sample size limits claims about patterns. Although overall results of the video observation task to rate dimensions of teaching were largely similar, there was much less variability in the judgement exercise in England, where participant responses were also found to be skewed toward the top ratings. There was a greater variation in the responses of participants in Scotland, particularly among the teacher educators. This was of interest given the phenomenon of scores clustering at the highest range which has been noted in prior research measuring teaching skills (Kraft & Gilmour, 2017). The realities and complexities evaluators must navigate when judging practices may explain this grouping of ratings. The authors suggested moving away from a focus on summative performance rating toward a more multidimensional approach allowing for identification of areas of practice to support more accurately. Kraft & Gilmour (2017) suggested a different question needs to be asked. Instead of asking ‘how effective is a teacher?’, they propose asking ‘how is a teacher effective?’ (p. 243) to provide a more precise picture of teaching effectiveness.

The second holistic consideration which emerges from the case studies is the centrality of complexity and associated dynamics that impact on judgement-making. The cross-case analysis identifies several characteristics of complex systems in the case studies. First, both case studies reveal interconnectivity among various actors – university-based teacher educators, school-based mentor teachers, and associate tutors – in making judgements about teaching effectiveness. The diverse perspectives of each group shape how teaching is assessed, and these perspectives are highly non-linear and interconnected. The concept of emergence is also evident in both cases, as complex judgement processes create outcomes that could not be predicted solely by observing individual components of the system. Additionally, both case studies highlight adaptation as a key feature of the evaluation process.

Feedback loops play a critical role in how judgements are made, refined, and applied over time. The iterative nature of feedback during student placements illustrates feedback loops where formative feedback is provided throughout ITE, leading to an adaptive process in which student teachers adjust their teaching practices based on continuous input. Judgements in both case studies demonstrate unpredictability due to the subjective nature of evaluations and the sensitivity to contextual conditions. Outcomes of teaching assessments can be influenced by the environment of each placement, such as the school's socioeconomic context, the school's resources, the mentor's own experience, or the student teacher's background. These factors can significantly affect how teaching effectiveness is judged, leading to unpredictable outcomes. Finally, both case studies show self-organization within the complex system of teacher education, where the various stakeholders naturally develop ways to collaborate and make decisions without a centralized, top-down process. This collaboration between university-based teacher educators and school-based mentors is a form of self-organization, where the system functions through a network of evaluators working independently yet interdependently to assess student teachers. Findings from the two case studies demonstrate that characteristics of complex systems – interconnectivity, emergence, adaptation, unpredictability, and self-organization – are central to how judgements of teaching effectiveness are formed. Together, stakeholders find ways to deal with inherent complexities instead of trying to mitigate them. The mixed methods approach captured some of the intricate and dynamic nature of judging teaching effectiveness, reflecting the non-linear and adaptive processes typical of complex systems.

9.2 Cross-Phase Analysis: Answering the Research Questions

In Phase 5, the final phase of this this mixed methods project (Creswell & Creswell, 2023), we looked to uncover the decision-making processes used by those who judge teacher candidates' readiness to teach, through a detailed investigation of what the judges specifically look for in order to make their decisions. A synthesis of findings from Phases 1–4 was thus conducted to achieve a richer understanding and to answer these three research questions. This involved examining significant patterns and relationships among findings from the different lines of enquiry and 'thinking upward' conceptually to draw meaningful conclusions (Yin, 2018, p. 197). The project was guided by three overarching research questions:

RQ1 What is the nature of shared judgement, consensus, and dissensus on observed teaching effectiveness among university-based teacher educators and school experience tutors/associate tutors and school-based mentor teachers?

RQ2 How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

RQ3 How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

Comparative analyses were used to explore the nature of shared judgement in determining teaching effectiveness and to ascertain patterns of consensus and dissensus. We looked to examine influences on judgement, levels of agreement, different weighting of cues, and

potential predictability in the ways judgements are made, as well as considering emergent insights from the judges. The empirical results from the two case studies are considered alongside findings from the systematic review of literature, the analysis of professional standards, and the Delphi panel.

9.2.1 The Nature of Shared Judgement (RQ1)

9.2.1.1 The Nature of Shared Judgement: Systematic Literature Review

The systematic literature review highlights that judgement of teaching effectiveness is influenced by various factors, such as evaluation tools, inter-rater reliability, and the differing perspectives of university staff, mentors, and tutors. The findings reveal that consensus is often found around observable teaching competencies, such as classroom management and student engagement, but dissensus arises when interpreting more subjective aspects like instructional innovation and reflective practices. The judgements of university staff are often theory-driven, emphasizing reflective practice and research-based pedagogy, while school-based mentor teachers prioritize practical, immediate classroom performance, leading to discrepancies in how teaching effectiveness is judged. For associate tutors, their role often blends both practical and theoretical perspectives, but they may align more closely with the mentors on classroom-based evaluations. The review found that consensus is more common in areas where evaluation instruments provide structured rubrics, but dissensus occurs in contexts where evaluators interpret results differently due to their roles and expectations. There was a high degree of variability regarding the competencies that should actually be looked for in an observational judgement, as evidenced by the variety of evaluation tools.

9.2.1.2 The Nature of Shared Judgement: Professional Teaching Standards Policy Review

The comparative analysis of teaching standards reveals that shared judgement, consensus, and dissensus on teaching effectiveness vary across the three nations, with each having distinct policy and cultural contexts that shape how judgements are made. In Scotland the standards emphasize professional autonomy, research-informed teaching, and reflective practice. The consensus among educators – both university-based and school-based – focuses on continuous professional development and an emphasis on holistic student development. Dissensus may arise in the assessment of practical versus theoretical teaching approaches, as school-based mentors prioritize immediate classroom effectiveness, while university staff emphasize reflection and research. In England the standards are more prescriptive and focus on specific competencies, such as behaviour management and curriculum knowledge. This results in consensus around performance-based measures but potential dissensus between mentors and university staff regarding innovation and teacher autonomy. University staff may favour more flexible, research-driven evaluations, while mentors might be more aligned with concrete, observable classroom practices. The Welsh standards emphasize collaboration, innovation, and professional development, aligning closely with UNESCO's *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019). There is greater consensus on the importance of professional growth and reflective teaching, but dissensus may occur in how innovation is interpreted in practice, with

university staff potentially focusing more on theoretical innovation and mentors emphasizing practical classroom solutions.

9.2.1.3 The Nature of Shared Judgement: Case Studies

Overall, judgement-making in this study was considered a careful and well-reasoned professional duty – and widely variable. Though the results also suggest that there is broad agreement about key considerations, with a consensus agreement that a level of consistency is essential. The first case study conducted with a TEP in Scotland reveals a varied but structured approach to assessing teaching effectiveness. Participants provided judgements based on a range of strategies, primarily focused on observable classroom cues, professional judgement, and suggestions for lesson improvement. There was general agreement that teaching effectiveness should be judged based on standards and observable classroom behaviours of the student teacher and the pupils. The dimension of the 'learning environment' was consistently rated as the easiest to judge across all groups, with shared agreement that this dimension had the clearest cues based on evidence from classroom observation. Associate tutors showed higher agreement than teacher educators and mentor teachers, especially in areas like 'learning environment' and 'content'. The use of professional judgement to evaluate teaching effectiveness was highly valued by all groups. There was variation in how different groups rated dimensions like 'instructional strategies' and 'assessment', with mentor teachers showing the greatest variability in ratings. Teacher educators were more divided on aspects of teaching, with some finding these elements difficult to evaluate without more contextual understanding. This variability indicates a level of dissensus among groups, particularly in how they perceived and rated less concrete aspects of teaching.

The case study in England presents a mixed pattern of consensus and dissensus among the three groups (i.e., university teacher educators, tutors, and mentor teachers) when evaluating teaching effectiveness. While there was a shared focus on using professional judgements based on classroom cues (e.g., teacher and pupil actions), significant variation existed across groups. This dissensus reflects differences in their roles and experience in observing and supporting teaching practice. Teacher educators and tutors used classroom cue utilization as their primary strategy but differed on specifics such as suggestions for improvement and perceived omissions during lessons. Mentor teachers, grounded in practical classroom experience, provided more critical assessments with higher variability in their judgement strategies. Thus, while a general consensus existed around core elements like 'learning environment', dissensus emerged in more nuanced pedagogical aspects like 'instructional strategies' and 'assessment'.

The two case studies on judging teaching effectiveness reveal several key points of consensus. First, participants agreed on the importance of looking for growth and development over time, rather than focusing on a checklist for implementation. Teaching effectiveness was assessed through multiple sources of evidence and formative judgements that lead up to a final summative judgement, signalling classroom readiness. Participants emphasized the role of professional standards and judgement in guiding evaluations,

recognizing the need for both objective measures and subjective, context-driven insights. Additionally, they noted the challenge of articulating judgement strategies, such as lesson improvement, indicating that evaluators often work within an idealized process that may not align with their actual practices. Collegial decision-making helped to clarify these nuances, with much of this work being difficult to express publicly. There was also a cultural element, particularly in Scotland, where evaluators tended not to prioritize strengths first, a contrast to approaches elsewhere. The case studies also highlight the limitations of the clinical model, which has been deemed unsustainable under some current school–university circumstances. Finally, the idea of the 'good teacher myth' was challenged, as participants agreed that teaching effectiveness cannot be fully predicted or defined by rigid frameworks.

A few areas of dissensus also emerged. One key area of debate concerned which aspects of teaching were the hardest and easiest to judge. Some evaluators found the learning environment the easiest to assess due to visible, concrete cues, while others struggled with assessment or instructional strategies, reflecting the challenge of interpreting deeper pedagogical processes. Another point of dissensus revolved around the balance between consistent standards and professional judgement. While there was agreement on the need for both, participants differed in how to best implement this balance – some favoured strict adherence to professional standards, while others argued for more flexibility to accommodate the complexities of individual classrooms. Finally, the participants debated the degree of inconsistency that should be allowed in judgements, with some accepting small variances as inevitable in a complex system, while others expressed concern that too much inconsistency could undermine the fairness and credibility of the evaluation process.

9.2.1.4 The Nature of Shared Judgement: Delphi Panel

The Delphi panel discussions highlight both consensus and dissensus in the evaluation of teaching effectiveness. The panel generally agreed that assessments of teaching effectiveness should be grounded in theoretical frameworks and encompass both observable competencies and dispositional traits, such as professional attitudes and moral commitments. This reflected a consensus on the value of holistic evaluation, where teaching is about not just classroom behaviour but also underlying dispositions that impact long-term teacher effectiveness. However, dissensus emerged around the importance of local context and individual experience. Different stakeholders, including university staff, mentors, and tutors, often emphasized distinct aspects of teaching. The panel also discussed the difficulty in measuring the 'it' factor, a quality believed to be essential but hard to assess consistently.

9.2.1.5 The Nature of Shared Judgement: Cross-Phase Summary

The findings across the literature review, policy analysis, case studies, and Delphi panel discussions highlight both consensus and dissensus in the judgement of teaching effectiveness. Consensus often centred on observable teaching competencies, such as classroom management and student engagement, with shared agreement on the importance of professional standards and judgements informed by multiple sources of evidence over time. Participants agreed on the need to look for growth and development in student teachers rather than focusing solely on checklist-based evaluations. There was also widespread recognition

of the importance of professional judgement to accommodate the complexities of teaching, with most evaluators emphasizing the need for flexibility. However, dissensus emerged in more nuanced areas, such as 'instructional strategies' and 'assessment', where university educators often prioritized reflective practice and theoretical frameworks, while school-based mentors focused on practical classroom performance. Additionally, disagreements arose around which aspects of teaching were the hardest and easiest to judge, with the learning environment seen as easier due to its observable cues, but assessment and instructional strategies proving more challenging. Finally, there was variation in how much inconsistency in judgement was acceptable, with some advocating for flexibility, while others feared it could undermine the fairness and credibility of evaluations.

9.2.2 Fostering Collaboration (RQ2)

9.2.2.1 Fostering Collaboration: Systematic Literature Review

The systematic review indicates that reliability in professional judgement, especially through standardized tools and training, plays a crucial role in improving collaboration between schools and universities. When evaluators (i.e., university staff, mentors, and tutors) are trained to use evaluation tools in a consistent manner, reliability increases. A shared understanding of how teaching effectiveness is assessed leads to greater trust between university educators and school mentors. This is essential for fostering a collaborative environment. When judgements are more reliable, dissensus diminishes, which minimizes conflict among those making the judgements. The review also suggests that reliable judgements help integrate practical classroom experiences (from mentors) and theoretical perspectives (from university staff). This leads to more effective and collaborative decision-making on student teacher evaluations.

9.2.2.2 Fostering Collaboration: Professional Teaching Standards Policy Review

Enhanced reliability of professional judgement across Scotland, England, and Wales can foster greater collaboration between schools and universities by creating a more aligned and transparent evaluation process. In Scotland, the reflective nature of the teaching standards, when combined with more consistent judgements that integrate both research-based and practical classroom outcomes, encourages a shared understanding of teaching effectiveness. This shared framework promotes collaboration through professional learning communities, where university staff and school mentors jointly engage in teacher development. In England, where prescriptive standards dominate, enhancing reliability through continuous professional development and standardized observation tools aligned with international benchmarks like the UNESCO *Global Framework* can bridge the gap between theoretical knowledge and practical application. This would result in more uniform evaluations and reduce tensions between university staff and school mentors, thus improving the collaboration between the two. In Wales, the focus on joint ownership of teacher education already provides a strong foundation for collaboration. Improving reliability by promoting shared assessment criteria and increasing communication between mentors and university educators ensures that both parties are interpreting and applying standards consistently. This would reduce dissensus and foster greater trust, enhancing the overall partnership between schools and universities. Thus,

across all three nations, enhancing the reliability of professional judgement creates a more cohesive and consistent evaluative framework, which bridges theoretical and practical approaches and strengthens collaboration through shared understanding, improved communication, and unified assessment practices.

9.2.2.3 Fostering Collaboration: Case Studies

Findings from the case study conducted in Scotland highlight several ways that enhanced reliability of professional judgement can foster greater collaboration between TEPs and schools. A key aspect of this enhanced reliability is the establishment of a sustained residency model, similar to a clinical model, where school-based mentors and university staff are more interwoven in the process of teacher education over time. Such a model provides student teachers with more opportunities to develop the competencies required for effective teaching, as both settings contribute consistently to their growth. Reliable and consistent judgements, particularly when based on agreed standards set by the General Teaching Council for Scotland, were seen as essential for fostering trust and transparency between university-based and school-based educators. When both groups apply consistent criteria and maintain open dialogue about assessment, it enhances mutual understanding, ensuring that professional judgement is aligned and fair. Moreover, ongoing dialogue – with the skills, time, and space to conduct meaningful discussions – was seen as crucial for maintaining reliable evaluations over time. By fostering professional dialogue between school-based mentors and university staff, both groups can better align their approaches to support student teachers, making evaluation a more collaborative process. The findings also underscore the importance of providing sufficient time to develop competencies, ensuring that judgements reflect the student teacher's overall growth rather than isolated observations.

Similarly, the findings from the case study in England reveal several opportunities for improving collaboration, with a key suggestion being the use of a sustained clinical model. Both university-based teacher educators and school-based mentors emphasized that consistency in evaluations is essential. A standardized and reliable judgement process would minimize dissensus between the parties, ensuring clearer communication and fostering better alignment in expectations. The findings also highlight the importance of a cohesive training approach for both university staff and mentors to ensure that professional standards are applied uniformly. Training both groups together would help synchronize evaluative practices, reduce variability in judgement, and support a shared understanding of the roles and responsibilities of each party. Additionally, by empowering school-based mentors with a more formal role in the evaluation process, the power dynamics that often favour university educators could be mitigated. This would ensure that mentors feel their assessments are valued equally, fostering a deeper collaboration between schools and universities. The sustained dialogue and co-construction of the evaluation process would help build a more reliable, transparent system that benefits both parties, ultimately leading to fairer and more comprehensive assessments of teaching effectiveness.

9.2.2.4 Fostering Collaboration: Delphi Panel

The Delphi panel identified several ways in which enhancing the reliability of professional judgement could promote collaboration between schools and universities. The use of co-designed, standardized assessment frameworks was suggested to create a common language for stakeholders. This would lead to more consistent evaluations and help reduce the differences in judgement which currently stem from varying professional backgrounds. Additionally, strengthening the role of mentors through better training and support could enhance the reliability of judgements, contributing to more aligned assessments between schools and universities. The experts emphasized the need for extended student-centred mentorship programmes that are fully integrated into teacher education, fostering deeper collaboration. The experts proposed that reliability of judgement-making should be part of a continuous feedback loop rather than a series of isolated assessments. This would allow for ongoing learning and development, encouraging more frequent collaboration and dialogue between school-based mentors and university staff. The panel also discussed the importance of co-development of language of student teacher dispositions and finding a better way to assess complex skills. They suggested descriptors of a teacher's practice are needed that are more congruent with the landscape of fact and have more clarity in markers that take a student teacher from the ordinary to the extraordinary. There was a sense that excellence was disappearing under the guise of egalitarianism, and this was potentially due to the democratization of the decision-making in teacher education. They suggested that it is not necessarily to the benefit of the profession for everyone to have a say. Of the many nodes of agreement among participants was a collective acknowledgement of the extent and range of the considerations necessary to ground an effective judgement as to early career efficacy. While all acknowledged the not inconsiderable challenge of competence proformas that reflect the complexity of the thing-in-itself, there was a collective belief that some such type of proforma (or rubric, checklist, etc.) was indeed required to maintain the integrity of judgements. The discussion surfaced a complex list of considerations on which to base an observation proforma that helps evaluators conceptualize the more difficult, but arguably more important, competencies a new teacher must enact. The panel suggested co-creation of an alternative set of competencies for making judgements could be co-developed with school partners. The different competencies discussed by the panel were collated from the discussion transcripts and included in Table 9.1.

Table 9.1

Different Competencies

Core dispositions and attitudes	Interpersonal skills and relationships	Best practices
Beliefs and values: • Belief in a reason for teaching • Knowing own values	**Communication:** • Dialogic • Intercultural communication	**Teaching and learning:** • Impact on pupil learning and development • Responding to the unknown and unexpected

- Personal language of working with pupils

Emotional intelligence:

- Reading emotions
- Accepting critical feedback
- Relationality
- Resilience
- Identity

Self-awareness:

- Knowing why something worked or didn't
- Understanding context
- Understanding how to judge themselves
- Metacognition
- Knowing own values
- Reflexive practice

- Metacognition of teaching practices
- Articulation of teaching practices
- Asking for help
- Self-advocate for your needs
- Negotiation
- Handling difficult conversations

Collaboration:

- How to teach in a team
- Creating spaces in which people can flourish

Empathy and nurture:

- Nurture
- Human interaction
- Building a sense of belonging

- Flexible
- Agile
- Critical engagement with AI
- Thinking competencies
- Contextual decision-making
- Coachability

Transferable skills:

- Improvization
- Initiative

Professionalism:

- Reflection in action and on action
- Internalizing the competencies
- Building a reservoir of experience
- Growing professional repertoire
- Notions of critique
- Identifying sources of support

The panel cautioned to avoid creating a 'new' laundry list of competencies for a new teacher to demonstrate. Some competencies identified may overlap between categories, reflecting the interconnected nature of effective teaching. Interestingly, many of these skills are reflected in the UNESCO *Global Framework* in the domain of teaching relations (i.e., collaboration, communication, and professional development), which includes dimensions of teaching that would not necessarily be observable in a lesson observation. This finding from the panel supported suggestions for a multifaceted approach in which multiple sources of evidence were gathered and critiqued in order to render a fair judgement of a teacher's practice.

9.2.2.5 Fostering Collaboration: Cross-Phase Summary

Enhanced reliability of professional judgement plays a crucial role in fostering greater collaboration between schools and universities by creating a more aligned, transparent, and consistent evaluation process. Findings across the systematic literature review, policy analysis, case studies, and Delphi panel discussions emphasize that when standardized assessment tools are used consistently by university staff, mentors, and tutors, trust and transparency between schools and universities improve. The use of sustained residency models and agreed standards allow for more consistent and reliable judgements, fostering mutual understanding and encouraging dialogue between school-based mentors and

university educators. Additionally, enhancing reliability through collaborative training and standardized evaluations can minimize dissensus, ensuring better alignment between theoretical knowledge and practical application. Co-designed frameworks and continuous feedback loops were suggested to create a common language for evaluators, promoting joint decision-making and reducing the impact of power imbalances. The co-construction of the teacher education experience, supported by ongoing professional dialogue, strengthens partnerships, ensuring both theoretical and practical perspectives are valued equally. Ultimately, reliable professional judgement enhances collaboration by ensuring that all stakeholders have a voice in the teacher education process, leading to more cohesive, fair, and effective judgements. The collaborative approach as envisioned is often not experienced as such by the teacher educators, tutors, and mentor teachers, who recognize the lack of recognition, investment, and time. There appears to be a lack of expediency to find a feasible way to address logistical and financial constraints that prevent use of these 'better' strategies to foster collaboration.

9.2.3 Power Dynamics (RQ3)

9.2.3.1 Power Dynamics: Systematic Literature Review

Power dynamics, as highlighted in the systematic literature review, shape the roles of university-based and school-based educators differently. University-based teacher educators typically hold more authority in formal evaluations, especially in contexts where reflective and research-based teaching practices are emphasized. The review notes that university educators often have the final say in the summative evaluations of teacher candidates, reinforcing their dominant role. Although school-based mentor teachers provide practical insights and continuous feedback, their role is sometimes seen as subordinate to the university's academic standards. This creates a power imbalance where mentors' practical judgements are not always given equal weight in final assessments. The review suggests that balancing power dynamics by giving mentors a more formal role in the assessment process could reduce tensions and improve the reliability of judgements. This could help align both theoretical and practical evaluations, leading to a more equitable partnership between schools and universities. Addressing power imbalances between university-based and school-based educators could foster more equitable judgement processes.

9.2.3.2 Power Dynamics: Professional Teaching Standards Policy Review

Power dynamics are evident through the politization of education, as evidenced in the divergent policy reforms of the devolved UK nations as well as the government entities involved in administering universal education. The ongoing reforms demonstrate how ideological differences, economic interests, and special interest groups (such as unions and businesses) have an influence. The entities that set entry and completion requirements for ITE, provide funding, and call for strike action all hold power. It is also interesting to note the exceptional power dynamic in Scotland, with the General Teaching Council for Scotland as an independent regulatory body whose work involves speaking up for high standards in the teaching profession and influencing policy.

Within the assessment process itself, power dynamics shape how university-based and school-based teacher educators contribute to determining teaching effectiveness; these dynamics differ across the three nations. In Scotland, the standards promote professional autonomy, which gives both university and school-based educators a voice in assessments. However, the power dynamic still leans towards university staff, who have a greater focus on reflective practice. Power struggles may emerge if school-based mentors feel their practical, on-the-ground experience is undervalued compared to academic perspectives. In the context of the TEP in England, the highly prescriptive nature of England's standards positions university staff in a more authoritative role, particularly in assessing theoretical knowledge and behaviour management. School-based mentors may feel their role is more supportive than evaluative, which can create an imbalance in the partnership. Shifting to a more collaborative approach, where both parties' input is valued equally, could address these power dynamics. Wales appears to have made progress in fostering joint ownership of the ITE programme, but power dynamics still exist, with universities holding more formal authority over final judgements of teaching effectiveness. Empowering mentors by giving them a greater role in decision-making and formal assessments would balance the power dynamics, leading to a more equitable evaluation process.

9.2.3.3 Power Dynamics: Case Studies

According to findings from the case study in Scotland, power dynamics in the judgement of teaching effectiveness between university-based and school-based teacher educators are shaped by several factors, including time, collaboration, and the roles each plays within the triadic model (i.e., university staff, school-based mentors, and student teachers). University-based educators often relied on formal professional standards and structured feedback, placing them in a position of authority in defining what constitutes effective teaching. This reliance on academic criteria sometimes limited the influence of school-based mentors, who drew on their hands-on classroom experience to evaluate student teachers. These mentors focused on contextual knowledge of the classroom, which can conflict with the more formal assessments provided by university staff. However, the study also highlights that partnerships between schools and TEPs can be mutually beneficial. Mentor teachers often viewed their role in teacher education as a professional development opportunity, where discussions about good teaching become valuable for their own growth. This dynamic benefits both parties, yet there remains a tension in how much power mentors have in influencing final judgements. Ensuring that school-based mentors are seen as equal partners in the co-construction of the school experience and in the development and modification of the partnership can help balance these power dynamics. By involving all partners in decision-making processes and giving them a voice in how assessments are carried out, the complementary roles of both groups can be fully realized, fostering a more equitable partnership.

In the second case study in England, power dynamics also significantly influenced the roles of university-based and school-based teacher educators. University staff typically relied on theoretical frameworks and classroom cue utilization for evaluations, whereas school-based mentors were found to provide a more critical, practical perspective grounded in day-to-day teaching effectiveness. This created a potential imbalance, as university educators'

judgements – being aligned with research and academic standards – were often prioritized over the practical insights of school mentors, even though mentors bring a more rigorous, hands-on assessment of teaching performance. Mentor teachers often found their role within this partnership beneficial for professional development, as it gave them the opportunity to engage deeply with concepts of good teaching. Yet, the power imbalance remained, with university educators sometimes holding greater authority over final decisions. To mitigate this imbalance, the study suggests giving school-based mentors more formal authority in the evaluation process, ensuring they are fully involved in co-constructing the school experience and contributing to the development and modification of the partnership. This would also involve providing mentors with the time and resources necessary to engage in the evaluation process meaningfully, ensuring that their contributions are fully integrated and valued. Additionally, the core responsibilities of each member of the triad should be clearly defined, with all partners, including the student teacher, having a say in shaping the learning and assessment process. By ensuring that all partners are involved in the co-construction, maintenance, and modification of these partnerships, schools and universities can create a more balanced and collaborative environment, where both practical and theoretical insights are valued equally in assessing teaching effectiveness.

9.2.3.4 Power Dynamics: Delphi Panel

The Delphi panel findings underscore the influence of power dynamics on the roles of university-based and school-based educators in judging teaching effectiveness. Findings note the political, economic, and administrative dynamics in each educational context as having an influence. University-based teacher educators often hold the authority in summative evaluations, as their judgements are perceived to be grounded in academic research and theory. This creates a power imbalance, where the more practical insights from school-based mentors may be undervalued in the final assessment of a teacher's readiness. Although mentors play a critical role in day-to-day teacher development, their evaluations are often seen as secondary to those of university educators. The panel emphasized the need to redefine the role of mentors, more carefully choosing who they are, and advocating for a more formal and equal role in the judgement process. To mitigate these power imbalances, the experts recommended a tripartite evaluation, involving active collaboration and dialogue between mentors, university staff, and the teacher candidates themselves. This format was perceived to ensure the practical experience of mentors is valued equally. The panel also suggested that self-efficacy is a form of power, and power dynamics can positively or negatively contribute to developing self-efficacy depending on how power is wielded. There was acknowledgement that asymmetrical power dynamics are not necessarily harmful but could potentially lead to unequal treatment or limit opportunities.

9.2.3.5 Power Dynamics: Cross-Phase Summary

Power dynamics shape the roles of university-based and school-based teacher educators in judging teaching effectiveness by positioning university educators in more authoritative roles, particularly in summative evaluations. In contrast, school-based mentors provide critical, practical insights from their day-to-day classroom experience, which are sometimes

undervalued compared to academic perspectives. This creates a power imbalance where mentors' input is often seen as secondary, despite their important role in shaping student teachers' development. The findings across the systematic literature review, policy review, and case studies emphasize the need for a more equitable partnership. Co-construction of the teacher education process, including shared decision-making, can help balance these dynamics. Furthermore, providing mentors with more formal authority in the evaluation process and ensuring their active involvement in shaping the teacher education experience can reduce tensions. By clearly defining the core responsibilities for those within the triadic model and involving all parties in the ongoing development and assessment process, partnerships can become more balanced and mutually beneficial. This approach, as supported by the Delphi panel, can foster a collaborative environment where both practical and theoretical judgements are valued equally, ultimately leading to more effective teacher education.

9.3 Conclusion

In this chapter, we have presented findings from the cross-case and cross-phase analysis of this project. The findings provide insights into the complex dynamics involved in judging teaching effectiveness in ITE programmes. The study highlights how power dynamics, professional judgement reliability, and shared evaluation frameworks shape the interactions between university-based educators and school-based mentors. By examining both consensus and dissensus in judgement processes across different contexts, this research sheds light on the importance of balancing theoretical knowledge and practical classroom experience to foster fair and accurate assessments. Enhanced reliability of professional judgement, achieved through consistent tools, co-designed frameworks, and collaborative dialogue, can bridge the gap between schools and universities, leading to more equitable and transparent partnerships. Furthermore, addressing the inherent power imbalances by empowering school-based mentors and promoting shared decision-making processes ensures that all evaluators play a meaningful role in shaping teacher education. Ultimately, this research emphasizes that fostering deeper, sustained collaboration between schools and universities is essential for building a stronger, more cohesive system of teacher education, grounded in fairness, transparency, and shared responsibility. In Chapter 10, we present an argument that linear ways of considering judgement-making have overlooked important facets for effective preparation of teachers in an incredibly complex and ever-changing system of education, and we explore the formulation and application of a conceptual model which emerged from the project.

10 A Model of Dynamic, Adaptive Systems Thinking in Teacher Education

In the multiple phases of this project, we have explored the challenges of judging teaching effectiveness from multiple angles and perspectives that continually revealed the magnitude of complexity involved in the exercise. In this chapter, we propose a conceptual model developed through the analytical process of theorizing (Morse, 1994, p. 32) to aid providers of teacher education programmes (TEPs) in navigating these complexities. The process of theorizing involved development of working propositions and testing them against the data to ensure explanations were indeed congruent (p. 33). Therefore, this chapter presents an argument that linear ways of considering judgement-making have overlooked important facets for effective preparation of teachers in an incredibly complex and ever-changing system of education.

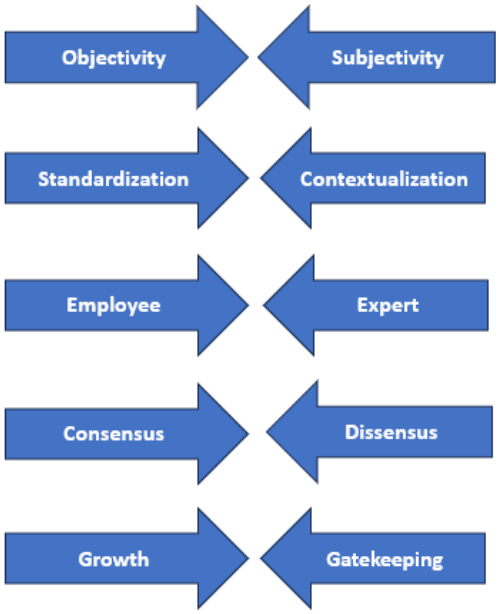
10.1 Development of a Model

Formulation of the conceptual model resulted from the work of teacher educators in the three participating institutions in a multi-phase project exploring the nature of judging new teachers' practices. We examined more closely impacting factors and looked for strategies and solutions to gain consistency and trustworthiness of judgements aimed to ensure new teachers' readiness to teach. In our exploration of professional judgement and professional standards in assessment of teachers' practices, a duplexity emerged. This duplexity was evident in findings of the systematic literature review (see Chapter 3), the review of professional teaching standards (see Chapter 4), and the case studies (see Chapter 5 and Chapter 6). Additionally, it was confirmed and validated by the Delphi panel of experts (see Chapter 8). The judgement process is not a duality in the sense that there are opposite or opposing themes, but instead that there are principal ideas composed of two parts, oftentimes operating together (see Figure 10.1). Both are needed and we thus see in the initial theorizing this duplexity revealed. While duality implies a straightforward division, duplexity conveys a more complex, ambiguous nature. Given these dynamic features, an adaptable decision model for judgement-making is necessary, one that is capable of adapting to changing contexts and situations.

As the convergence of findings coalesced in this project, the challenges of complexity continued to make themselves known. Complexity theory is thus at the heart of the model as a framework for thinking about how systems change, develop, evolve, and emerge (Davis & Sumara, 2008), in our case applied to judging the effectiveness of teaching. Martin et al. (2019) brought together a list of features complex systems display that indeed made themselves known in our research. These noted features are clearly evident in education – self-organization, emergence, nested, dynamic, difficult to predict outcomes, interactions, ambiguously bounded and positive and negative feedback loops (pp. 3–4).

Figure 10.1

Initial Theorizing of Duplexity According to Emerging Project Findings



The authors therefore acknowledged difficulties developing understandings in areas such as teacher education; the issues we explore are questions about patterns of complexity related to systems outcomes (such as teachers' practices and pupil attainment), the factors we investigate are not deterministic and cannot necessarily be understood through analytic means, and outcomes from social systems involve humans with beliefs and agentic actions which are difficult to predict (pp. 1–2). Biesta (2020) has reminded us that education itself, every encounter, always impacts on the student teacher as an individual; thus, teacher education can serve to either enhance or restrict capacities and capabilities. Additionally, as Cooksey (1996) noted regarding social judgement theory (SJT), and as the systematic review reconfirmed (see Chapter 2), judgement-making appears to remain a best estimate of the right choice under specific constraints and always runs the risk of being in error. SJT acknowledges there is a latitude of acceptance – a range of ideas we find acceptable and unacceptable. Furthermore, decisions of judgement are impacted by the simultaneity of influences from different levels, which prompts variability (Martin et al., 2019). Even small influences (e.g., how an evaluator grounds a judgement they observe) can have a cascading, consequential effect (e.g., a student teacher receiving licensure or not). Inconsistency itself varies with the predictability of the judgement task; classrooms happen to be far from predictable. The degree of ambiguity and variation with which decision makers can cope among an intertwined set of probabilistic relationships also varies from one context, and one evaluator, to another. We often see a step-by-step approach or deployment of heuristics to manage this complexity.

Therefore, while duplexity was evident in two seemingly opposite dynamics, being simultaneously necessary in making judgements of teaching effectiveness, it was clear the complexity of the task required additional reflection. The features of complexity of learning systems which arose in this study align with prior empirical work confirming that a variety of

systems at different levels influence teacher learning and decisions in teacher education (Martin et al., 2019, pp. 7–9). Faul and Savage (2023) have defined a system as 'group of interconnected components with shared purpose that together achieve more than the sum of their parts' (p. 8). The authors note that an education system can be exemplified in the micro – as in classrooms, or in the macro – as seen in a network of teachers or schools; within each of these and across each lie other systems. As a human-centred system, education does not lend itself to technical solutions that can be applied; elements are always changing, and interaction amongst the elements occurs with 'unpredictable and unintended consequences' (p. 8). As Davis & Sumara (2008) observed, education is transphenomenal, transdisciplinary, interdiscursive, and pragmatic. It emerges and adapts through interactions and thus is constantly in flux. One must simultaneously examine teacher education in its own right and pay attention to the conditions of its emergence – that is, examine teacher education's own particular coherence and specific rules of behaviour as well as examine the agents that come together and the context of their co-activities. Teacher education is fundamentally nested within other complex systems (e.g., political, welfare, health), and as Davis & Sumara (2008) identified, these complex systems are idiosyncratic, recursively elaborative, and ever divergent in possibilities (p. 42). Additionally, teacher education involves various disciplines across the hard and social sciences, which are sometimes seen as incompatible if not contradictory; complexity thinking provides a means around these apparent impasses – discourses are presented as complementary rather than oppositional, thus reflecting the noted duplexity. Having a way to frame a matter helps to channel change. We therefore respond to the call from others in the field to build on current work to address how to 'mesh the perspectives of complexity theory with the normative requirements found in teacher education' (Martin et al., 2019). As Hodgson et al. (2018) put forward, we take up principled normativity to guide actions through giving up a desire for judgment and certainty and moving away from procedural normativity. While there is no perfect frame to provide guidance or explanation, it is essential to have a compelling alternative to offer in this engagement of ideas. Thus, we put forward for consideration the Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education (Figure 10.2).

10.2 An Emerging Model

10.2.1 Key Assertions

We present this conceptual model to inform judgement-making in an attempt to tease out complexities, think about decisions in different ways, and reorientate teacher education for the uncertainty of future challenges (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2021). The proposed model is grounded on three key assertions about teacher education: first, teacher education is classed as a complex system as it is simultaneously transphenomenal, transdisciplinary, interdiscursive, and pragmatic; second, learning to teach is transformational, involving ever-evolving yet fundamental changes in nature, quality, or structure; third, the goal of teacher education is future-oriented towards the possible and ensuring conditions for thriving in the 'as-yet unimagined' (Davis & Sumara, 2008), yet resources are finite. We propose a model where those in teacher education can step outside limiting frames and consider efforts for high-quality provision within the

presupposition of social sustainability. Although in teacher education we are committed to the United Nations (UN) Sustainable Development Goal 4 (SDG 4) to prepare high-quality teachers who deliver an 'inclusive and quality education for all' (UN, 2022), in teacher education we seem to have not actually taken into account the 'sustainable' part of this goal – that is, 'meeting the needs of the present without compromising the ability of future generations to meet their own needs' (UN Brundtland Commission, 1987, p. 16). The findings from the phases of this project stimulated ruminations regarding how, in teacher education, do we prepare new teachers to meet the needs of the present, within the boundaries of our resources, without compromising the ability to educate teachers in and for the future?

This concurrent and convergent exploration (see Chapters 1–9) reveals two key notions about judging the practices of student teachers: fairness and complexity. The study also reveals the often reductive, and even illusory, methods taken to gain validity and reliability, some of which demonstrate little benefit or impact on outcomes. Consequently, we began to build an alternative approach to thinking about consistency in judgement-making in teacher education, established on these core concepts of complexity and fairness grounded in fundamental concepts of sustainability. In proposing the model, our purpose is not to explain every aspect of teacher education or the multitude of influencing factors that impact the judgements we make regarding teaching effectiveness. Rather, we wish to propose a conceptual tool that might help teacher educators, schools, practitioners, researchers, and policymakers better see the complexity of teacher education and need for just and fair practices in evaluating and conveying effectiveness of teaching in new and productive ways. Our aim in this report is to stimulate discussion about the value of the model in understanding affordances and constraints in the complex decisions made in provision of high-quality initial teacher education (ITE).

The model is a visual framework that uses two concentric radar charts to depict a doughnut shape which displays comparison of multiple features. The visual representation of the model was inspired by Raworth's (2017) economic sustainability compass, the essence of which focuses on

> a social foundation of well-being that no one should fall below, and an ecological ceiling of planetary pressure that we should not go beyond. It encapsulates the ideas of fairness, complexity, and dynamic adaptability. Between the two concentric circles lies a 'safe and just space for all'. (p. 11)

A sustainable approach is about coming into dynamic balance by eliminating shortfalls and overshoot at the same time. It was the bringing forward of fair and just practices by participants in the case studies (see Sections 5.5 and 6.5) that clearly reflected the social pillar of sustainable development. Additionally, the continued variability of participants' responses to the influence of context and continued pursuit of eliminating bias in judgement-making evident across all phases of this project spurred the restructure of viewing factors as a balance or scale (see Figure 10.1) to reflect more accurately the inherent complexity of judgement-making in an educational system (See Figure 10.2). We put forward essential principles to guide us towards a systems thinking perspective. The model acknowledges the complexity of

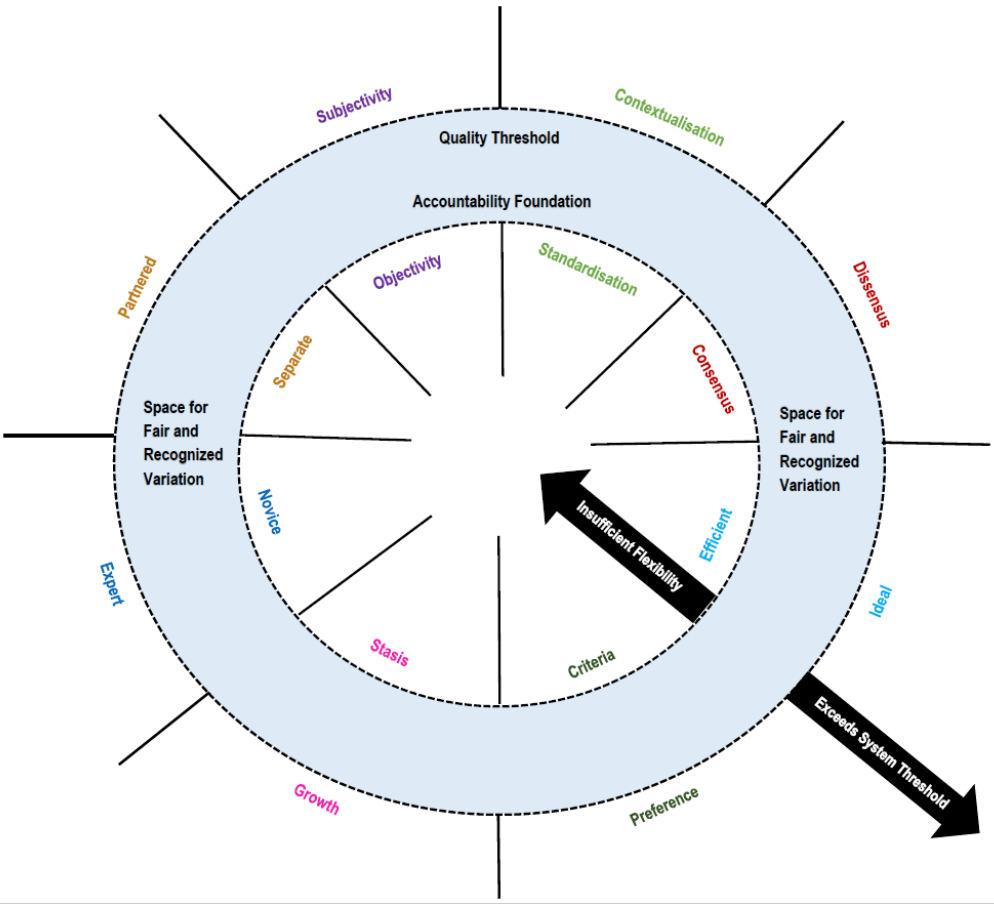
evaluating teaching and emphasizes the ongoing adjustments and adaptability required. It provides a way to take into account changing contexts and situations and the trade-offs we make when faced with complex decisions that forego clear solutions or one right answer. Essentially, conceptualizations of fairness and socially situated practices pointed towards an urgent need to approach teacher education in the context of SDG 4 in a way more commensurate with the definition of sustainability.

10.2.2 Concept 1: Eight Basic Factors

The eight basic factors reflected in the model are considerations for a starting point of aspects involved in judging teacher effectiveness within the provision of high-quality teacher education. In development of this model, we have focused closely on judgement-making and evaluation of new teachers' practices and considered the context, judgement-making ecologies, and processes from the case studies as presented. As an emerging model, we expect these to be refined with time, and they are likely interchangeable based on the decisions being made. However, the factors in Figure 10.2 do reflect characteristics of established best practices in teacher education (Australian Council for Educational Research [ACER], 2014; National Academy of Education, 2024).

Figure 10.2

Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education



Note. A full-page version of the model is included in Appendix A10.1.

Partnered – Separate

In the endeavour to best prepare teachers, we aim to work in partnered collaboration with stakeholders, in particular with schools and through meaningful classroom experiences. While it is important for teachers to understand themselves, individuals can also be narrowly self-interested. Additionally, TEPs need to know themselves well; often this occurs through the self-evaluation and continuous improvement processes of periodic review. However, we are social and reciprocating creatures. We are not isolated but interdependent, and we strive for partnerships which are mutually beneficial. There is a full continuum of partnership, from teacher education being totally isolated to programmes being fully embedded in schools. Partnerships are often identified through their mechanisms for offering feedback and dialogue. Partnerships for clinical experiences have been identified as one of the keys to high-quality teacher preparation (American Association of Colleges for Teacher Education, 2018).

Subjectivity – Objectivity

Objectivity and subjectivity are fundamental concepts in judgement-making, often perceived as opposing forces. However, they are interconnected and coexist within the teacher evaluation process. In teacher education, we strive for objectivity while acknowledging the influence of subjectivity. This balance is evident in our commitment to fair and unbiased assessments while also recognizing the role of personal perspectives and experiences in teaching and learning. We also acknowledge the elusive nature of qualities like responsibility, respect, integrity, and caring/humanity that student teachers are asked to exhibit and evaluators to identify. Conderman and Walker (2015) noted the high level of subjectivity attached to dispositions, which are not easily observable and quantifiable by nature, leading to inconsistencies where raters may not be seeing the same thing. By contextualizing evaluations and following fair processes that account for individual needs and equity, we demonstrate the importance of this balance. While objective criteria are ideal, subjectivism inevitably influences evaluations, as illustrated in the project's case studies.

Standardization – Contextualization

Standardization and contextualization are essential principles in evaluation of teaching. While standards provide a general framework for teacher performance, contextualization recognizes the unique circumstances of individual teaching contexts. Standardization ensures fair and consistent evaluation, regardless of school placement or classroom context. Contextualization acknowledges the variability of teaching challenges and opportunities. A balanced approach requires both standardized criteria and contextual consideration. By combining these principles, teacher education can provide fair, accurate, and informative judgements of teaching effectiveness. In Chapter 4 we explored the role of teaching standards in depth. The UNESCO *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019) acknowledges that, by necessity, standards involve general statements which broadly demarcate teachers' work and practices, and these should be made context specific.

Consensus – Dissensus

We see the need for consensus which fosters a sense of unity and cooperation, efficiency, and a common standard, and simultaneously we value how disagreement can spark ideas and creative solutions. As Moss and Schutz (2001) affirmed, dissensus is an essential natural resource that should be acknowledged and nurtured alongside the search for consensus or agreement. Across prior research (see Chapter 3) and the empirical components of this study (see Chapters 5–8), we see confirmation of needing both. What makes a degree of dissensus of value is in the promotion of critical thinking and ingenuity, the space for curiosity and innovation that multiple perspectives can bring. There is a continued dialogue about the necessity for consensus regarding readiness for teaching responsibilities. A key to balancing consensus and dissensus often lies in respectful dialogue and open-mindedness. Judging teaching effectiveness requires deliberating, thinking, and clarifying reasons for decisions. In a circumstance with no definitive right or wrong answer and with individuals having different value judgements, deliberation can be useful to air different perspectives and if not achieve consensus, then achieve a shared position – where all feel listened to.

Efficient – Ideal

TEPs often face the challenge of balancing efficiency with ideality. While efficiency can lead to cost-effective and time-efficient programmes, it may compromise certain aspects of an ideal programme, and perhaps even core values. For example, the Delphi panel (see Chapter 8) discussed in depth the implications of austerity and ramifications of the global teacher shortage (UNESCO, 2024). The reality is that shorter programmes, reduced coursework, or a focus on only practical aspects of the profession can save time and resources; however, this might compromise a strong foundation in educational theory and research as well as a deeper exploration of subjects and more in-depth practicums. Similarly, standardized curricula and assessments can streamline processes, yet a learner-centred approach affords tailored instruction and personalized support to better address individual needs and learning styles. Limited resources may necessitate trade-offs between programme components noted as essential for providing a high-quality teacher education (ACER, 2014; Cochran-Smith, 2021; Darling-Hammond et al., 2023; National Academy of Education, 2024). It is critical to also note that what is considered ideal or efficient is also influenced by what a society has resolved as the functions of teacher education – for example, as qualification, socialization, and/or subjectification (Biesta, 2015). Qualification, as Biesta defined it, refers to the preparation of individuals for their future role in society, with ITE equipping them with the knowledge, skills, and competencies necessary to participate effectively in the workforce and contribute to the economy. Socialization involves preparing new teachers in the way they internalize cultural and social expectations that shape their behaviour and identity. Additionally, subjectification involves the formation of personal identity, values, and beliefs and the ability to make informed choices and take responsibility for one's actions. While these three functions are distinct, they are also interconnected. For example, qualification can contribute to socialization by preparing individuals for specific roles within society, and subjectification can be influenced by both qualification and socialization as individuals develop their own identities and values within a particular social context. The extent to which

one of these is considered in context to be the ideal may help to determine the 'sweet spot' of acceptable variation. Ultimately, the goal of a TEP is to produce highly qualified teachers. Balancing efficiency with ideality requires careful consideration of these compromises and a commitment to providing ITE that meets the needs of both students and the profession.

Preference – Criteria

We can have an image of an ideal for teacher preparation and the most well-qualified and prepared new teacher that meets all our preferences and internal criteria. Yet we operate with limited and often diminishing resources in a complex system that can rarely guarantee a particular outcome amid multiple nodes of external accountability. We must therefore seek to balance the specified standards and expectations for teacher performance (see Chapter 4) with recognition of the importance of individual preferences and choices in teaching. Teaching does holds itself to account as a profession guided by ethical principles (General Teaching Council Scotland, 2012) and teacher education is subject to accountability measures (Cochran-Smith, 2021). When a profession demonstrates influence over others, it necessarily takes on ethical obligations and agrees to a social contract of just practices. Teachers are viewed both as public sector employees who are doing a job in the workforce and as experts in their profession who have valuable tacit knowledge about teaching and learning. As a profession, teaching commits to act in service, respect autonomy, be prudent, and work with humility (Raworth, 2017, p. 161). At the same time, the current social context demonstrates fluid values in place of fixed preferences, which contributes to intricacy of balancing partialities and external criteria. As Raworth (p. 97) reminded, Adam Smith identified an individual's self-interest and concern for others combined with their diverse talents, motivations, and preferences to produce complex moral characters, who show behaviours which could not easily be predicted. As we explored in the Delphi panel (see Chapter 8) what teacher education is actually accountable for is jointly determined. This reflects how criteria for accountability emerge from deliberations about commitments and goals as well as professional and national values.

Stasis – Growth

Additionally, education as a whole system is highly resistant to change and defaults to the status quo, yet teacher education aims for continuous improvement. Stasis might even seem counterintuitive in a field that is constantly evolving. Stasis provides a solid foundation, while growth ensures that teachers are adaptable, innovative, and committed to lifelong learning. In this study, we confirmed two principal uses of judgements of teaching effectiveness: for individual growth and professional development, and also to meet the need for evaluation measures in TEPs as a defender of quality (i.e., entry and exit of TEPs). Qualification (i.e., gatekeeping) is the most dominant reason judgements of teaching effectiveness are made; however, this appears often to be at the expense of other purposes (see Chapter 3). There was a tension between high-stakes consequential outcomes of judgements and educative uses of evaluation for growth. TEPs may be challenged to consider if knowledge, skills, and dispositions to teach should be precise, confined, and measured analytically according to operationalized indicators for a high degree of reliability, or if these

can be relatively broad, such as the holistic ability to gracefully teach increasingly diverse learners, as suggested by the Delphi panel (see Chapter 8). A balanced approach that combines both elements can help to produce highly skilled and effective educators.

Novice – Expert

The concept of novice and expert applies to all the stakeholders involved in the judgement-making process along the full continuum of the profession. The student teacher is most notably on their journey from novice to expert, expected to demonstrate an increasing level of competence and sophistication of practices throughout their preparation. In a postgraduate diploma programme, this is most often over the course of 1 year, in which an individual must demonstrate they are classroom-ready. What is expected to demonstrate competence and when is of course continuously in discussion (see Chapters 3–8). Additionally, mentor teachers in schools who partner with ITE for practicum experiences may be new to their role as a school-based mentor teacher, yet have much practical wisdom (Yacek & Jonas, 2023). They themselves are still developing their own practice – a process which Danielson (2007) has noted as taking about 5 years – while also developing their skills of mentorship. Finally, teacher educator is itself a distinct and critical role in ITE (Goodwin, 2012; Goodwin & Kosnik, 2013) with its own unique set of skills and knowledge. A highly accomplished teacher may excel in being a teacher yet still be developing their expertise about curriculum development, assessment, or instructional design to become an effective teacher educator. This development often occurs at the same time as entering full time into the university academy, which in itself can be difficult. The dynamic nature of education can leave many of these experts still feeling like novices at times. The field of education is constantly evolving, with new technologies, pedagogical approaches, and curriculum standards emerging regularly. Even experienced educators may find themselves learning new skills or adapting their practices to meet the changing needs of students and society.

8.2.2.1 Eight Interconnected Facets

These core elements are interconnected and interdependent, and they work together to create a dynamic and adaptive system that supports the development of effective teachers. Therefore, while the model reflects separated factors, these may actually be overlapping and are not necessarily equally distributed around the graph in terms of importance. A system may have an overemphasis on efficiency, for example, which takes up a quarter of the chart; an amplification of any one factor will necessarily reduce the rest, although not necessarily proportionally. As an illustration, if cost-efficient operations and ‘value for money’ constitute 25% of the factors considered, it may reduce or squeeze out a partnered approach or mentorship support for new teachers, therefore changing the shape of the model. Ultimately these eight factors are considered in view of the goal of preparing the teachers we need for the education we want.

10.2.3 Concept 2: Foundation and Threshold

The concentric circles in Figure 10.2 depict a doughnut shape which displays comparison of the eight basic factors. The circles are bounded by the foundation and threshold which set a

minimum quality standard for acceptable practices and outcomes, while acknowledging the need for flexibility within the system. There might be consequences for falling below or exceeding certain thresholds. At the edge of these circles are the permeable boundaries which represent the accountability foundation and the quality threshold, which in the model set a minimum quality standard. Too close to the centre or within and we see an excessive degree of process standardization that moves towards the dogmatic; processes become ridged, simplified, narrowed, and a default perhaps to what has been known. For example, a strict, word-for-word application of professional teaching standards without considering the degree of sophistication expected or school/community/classroom factors (e.g., deprivation, denominational, additional support for learning) would demonstrate insufficient flexibility (for more on professional standards, see Chapter 4). In this space below the foundation, there may be reduced variability, cost reduction, a perceived better quality, and consistency; however, a shortfall of the inner circle likely indicates insufficient flexibility to consider factors of complexity. Flexibility allows processes to respond to specific demands, to be agile, and to use human judgement-making well. The accountability foundation therefore sets a minimum acceptable level of consistency a TEP or educational system finds acceptable. This is likely to vary depending on which system one is in (e.g., Scotland, England, or Wales). Beyond the outer ring – the quality threshold – lies an unacceptable degree of inconsistency. Variability is part of human nature, yet large degrees of variability can lead to incorrect conclusions or, as participants in our study noted, the standard and reputation of the teaching profession and the education of children and young people being at risk (see Tables 5.14 and 6.14).

10.2.4 Concept 3: The Inner Ring

Between the two sets of boundaries lies the optimal place for action, or the ‘sweet spot’. We have termed called this ‘the space for fair and recognized variation’. As findings from the case studies emphasized, there is not necessarily one way to judge, therefore some degree of disagreement is inherent and acceptable; however, a baseline is necessary and a way to apply idea of limits is needed. This need for both prompted the notion to design the model to allow for responsiveness to differences so long as the standard for accountability as defined by stakeholders in the system continues to be met and the threshold of adequate variability has not been exceeded. The inner ring acknowledges that while there will never be complete agreement (which may not even be desired), there is an acceptable degree of inconsistency the system or TEP can tolerate. We don’t necessarily need to have perfect rater agreement or attain a particular statistical validity score for decisions of professional judgement to be deemed dependable (see Chapter 3). In addition, there is space at the edge, the liminal space, for emergent practices, innovation, dialogue and debate, trial and error. It is often at the edge of the boundary that a sense of possibility and transformation occurs, as the Delphi panel explored (see Chapter 8).

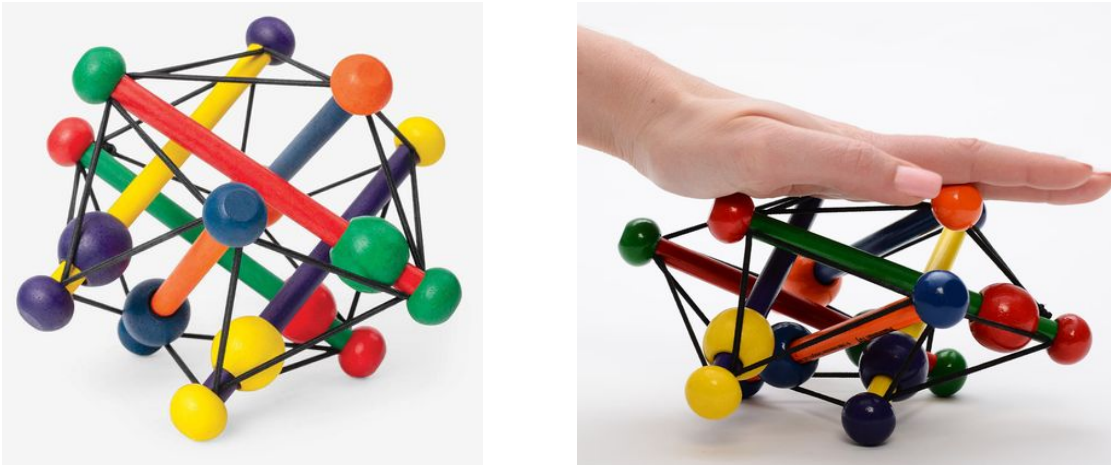
10.2.5 A Novel Frame: Teacher Tensegrity

While the Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education is put forward here as a two-dimensional concentric radar chart, the way these features, or

nodes, relate and interact is more accurately represented in the four-dimensional. This is more closely reflected to the image of the Manhattan Skwish toy shown in Figure 10.3. The Manhattan toy is actually sold as a 'powerful tool for play' for children; we value how it gives rise to 'playing' with ideas and imagining interactions in complex systems. Each component of the model is in a different degree of tension or flux in connection with another, which a chart designed for displaying multivariate data cannot fully describe; in this emergent model, all the features will not necessarily have an equal amount of importance. And it is often the dynamic spaces in-between these nodes that the degree of fair judgement-making occurs, and subjectivism that can impact professional judgement can be dealt with in way that protects boundaries of competency necessary for safeguarding pupil learning.

Figure 10.3

Manhattan Skwish Toy



Note. Images of the Manhattan Skwish toy captured from www.jojomamanbebe.co.uk

Thinking of the complex influences on judging teaching effectiveness in this way offers a structural principle of tensegrity, a novel way to reflect on our judgement-making processes in teacher education. What may contribute to new teacher's being able to demonstrate 'the what' of teaching might be what we will term 'teacher tensegrity'. Tensegrity is a structural principle where a structure is held together by a network of continuous tension elements (like cables or strings) that offset discontinuous compression elements (like struts or poles). The concept itself incorporates the tension and release of strength in building for architectural purposes, but the principles of this concept are first present in nature, reflecting balance, efficiency, and adaptability (Ingber, 1998). These principles are the reason that objects can absorb the impact of a hard landings. We would like to think that our teacher education system can be similarly resilient to handle 'hard landings'. Application of this principle in the field of education was recently made by Berise Heasly (2021), who coined the term 'Edu-tensegrity' to emphasize the use of trilectic, democratic thinking skills in classroom settings and encouraged the blurring of rigid boundaries (pp. 76–77) given the varied elements in the whole world of education involved in the process of decision-making. Heasly brings the

concepts of resilience, sustainability, and securitability broadly into the language of education and describes the flexible response to chaos as a mechanism toward gaining justice.

We found this concept of tensegrity and Edu-tensegrity a fascinating exploration of complexity and meaningful for us to take up in the context of teacher education. We were thus prompted to consider assessment of pre-service teachers' skills in their TEPs through the lens of six foundational principles that define the unique properties of tensegrity.

The first principle is that of continuous tension, a foundational network (e.g., the cables in Figure 10.3) that maintains a structural form, contributes to stability, and allows for even distribution of stresses. In teacher education, we maintain structured, partnered, programmes that operate within the relatively stable university structure, programmatic provision, quality assurance process, and funding model.

The second principle is discontinuous compression; in the physical model, this refers to the compression components, such as struts or rods, which are distinct in that they do not make direct contact with each other but are instead suspended within the tension network, removing the need for rigid connections. In teacher education, connections such as practicum experiences in multiple schools and a spiralled curriculum provide structured opportunities to evaluate pre-service teachers' skills and knowledge.

The next principle is pre-stressed tension, where the tension elements are initially stretched or tightened; this pre-tensioning creates a built-in stress within the structure, making it more resistant to external forces. In teacher education, the programme's foundation is built on pre-existing knowledge and skills and a foundation in a teacher's identity. TEPs often assume a certain level of academic preparation and personal qualities in their applicants; these are often defined in entry requirements and influence professional growth. These are frequently founded on predictive validity.

The fourth principle is self-equilibration; tensegrity structures automatically distribute internal stresses across the structure, which allows them to adapt to varying loads without losing structural integrity. In teacher education, evaluations should be balanced and fair and adapt to individual needs and strengths. A well-designed assessment system should consider the diverse backgrounds and learning styles of pre-service teachers and take context into consideration.

The fifth principle is that of minimalism and efficiency – that is, utilizing the minimum amount of materials to achieve maximum structural strength. Judgements of new teachers' practices should be focused and efficient, avoiding unnecessary complexity. Evaluations should be designed to measure the most essential competencies and skills required for effective teaching.

Finally, the sixth principle is scalability and modularity; tensegrity structures can be easily adapted or expanded in size and complexity according to specific requirements. In a similar way, a robust and effective evaluation process should be adaptable to different programme sizes and needs. Evaluation components should be modularized, allowing for flexibility and customization. In these ways, the tensegrity framework provides a useful lens for

understanding the sophisticated work of a teacher, the complexity in which it occurs, and the need for both flexibility and fairness in judging pre-service teachers' practices to create a robust and effective assessment system. Needed is an orientation towards strength and flexibility that tensegrity exhibits.

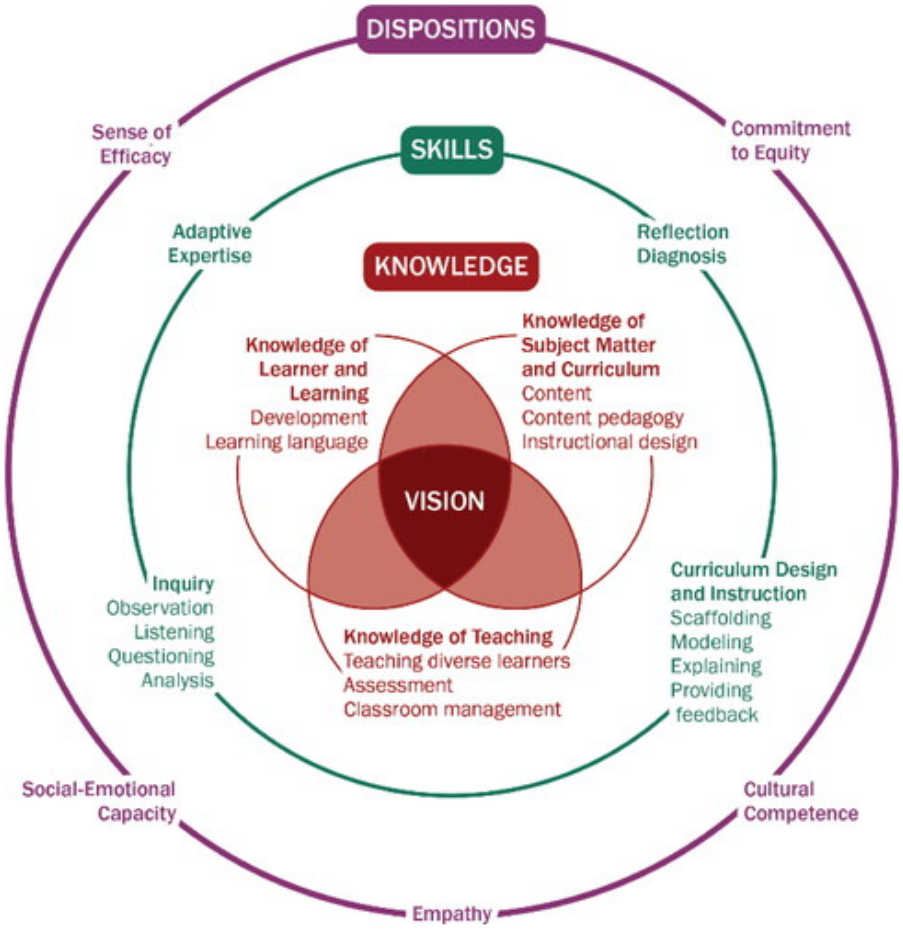
10.2.6 A Model within a Context for the 'What'

This conceptual model does not exist on its own; as noted in the theoretical framework of SJT, it is socially situated. The model is nested within multiple complex social contexts (in our study, the three nations of the UK) and ecological systems (Bronfenbrenner, 1979). The judgements of new teachers' practices are impacted by characteristics of the individual and relationships with family, peers, community, and school (microsystem), interactions and relationships between these individuals and groups (mesosystem), societal values, educational reforms, funding, and research (exosystem), and educational policies and governing legislation, national standards, and cultural diversity (macrosystem). All interactions are influenced by historical events, societal shifts, technological advances, personal experiences, and personal growth (chronosystem). *Teacher tensegrity* is a concept reflecting the ways this complexity could be responded to in sustainable ways. Teacher education, then, is charged with equipping educators with the skills to adapt to a constantly evolving educational landscape. It is within this thoroughly complex space that fair judgements of new teachers enacting the 'what' of teacher education are being made (see Figure 10.4). Clear characteristics of complexity are evident in the continued pursuit to define the competencies an effective teacher should demonstrate. Darling-Hammond et al. (2023) have engaged in expansive research on the fundamental concepts that educators must understand regarding learning and teaching (Cantor et al., 2018; Osher et al., 2018; Darling-Hammond et al., 2019), describing these as deeply embedded in sociocultural contexts.

To this end, Darling-Hammond et al. (2023) noted several challenges the teacher education must take into account for teacher candidates to learn to teach: (a) to frame their own prior experiences to address what Lortie (1975) termed 'the apprenticeship of observation'; (b) to not only learn to think like a teacher but also act like a teacher (Kennedy, 1999) – this is to not only learn what to do but be able to actually do it; and (c) to address the 'problem of complexity' (Darling-Hammond & Bransford, 2005). The layers of accumulated complexity are the very space where fair and accurate judgements of teaching effectiveness are attempting to be made.

Figure 10.4

The 'What' of Teacher Education



Note. Darling-Hammond et al. (2023, p. 4).

10.3 A Hypothetical Application in Teacher Education

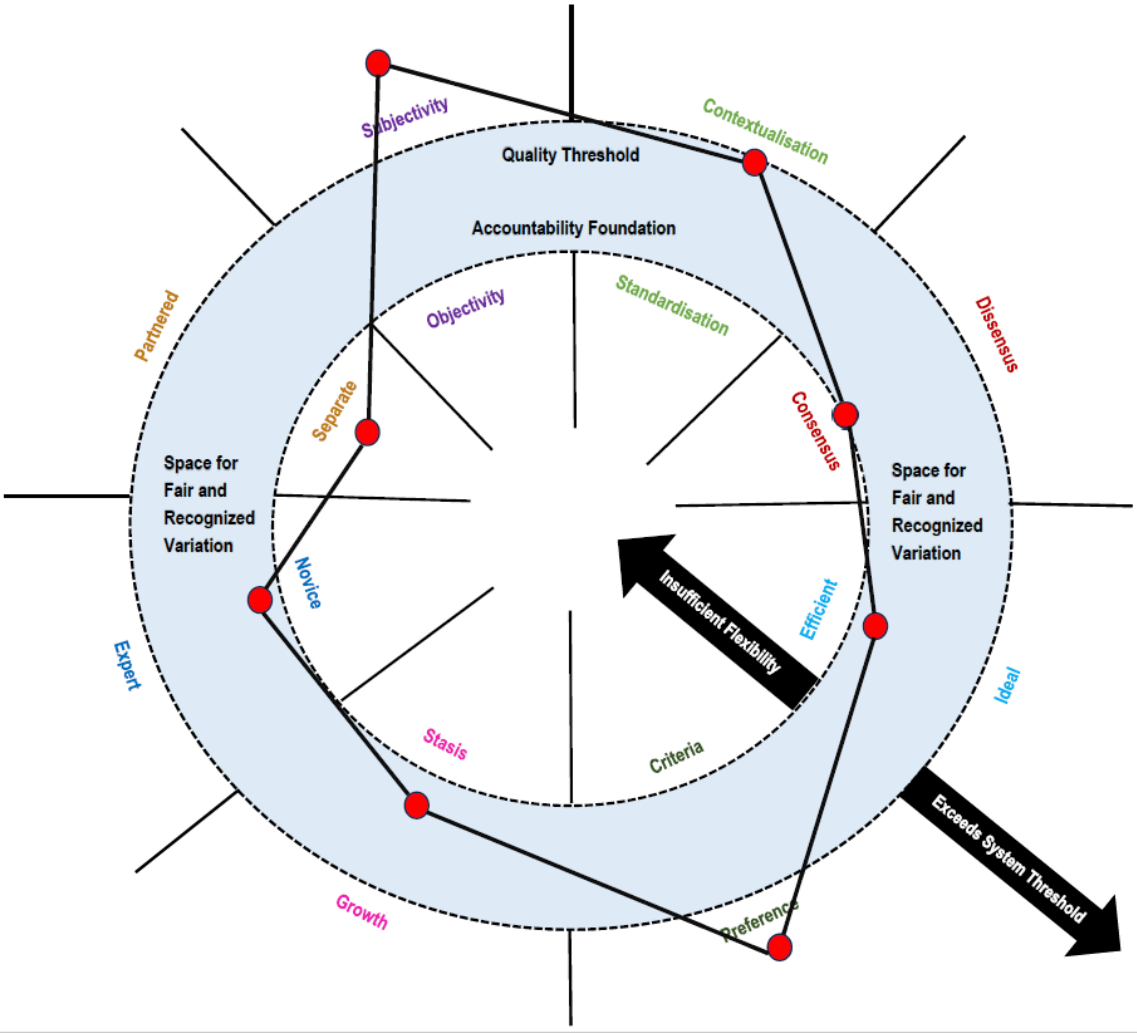
We now provide a hypothetical example of how the Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education can be applied within a TEP. The purpose of this example is to illustrate the interrelationships among principal ideas composed of multiple parts, oftentimes operating together, and the ongoing tensions and adaptability needed to take into account changing contexts and situations to support judgement-making. The example is based on a real ITE programme and involves practices of TEPs when evaluating student teachers during school experiences, an element of preparation common across programmes in the UK and beyond. A hypothetical example is used in order to openly explore implications of potential shortfalls and overshoots; the model could be utilized in a similar way to explore the case studies involved in this project.

Findings from the multiple phases of this project examined the complexities of teaching effectiveness judgements from several perspectives. Our analysis reveals the multifaceted nature of this process as experienced by university-based teacher educators, school experience tutors, and school-based mentor teachers, and it highlights the complexity of the

task behind what may seem to be simple ratings of teaching effectiveness. Results illuminate indicators of complexity which define the nature of shared judgement and challenge consistency and reliability. We therefore chose to trial application of the model in the current evaluation process used in judging student teachers' demonstration of professional teaching standards at one anonymous university-based provider of ITE (see Figure 10.5). The simulated application of the model was carried out by the Principal Investigator and a second project member familiar with the programme and its context; we utilized internal documents and website information to conduct the application of the model using the eight basic factors.

Figure 10.5

Example Application of the Duplexity Model to Judgement-Making Processes In ITE



Partnered – Separate

According to written processes in handbooks and role remits, the process of making a judgement of a student teacher's teaching was to occur jointly with an observation evaluation agreed by the mentor teacher and the school experience tutor. In reality, the judgement was being made by the tutor, often with the TEP's evaluation form filled out before even arriving at the school for an observation. The form was sent to the mentor teacher via email to review, add to, and edit as they saw fit. Due to significant time pressures of classroom teachers, many

were happy to follow the judgement of the tutor. In this way, processes were not followed, compromising the fairness of the evaluation; the judgement was not a partnered effort.

Subjectivity – Objectivity

The mentor teacher and tutors all articulated a commitment to unbiased assessment practices in their agreement to the role. Additionally, both groups of individuals attended training sessions regarding the evaluation process in order to implement these consistently. However, the observation proforma included all of the professional teaching standards, which included some values-based qualities such as responsibility, respect, integrity, and social justice. There is a noted high level of subjectivity attached to evaluating dispositions (Conderman & Walker, 2015), which are by nature not easily observable and quantifiable. Additionally, the proforma included tick boxes with binary options (e.g., pass or fail) and space for qualitative comments. No actionable descriptors or clear expectations of what an evaluator would be looking for were included. Additionally, when sampling of evaluators comments were reviewed, these were found to reflect much personal judgement and a high degree of variability. Not all judgement decisions were supported.

Standardization – Contextualization

Observations at this TEP are carried out to judge the student teachers' effectiveness according to the entire set of professional teaching standards. These are noted as holistic aspects and not descriptive practices. While the process includes clear protocol, due to the size of the programme and the number of individuals involved, there appears to be a great deal of variation in how the processes are actually carried out in practice, therefore a standardized process cannot clearly be guaranteed. The level of contextualization could be quite high and may be nearing the point where there is too much variation to be categorized as a fair or consistent process.

Consensus – Dissensus

There is a very high degree of consensus evident, however as noted above in relation to a partnered approach, this could be due to the fact the tutors are making the judgements on their own, so there is no opportunity for dissensus. In this factor, the consensus gives warrant to consider if space for curiosity and innovation that multiple perspectives can bring is missing. In the absence of dialogue to make a joint judgement, there is perhaps little opportunity for balancing consensus and dissensus, which often lies in respectful dialogue.

Efficient – Ideal

In an ideal approach to judgement-making, the ITE provider would prefer to include pre- and post-observation meetings with the student teacher and their mentor, examine multiples sources of evidence, speak with pupils who have been taught, conduct additional formative observation visits, and participate in joint mentor training and evaluation calibration exercises to develop judgement-making skills. In reality, there are separate 1-hour online training meetings and one formative assessed visit and one summative assessed observation. As noted, often the evaluation is completed by the tutor. Additionally, part-time tutors are

conducting the observations instead of full-time ITE staff, and teachers are given little time off timetable to dedicate to their role as mentor. This leaves the programme operating at the boundary of efficiency instead of exhibiting best practices in school-based experiences, and it may need to reexamine the trade-offs that could be made to bring processes back into a more fair space.

Preference – Criteria

In this particular case, the format of the proforma seems to have brought in a stronger influence of the personal preferences of the tutor and classroom teacher, which seems to outweigh the formal criteria for judgement (i.e., teaching standards). This was evident in the qualitative comments. It is unclear if an assurance of quality teaching could be made given the level of personal judgements instead of professional judgements being made.

Stasis – Growth

The formative observation visit and personalized growth plans to work on areas needing improvement are indicators of a growth-oriented system. It is unclear how growth is perceived, however, as the programme protocols indicate that all the standards must be demonstrated at a satisfactory level halfway through the programme. This seems to be an indicator of the tension between high-stakes consequential outcomes of judgements and educative uses of evaluation for growth, as all standards must be passed in order to complete the programme and earn a qualification.

Novice – Expert

While the student teachers themselves are considered novices in the context of their initial preparation, both mentor teachers and teacher educators are considered experts. They themselves are still developing their own practice, a process which, as mentioned earlier, Danielson (2007) suggested takes about 5 years, while also developing their skills of mentorship. There is no clear indicator of how much classroom experience the mentor teachers have, and rarely do they, or the teacher educations, have any formal preparation for their roles; in this way they are both experts and novices.

Based on the application of the model, it is evident that subjective factors and individual preferences have likely surpassed the threshold necessary to consistently ensure a specific quality. Furthermore, several factors are situated near the boundary of the inner circle, indicating that there may be insufficient flexibility within the system. While a majority of factors fall within the acceptable range for just and fair variation, the programme should consider strategies to address instances of overshoot and identify areas where flexibility can be increased. The simulated application of this conceptual model offers valuable insights into its functionality. It has highlighted specific areas where the teacher preparation programme may need to reconsider and adjust its processes, particularly regarding the observation proforma. This model enables experimentation with various variables and parameters, allowing for an examination of their potential impact on outcomes and the ability to observe complex interactions.

10.4 Implications for a Change in Thinking

We can listen to what this dynamic system of judgement-making in teacher preparation tells us and explore how its characteristics and our values can work together to bring forward something better. Perhaps the space for fair and recognized variation reflects what is desired in a new social contract with and for teacher education (UNESCO, 2021). In response to the outcomes of this research, we therefore put forward six principles for embracing complexity and draw on insights of sustainability for a transformative approach to teacher education. The principles are:

1. Change the aim
2. Context matters
3. Nurture human nature
4. Systems thinking
5. Intergenerationality
6. Focus on the right kind of growth

First, it is necessary to change the aim. Teacher education is not primarily for a process of social efficiency (Schiro, 2013) in which a factory model of teacher replication is desired to meet the needs of the job market and the economy. If entry and exit requirements for teacher education are found to keep the demographic of teachers homogenous, a process that better helps to reflect the full diversity of learners in their teachers is desirous. Needed is a pivot from seeing teachers as a standardized product to seeing them as sophisticated individuals focused on enabling human beings, including themselves, to thrive. As Biesta (2020) observed, effectiveness is considered a process value, and effective 'for what' and 'for whom' should be a consideration of TEPs in the exploration of judging effectiveness. The challenge now is to agree an acceptable foundation and threshold quality indicators which can thrive in balance. There are no prescriptions. The right aim can assist in allocating finite resources and funding enhancement in education by knowing better where to invest first based on shortfalls and overshoots.

Second, the catchphrase 'context matters' is ubiquitous in the field. But what actions would we see that would demonstrate this is true in a teacher's practice? We are challenged in the in-between spaces of variation and liminality to clarify the way a teacher enacts a response to the infinite combinations that amalgamate to structure a classroom, school, authority, or national education context. All communities involve nested economic, environmental, and social systems, so it becomes important to understand the interconnections. Educators are consistently challenged to consider *how much* and *in what ways* context matters. How does a new teacher demonstrate application of understanding Bronfenbrenner's (1979) ecological systems theory? The skills to teach in response to a particular context are perhaps the competencies, what the Delphi panel members suggested as the 'it factor', that we are really looking for in judging effectiveness and classroom readiness. These just may be the skills of the in-between; the skills required to take theory and actually *enact* the application in the practicalities of practice are different than competencies themselves.

Third, nurture human nature. As Professor of Environmental Biology Robbin Wall Kimmerer (2015) put it: 'Isn't this the purpose of education, to learn the nature of your own gifts and how to use them for good in the world?' (p. 239). In a change from efforts to eliminate variability and differences, we shift focus to cultivating future educators who are not only knowledgeable but also compassionate and ethical, and who promote human flourishing in their classrooms and communities. We likewise desire for teacher educators to do the same. We expect to see this evidenced within teacher preparation, in particular through assessment measures focused on growth and development. A focus on a teacher's ability to such things as enlargening people's capacities would require a different way of judging effectiveness. Importantly, Valentine et al. (2021) argued: 'Changing focus to look at what is "fair" human judgement in assessment, rather than what is "objective" human judgement in assessment allows for the embracing of many different perspectives and allows for the legitimizing of human judgement in assessment' (p. 2). In their work, the authors proposed that fair judgement decisions are transparent, credible, fit for purpose, defensible, and supported by individual (e.g., evidence, boundaries, agility, expertise, narrative) and system (e.g., procedural fairness, documentation, multiple opportunities, multiple assessors, validity evidence) factors.

Fourth, we have attempted to elucidate intricacies involved in rendering sound judgements of teaching effectiveness and to explore how systems thinking can contribute to considering equitable degrees of variation. While education is often characterized as a system, this perspective is seldom incorporated into decision-making and design processes. This prompts a shift from a narrative of linear functionality and efficient performance to one of complexity. Systemic thinking entails deconstructing and reconstructing processes, examining causal relationships and mechanisms, exploring the interplay of structural forces and human agency, and endeavouring to explain diverse dynamics of learning (Bermudez, 2015). Systems thinking encourages us to pose broad questions such as why did this occur, how does this function, has this always been the case, how do these elements interconnect, and who benefits and who suffers. Adopting a complexity perspective challenges short-sightedness and fragmentation, acknowledging the unequal distribution of costs and benefits in conflict resolution. Systems thinking and forecasting both necessitate a specific skill set as teachers employ this approach themselves and guide their pupils to do so as well. Bermudez (2015) identified several cognitive tools essential for a systems thinker: interconnectedness rather than disconnection; circularity instead of linearity; emergence in place of silos; wholes rather than parts; synthesis as an alternative to analysis; and relationships rather than isolation. Systems thinking also involves the judicious use of feedback loops (Bermudez, 2015; Raworth, 2017) as we learn from the outcomes of each action taken. When the output of a process is employed as input to the same process, it generates a positive or negative cycle of cause and effect, influencing the overall behaviour of a system. Establishing feedback loops between ITE providers and schools is paramount. Despite some robust partnerships, TEPs are often slow to adapt their programmes to the specific needs of schools. Throughout the system, there are few formalized structures requiring universities to respond to feedback from schools and vice versa. The ITE programmes analysed in Chapters 5–7 exemplify the necessity and potential benefits of continuing to develop strong feedback loops that inform

and are informed by research, context-specific practices, and professional judgement. Such feedback loops would ensure the integration of teacher candidates' professional experiences with their teacher education coursework and future professional responsibilities.

Fifth, in teacher education we need to consider the intergenerationality of teacher preparation and long-term effects if quality is compromised. Underprepared or unsupported new teachers may become less effective mentors for future educators, creating a negative feedback loop. Non-university-based TEPs with less rigorous standards can also contribute to this issue. Given the global teacher shortage (UNESCO, 2024), the immediate and long-term consequences of such policies are particularly concerning. Recent examples like Glasgow City Council's proposed plans to cut 450 teachers over the next 3 years (McCool, 2024) highlight the potentially detrimental effects of teacher shortages on pupil outcomes, particularly for vulnerable populations. These shortages can lead to increased workloads, difficulty recruiting diverse educators, and higher turnover rates, all of which can disrupt pupil learning and create instability in classrooms. Moreover, a decline in the teaching profession due to factors like instability or underfunding can create a self-perpetuating cycle. Fewer qualified teachers may lead to decreased funding for TEPs, further impacting their ability to provide quality instruction. This can result in a shortage of faculty members to teach aspiring teachers, creating a vicious cycle. To address this issue, we must adopt a more sustainable approach to teacher education. This includes investing in ongoing professional development, mentoring, and opportunities for teachers to transition into new roles. An adequate partnered response requires full participation of all the actors to prevent a cascading effect. We require a change in thinking that again gains purchase on concepts from sustainability. By creating a more supportive and rewarding environment for educators, we need to work to ensure a high-quality teaching profession for generations to come.

Finally, focus on the right kind of growth. This includes growth in pupil learning and development and continuous refinement and sophistication of professional knowledge and practice of teachers and teacher educators. This does not include value-added measures or exam scores, and it does not include one-word ratings of effectiveness. In terms of reliability and consistency, we have been fixated on what is easier to measure, quantify, and compare (Biesta, 2020) rather than growth in human capacities and prosperity. For teacher education, it does not necessarily entail common measures of success in the higher education sector (Hubball & Dawson, 2014), such as increasing enrolment, external funding targets, national and international rankings, or Research Excellence Framework (REF) outputs. We desire TEPs that make all involved thrive, whether or not we 'grow'. A more appropriate measure could be the impact cases of REF, which involve the impact academic research has on society (UK Research and Innovation, 2023) with indicators focused on growth and change. Teacher educators' contribution towards the 'impact' dimension of research can be substantial because of the location of their research, within schools and the public, and with many direct links to and influence on public policy and practice. In congruence with a revised view on determining growth, we may need to reconsider how we deploy our universal human resources of time, knowledge, skill, care, empathy, teaching, and reciprocity (Raworth, 2017).

Systems can't be controlled, but they can be designed and redesigned. This model and thinking promise no immediate answer for what to do next, but these concepts are fundamental to a different way of thinking about teacher preparation than our current circumstances demand. Overall, our Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education proposes a more nuanced way to enact teacher education, influence how we evaluate teachers, and consider a wide variety of factors, thus allowing for adaptability while still maintaining accountability. It offers a framework for a more comprehensive and evolving approach to judgements of teaching effectiveness.

10.5 Conclusion

In this chapter, we have argued that linear ways of considering judgement-making have missed important considerations for quality as well as improvement of teacher preparation as part of an incredibly complex education system. The conceptual model presented is an attempt to tease out some of these complexities and to suggest that judgement-making necessitates an adaptive and duplexed approach which requires a change in thinking about several educational concepts. We have also introduced the term 'teacher tensegrity' to describe the tension, flexibility, and strength needed for effective teaching and for high-quality teacher preparation. Further, we have applied the conceptual model to examine the procedures and processes of judging teaching effectiveness in one ITE programme to explore potential shortfalls or overshoots of the foundation and threshold boundaries. For teacher education to thrive, we suggest the need to embrace ambiguity rather than avoid it and to get ahead of the curve of uncertainty before damage is done to pupils, schools, and teachers. We need to not simply react to what is imposed on us. This argument has implications for TEPs, partnering local authorities, and policymakers, which are presented in Chapter 11.

11 Conclusions and Recommendations

This research project, funded by the Society for Educational Studies, titled *Reliability and consistency in judging new teacher practices – why does it matter?*, sought to investigate the significance of reliability and consistency in judging new teacher practices within the context of initial teacher education (ITE). Employing social judgement theory (SJT) as a theoretical framework, the multi-phase study explored the nature of judgement-making processes, consensus formation, and power dynamics among university staff, associate tutors, and school-based mentor teachers. Through three key research questions, the project aimed to understand the shared judgements of teaching effectiveness, the potential for enhanced collaboration between schools and universities, and the influence of power dynamics on the roles of teacher educators in judging teaching effectiveness.

11.1 Why Does It Matter?

Case study findings in Phase 3 of the project underscore the importance of consistent and reliable judgements of teaching effectiveness in ensuring equitable treatment of all teacher candidates. Fairness, characterized by impartiality, justice, and equity, necessitates treating individuals without bias or favouritism. In the context of evaluating future teachers, this entails applying consistent and equitable standards across all candidates, using clear and measurable criteria, and assessing the actual skills and knowledge required for effective teaching. Ensuring equitable treatment involves avoiding personal biases, providing equal opportunities for demonstration of abilities, and addressing potential biases in evaluation methods or materials. To maintain fairness, transparency and communication are essential, including clear communication of evaluation criteria, constructive feedback, and a transparent process for challenging decisions. By adhering to these principles, judgement-making processes can better ensure that student teachers have a fair chance to demonstrate their readiness for the classroom.

The project findings highlighted several reasons why consistent and reliable judgements of teaching effectiveness matter. First is *fairness for teachers*; inconsistent judgements can be arbitrary and lead to some teachers being unfairly disadvantaged in areas like career progression or receiving support. Next is *reliability of selection*. If judgements are not reliable it is difficult to recognize the future teachers who will demonstrate effectiveness within the profession. Ultimately, the goal of effective teaching is *pupil learning*; consistent judgements can help ensure teachers are developing the knowledge, skills, and dispositions that benefit pupils. Every child and young person deserves a rich learning experience. Furthermore, reliable judgements can provide a clear picture of strengths and weaknesses, facilitating targeted support and continued *professional development* for new teachers. Consistent judgements can also help bring *credibility to the profession* through clearer standards and expectations. The integrity and credibility of the teaching profession hinges on the exercise of professional judgement guided by meaningful sources of evidence. As such, ongoing reflection and refinement of judgement practices are essential to ensure the fidelity of teacher preparation.

Strategic and systemic changes to our judgement-making practices for determining teaching readiness must be student-centred, collaborative, equitable, data-driven, and future-focused. As demonstrated by the complexity of defining and judging teaching effectiveness, which is influenced by dynamic interactions between various factors, the teacher education system necessitates a reciprocal interplay among stakeholders. To implement effective solutions, we must focus on underlying rationales and reasoning strategies. Student-centred values, while important, do not always align with the best interests of teacher candidates. Therefore, a redesign of judgement-making processes is necessary to maintain effective elements and disrupt ineffective ones, ultimately improving teacher preparation. This can only be achieved through collaboration among stakeholders to ensure equity, dependability, and stronger consequential validity in shared judgement-making. By conducting ongoing research and analysing actionable findings, we can better understand the root causes of consensus, dissensus, and power dynamics, enabling us to maintain a future-oriented approach. The status quo is comfortable, and the more complex the issues the more likely it will receive pushback; we hope to remain future orientated on contributing to the 'life-entangled journey of teacher development' (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2021, p. 84).

Overall, we have concluded that transformation of assessment of teaching effectiveness is important for several reasons. The experiences of student teachers during their mentorship period significantly influence their future teaching practices and self-perception. The established expectations and feedback they receive can shape their resiliency, their ability to cope with challenges, and their overall commitment to the teaching profession. These experiences contribute to the institutional memory of teacher education programmes (TEPs), influencing the generational experiences of subsequent cohorts. Thus, the quality of mentorship and assessment during this formative period is not merely a matter of ensuring competent teachers but also a critical factor in shaping the long-term trajectory of the teaching profession.

11.2 Recommendations

Taken together, findings from this project support several recommendations for improving the judgement-making process of teaching effectiveness for TEPs, school partners, and policymakers. These recommendations are predicated on the principles of Sustainable Development Goal 4 that high-quality teaching, for all students, in all circumstances is a right (United Nations, 2022).

11.2.1 Teacher Education Programmes

Programmatic Revision:

1. Examine entrance requirements and evaluation processes to ensure they do not narrow the talent pool.
2. Eliminate ineffective programmatic requirements in ITE that do not demonstrate predictive validity of a positive impact on pupil learning and development.

3. Revise ITE structure and curriculum with a focus on creating opportunities and learning experiences in which future teachers develop skills needed to deal with complexity and uncertainty and to translate theory into their practice.
4. Prepare student teachers for systems thinking through using systems thinking.
5. Explore opportunities to expand the amount of time prospective educators spend in clinical experiences in which future teachers can sustain relationships needed to develop the sophisticated skill set required for effective teaching in increasingly complex classrooms (e.g., multi-year residencies, 1-year mentored residencies).
6. Conduct a collaborative research study to examine the effectiveness of the 1-year Postgraduate Diploma in Education/Postgraduate Certificate in Education model of teacher preparation.

Evaluating Teaching Effectiveness:

7. Standards should be calibrated to better reflect different levels of experience rather than a one-size-fits-all approach across the continuum of the professional teaching career.
8. Develop judgement processes with explicit performance expectations (e.g., look fors).
9. Create opportunities for developing clinical judgement skills.
10. Adopt, revise, or create evaluation measures of teaching effectiveness that better address the complexities of teaching and allow for a fair degree of dissensus.

Teacher Educator and Mentor Teacher Development:

11. Expand professional development opportunities for teacher educators and mentor teachers.
12. Create a diploma, certificate, or endorsement for teacher educators and for mentor teachers.
13. Emphasize the role of mentor teachers in their subject area expertise, as historians, artists, mathematicians, etc.
14. Create a specialized TEP advisory board focused on clinical partnerships and practice.

Partnerships and Collaboration:

15. Form Research Practice Partnerships where researchers and practitioners work together to address educational challenges and improve student outcomes.
16. Jointly make placement decisions.
17. Only place teacher candidates with mentor teachers who demonstrate high-quality instruction.
18. Partner with schools to identify preparation gaps and opportunities.
19. Develop a comprehensive probation process for new teachers that involves TEPs.
20. Reduce bureaucratic workload for all involved in partnership to prepare teachers.

21. Explore team teaching/co-teaching models involving school-based mentor teachers and university-based teacher educators.
22. Gather actionable feedback from ITE graduates and mentor teachers to inform programme improvements.
23. Use systems thinking to incorporate feedback loops into teacher education.
24. Collaborate with other TEPs nationally and internationally to inform continuous improvement efforts.

11.2.2 School Partners

Effective Mentoring and Instruction

1. Place future teachers with school-based mentor teachers who have demonstrated exceptional teaching practices and are committed to working with teacher candidates.
2. Ensure schools where teacher candidates are placed provide high-quality, research-based instruction, effective social emotional learning, and evidence-based interventions to address the needs of all pupils, including those at risk.
3. Employ a university-based teacher in residence to facilitate collaborative approaches and discussions, potentially considering residency models.
4. Pair newer school-based mentor teachers with more experienced colleagues and offer differentiated programming and support for both groups.

Strategic Planning and Collaboration

5. Ensure the school's overall strategic plan includes a personnel strategy and specific goals for talent management and the development of teacher candidates, aligned with the school's mission, vision, and strategy.
6. Explore flexible or non-traditional work arrangements for school-based and university-based teacher educators to enhance collaboration and efficiency.
7. Identify and celebrate highly effective student teachers and teachers, and provide them with opportunities to serve as educational ambassadors for the profession.

11.2.3 Policymakers

Enhancing Professional Standards and Support

1. Ensure professional standards for educators, including those related to teaching and headship, are clear, accessible, and applicable to diverse teaching contexts. Include specific responsibilities for mentoring and educating future teachers.
2. Provide fair compensation to teachers who serve as mentor teachers during ITE preparation experiences.
3. Adjust funding models to provide fair compensation to TEPs for the actual costs associated with strong clinical experiences.

4. Promote innovative teacher preparation programmes that address barriers and improve educator outcomes through increased funding, research, and recognition of promising initiatives.
5. Invest in research on assessment practices, the validity and reliability of accountability measures, and the data points used to determine the quality of teacher preparation.
6. Commission longitudinal research to develop, implement, and evaluate valid and reliable evaluation tools.

Strengthening Teacher Education Systems

7. Offer comprehensive technical assistance to TEPs and schools to support the development, implementation, improvement, and expansion of TEPs nationwide.
8. Establish a system-wide academy that provides professional development, networking, and mentorship opportunities for new mentor teachers to strengthen their skills.
9. Include the voices of student teachers on government committees involved in teacher education and professional development.
10. Examine and revise policies that may hinder or discourage experienced teachers with extensive tacit knowledge from entering teacher education roles.
11. Ensure that compensation for university-based teacher educators is competitive with the broader education system and consider relevant work experience when determining starting salaries.
12. Convene groups of teacher educators and connect them with their Members of Parliament to advocate for policies that support teacher education and professional development.

11.3 Limitations

Although efforts have been made to address the constraints of each of the methodologies reflected in this report, limitations of these approach were anticipated and must be addressed. As the experiences of those involved in making judgements and requirements for determining effectiveness vary, it can be difficult to investigate reliability and validity across multiple TEPs, each situated in complex contextual settings. It is therefore important to consider the applicability of findings and conclusions from the research presented in this report.

11.3.1 Systematic Literature Review

Results obtained through the systematic review of literature in Chapter 3 are only as reliable as the methods adopted in the original primary research. Consequently, even though quality of the original research was an inclusion criterion, any inherent issues in research design remain and may have influenced results. This study included research and practices of TEPs reflective of multiple countries, yet only examined research published in English and inclusive of the search criteria. The nature of systematic reviews means some non-included work may have been found relevant if framed in a different way, therefore exploring research beyond the inclusion parameters may have identified further sources.

11.3.2 Teaching Standards Policies Analysis

Although efforts have been made to address the nuances of the novel, blended methodology employed in Phase 2, limitations were anticipated. The reality remains that the professional standards investigated are in a constant state of implementation and situated in complex and dynamic contextual settings. UNESCO's *Global Framework of Professional Teaching Standards* (Education International & UNESCO, 2019) was utilized to anchor the comparative analysis. If a different set of recognized standards had been utilized (e.g., Danielson's Framework for Teaching, 2007; Marzano's Instructional Framework, 2017), it is possible an adjusted alignment could have emerged. Also, had a researcher outside of any of the consistent policy contexts engaged in the crosswalk exercise, the probability remains that differences would have occurred. While it was not possible for the researchers to stand outside of the policies being studied, free of values and meanings, the researchers remained aware of their own values, beliefs, and feelings. It is therefore important to consider the applicability of findings and conclusions.

11.3.3 Case Studies

Although the case study design is particularly situated for investigating a complex educational phenomenon and advancing the knowledge base, limitations have been identified. Saturation of data for rich, thick description is limited due to purposeful selection of the participants and institutions. It is important to note that data within the case studies are based on a small sample of participants, and the three cases in this study are only a portion of the ITE staff, tutors, and school-based mentor teachers from each participating institution. Thus findings are not necessarily representative or generalizable outwith the institutions. However, the data does provide some valuable insights into the factors that are considered important when making judgements about student teachers' practices. Also, tripartite conversation is an essential feature of programmes in the case studies, but this was not replicated in this study. Although efforts have been made to address bias, it remains an inherent issue in case study research. The case studies only address a part of the whole decision-making process, a decision with a significant moral and ethical dimension embedded in long-established and complex processes and settings; limitations on seeing the entire sociocultural context are inherent. Finally, SJT was selected as the appropriate theoretical framework; different conceptions of judging readiness resulting from an alternative theoretical grounding could give rise to different ways of investigating judgement-making.

11.3.4 Delphi Panel

The Delphi panel is an established method to obtain consensus among experts on a particular topic and has several strengths (Bolger & Wright, 2011); however, limitations also need to be taken into account. The method is valuable to create a discussion among individuals with different backgrounds and professions which moves towards clarification and consensus. The panel members in this study demonstrated a strong degree of agreement on the core topic; the absence of strong debate has the possibility of missed interpretations or ideas. We formed a panel to include perspectives of university-based teacher educators, school leaders, researchers, and teachers with a variety of expertise (i.e., academic research, systems change) from six countries, three being from Scotland and none from the UK nation of Wales.

Answers from the panel might thus be influenced by these contexts and consensus building could vary with a different panel of experts. Furthermore, the panel discussion considered judgement-making in application to traditional undergraduate and postgraduate certificate programmes situated within the UK. Consideration of non-university-based TEPs or bringing together experts from other countries might have resulted in a different focus. We presented the panel members with a brief in advance of the first round of discussions with topics and queries based on literature, professional standards policies, and the empirical data collected in other phases of the project. This could have 'primed' panel members to think in a certain direction.

11.4 Future Research

The findings presented in this report provide a solid foundation for future research. To delve deeper into this topic, it is recommended that future studies focus on specific aspects raised across the project phases. Notably, in the systematic review of literature, no studies of student teacher assessment were identified from programmes outside of universities (i.e., alternative provisions); all took place within the context of university-based teacher preparation programmes. This is a notable finding, indicating the need for further investigation of the work of alternate education provisions and methods of judging classroom readiness. Our examination by country of origin and utilization of the 11 examined student teacher evaluation tools revealed that all but one instrument was created in the US. It would be of interest to examine the validity and reliability of those created/used in the UK to deepen knowledge of assessment quality and how evaluations of teaching are conducted in the devolved nations.

The comparative analysis of professional teaching standards could be further developed to gain from 'policy learning' of professional teaching standards globally, such as in the seven international jurisdictions with what Sato and Abbiss (2021) termed 'highly developed teacher education systems' (i.e., New South Wales, Australia; Alberta and Ontario, Canada; Shanghai, China; and Singapore). Additionally, since there is evidence of a strong professional standards initiative in the US (Sachs, 2005), which is also the context in which much research is being carried out on the use of stands-based assessments of future teachers (Anderson et al., 2024), it would be of interest to expand comparative work to include the *InTASC Model Core Teaching Standards* (Council of Chief State School Officers, 2013), which are utilized across the US. Beyond considering an alternative set of anchoring standards and global comparison, future research could further develop and confirm these findings by extending the comparative analysis to include Northern Ireland. This would encapsulate the four nations of the UK and could bring forward additional insights regarding professional teaching standards.

Despite the valuable insights gained from the case studies, several areas warrant further investigation. Future studies could delve into the reasons behind the differences in opinion between university-based teacher educators, tutors, and mentor teachers to better understand their perspectives. The variation in views on the contextual nature of judgements suggests a need to further explore how to effectively account for context in teacher evaluation; this

would delve deeper into the factors influencing judgement, particularly those with varied agreement among groups. To better understand the impact of different evaluation practices, longitudinal studies could be conducted to track teacher performance and growth over time; research tracking the effectiveness of different evaluation approaches over the course of an entire TEP is essential for understanding the impact on teacher development and student outcomes. Future research could repeat the case studies, but with a formal situational sample (i.e., real student teachers during their placement or live video feed) instead of a video of a lesson.

The Delphi panel triangulated findings across phases of the project and provided key recommendations. While this panel was comprised of international experts with a variety of roles in ITE, it would be of interest to repeat the Delphi process with other panels, considering the same findings. It would be of interest to limit participants to one of the three groups involved in the case studies (i.e., teacher educators, tutors, and mentor teachers) and to form a panel from each separate nation and then compare the results. It would also be useful to repeat the panel with student teachers themselves. And it would potentially be of interest to expand the iterative round for consensus building to occur across multiple days.

Development of the *Duplexity Model of Dynamic, Adaptive Systems Thinking in Teacher Education* also provides ample opportunity for future research. As an emerging conceptual model, it would be of interest to continue to apply conceptualizations with additional simulated and real practices in teacher education in different contexts to further explore functionality. Additionally, further exploration into what specifically constitutes the threshold barriers for any given ITE programme in its specific context would be of interest. It would be advantageous to further explore the variables and confirm or refine these based on feedback from those who experimented with application. It could be of interest to put the model forward to a Delphi panel as well. Finally, given that one of our partner institutions in this project was unsuccessful in its bid for re-accreditation, it would be of interest to apply the model in an effort to understand better what occurred and identify any potential shortfalls or overshoots.

Ultimately, the question of judging new teachers' effectiveness remains complex and multifaceted, requiring further exploration. The findings of this project provide a solid foundation for future research in this area. By addressing the identified gaps, forthcoming studies can contribute significantly to our understanding of judging teaching effectiveness and guide transformative practices.

References

References marked with an asterisk indicate studies included in the systematic analysis (see Chapter 3).

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 20-34.

American Association of Colleges for Teacher Education. (2018). *A pivot toward clinical practice, its lexicon, and the renewal of educator preparation: A report of the AACTE Clinical Practice Commission*. http://www.kacte.org/assets/whitepaper-12-21-2017.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, R. (2018). Historical perspectives. In T. G. K. Bryce, W. M. Humes, D. Gillies & A. Kennedy (Eds.), *Scottish Education* (pp. 99–107). Edinburgh University Press.

Anderson, S. K. (2023, 19 November). *Navigating the reform agenda: A few considerations for government's role*. Learning and Unlearning: Scholarship of Teaching and Learning in the School of Education. https://learningandunlearning.gla.ac.uk/index.php/2023/11/19/navigating-the-reform-agenda-a-few-considerations-for-governments-role/

Anderson, S. K., & McMahon, M. (2024). A partnered approach to quality assurance in educator preparation. In M. Monaco, E. Thomas Horne, & C. Tanguay (Eds.), *Working Together for Success: A Guide to Effective Stakeholder Engagement and Collaboration.* Information Age Publishing. (In Press)

Anderson, S. K., Ozsezer-Kurnuc, S., & Jain, P. (2024). Judging student teacher effectiveness: a systematic review of literature. *British Journal of Educational Studies*, 1–33. https://doi.org/10.1080/00071005.2024.2374070

Anderson, S., & Tonner, P. (2023). A post-critical pedagogy for sustainability: Engaging the head, heart, and hands. *Open Scholarship of Teaching and Learning*, *2*(3), 163–172. https://doi.org/10.56230/osotl.54

Andrews, L. (2011). *Teaching makes a difference* (speech at the Reardon Smith Lecture Theatre, Cardiff, 2 February). https://www.iwa.wales/click/2011/02/teaching-makes-a-difference/

Andrews, L. (2014) *Ministering to education: a reformer reports*. Parthian Books.

* Ata, A., & Kozan, K. (2018). Factor analytic insights into micro-teaching performance of teacher candidates. *International Online Journal of Education and Teaching*, *5*(1), 169–178. http://iojet.org/index.php/IOJET/article/view/264/225

Asher, L. (2018) How ed schools became a menage to higher education. *The Chronicle of Higher Education*. Available at https://www.chronicle.com/article/how-ed-schools-became-a-menace/

Australian Council for Educational Research. (2014). *Best practice teacher education programs and Australia's own programs*. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1014&context=teacher_education

* Basit, I., & Khurshid, F. (2018). Satisfaction of prospective teachers and teacher educators about the quality of teacher education programs. *Journal of Research in Social Sciences*, *6*(2), 168–188.

Bastian, K. C., Patterson, K. M., & Carpenter, D. (2022). Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy*, *36*(7), 1583–1611. https://doi.org/10.1177/0895904820951126

Baumfield, V. M., Conroy, J. C., Davis, R. A., & Lundie, D. C. (2012). The Delphi method: Gathering expert opinion in religious education. *British Journal of Religious Education*, *34*(1), 5–19. DOI: 10.1080/01416200.2011.614740

* Beare, P., Torgerson, C., Marshall, J., Tracz, S., & Chiero, R. (2014). Examination for bias in principal ratings of teachers' preparation. *The Teacher Educator*, *49*(1), 75–88. DOI: 10.1080/08878730.2013.848005

Beck, A. D. (2016). *Policy processes, professionalism and partnership: An exploration of the implementation of 'Teaching Scotland's Future'*. PhD thesis, University of Glasgow.

Beck, A. D. (2023). Translating teachers as leaders of educational change: Briefcases, biscuits, and teacher participation in policymaking. *Journal of Educational Administration and History*, *56*(1), 22–38. https://doi.org/10.1080/00220620.2023.2275126

* Behizadeh, N., & Neely, A. (2018). Testing injustice: Examining the consequential validity of edTPA. *Equity & Excellence in Education*, *51*(3–4), 242–264. DOI: 10.1080/10665684.2019.1568927

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021, May). Preparing, conducting and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX*, *28*(8), 1–20. DOI: 10.1016/j.mex.2021.101401

* Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment*, *23*(4), 229–249. http://dx.doi.org/10.1080/10627197.2018.1513788

Bell, C. A., Dobbelaer, M. J., Klette K., & Visscher A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bermudez, A. (2015). Four tools for critical inquiry in history, social studies, and civic education. *Revista de Estudios Sociales*, *52*(52), 102–118. DOI: 10.7440/res52.2015.07

Biesta, G. (2015). What is education for? On Good education, teacher judgement, and educational professionalism. *European Journal of Education*, *50*(1), 75–87. https://doi.org/10.1111/ejed.12109

Biesta, G. (2020). *Educational research: An unorthodox introduction*. Bloomsbury.

Boguslav, A., & Cohen, J. (2024). Different methods for assessing preservice teachers' instruction: Why measures matter. *Journal of Teacher Education*, *75*(2), 168–185.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change*, *78*(9), 1500–1513. http://doi.org/10.1016/j.techfore.2011.07.007

Borremans, L. F. N., & Split, J. L. (2023). Towards a curriculum targeting teachers' relationship-building competence: Results of a Delphi study. *Teaching and Teacher Education*, *130*, 1–21. https://doi.org/10.1016/j.tate.2023.104155

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Teacher preparation and student achievement: Working paper 14314*. National Bureau of Economic Research. http://www.nber.org/papers/w14314

Bransford, J., Darling-Hammond, L., & LePage, P. (2005). Introduction. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 1–39). Jossey-Bass.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.

Brown, B. (1968). *Delphi process: A methodology used for the elicitation of opinions of experts*. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/papers/2006/P3925.pdf

* Brown, E. L., Suh, J., Parsons, S. A., Parker, A. K., & Ramirez, E. M. (2015). Documenting teacher candidates' professional growth through performance evaluation. *Journal of Research in Education*, *25*(1), 35–47.

Bruner, J. (1960). *The process of education*. Harvard University Press.

Bryman A. (2016). *Social research methods*. Oxford University Press.

Cameron-Jones, M., & O'Hara, P. (1994). What employers want to read about new teachers. *Journal of Education for Teaching*, *20*(2), 203–214. https://doi.org/10.1080/0260747940200208

Campbell, C., & Harris, A. (2023, 31 May). *All learners in Scotland matter – national discussion on education: Final report*. Scottish Government. https://www.gov.scot/publications/learners-scotland-matter-national-discussion-education-final-report/

Cantor, P., Osher, D., Berg, J., Steyer, L., & Rose, T. (2018). Malleability, plasticity, and individuality: How children learn and develop in context. *Applied Developmental Science, 23*(4), 307–337. https://doi.org/10.1080/10888691.2017.1398649

Carter, C. C. (2008). Voluntary standards for peace education. *Journal of Peace Education*, *5*(2), 141–155. https://doi.org/10.1080/17400200802264347

* Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools. REL 2014–024. Regional Educational Laboratory Mid-Atlantic. https://files.eric.ed.gov/fulltext/ED545232.pdf

* Choi, H. S., Benson, N. F., & Shudak, N. J. (2016). Assessment of teacher candidate dispositions: Evidence of reliability and validity. *Teacher Education Quarterly*, *43*(3), 71–89.

Clapham, A., Richards, R., Lonsdale, K., & la Velle, L. (2023). Scarcely visible? Analysing initial teacher education research and the Research Excellence Framework. *London Review of Education*, *21*(1). https://doi.org/10.14324/LRE.21.1.24m

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education, 54*(1), 3-5.

Cochran-Smith, M. (2009). 'Re-Culturing' teacher education: Inquiry, evidence, and action. *Journal of Teacher Education*, *60*(5), 458–468.

Cochran-Smith, M. (2021). Rethinking teacher education: The trouble with accountability. *Oxford Review of Education*, *47*(1), 8–24. https://doi.org/10.1080/03054985.2020.1842181

Cochran-Smith, M., Ell, F., Ludlow, L., Grudnoff, L., & Aitken, G. (2014). The challenge and promise of complexity theory for teacher education research. *Teachers College Record, 116*(4), 1–38. https://doi.org/10.1177/016146811411600407

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Colón, E. P., Dassa, L. M., Dana, T. M., & Hanson, N. P. (2024). Agree to disagree: Multiple methods to assess rater agreement during student teaching. *Action in Teacher Education*, 1–18. https://doi.org/10.1080/01626620.2024.2344554

Commission on Accreditation of Athletic Training Education. (2020). 2020 *Standards for accreditation of professional athletic training programs crosswalk*. https://caate.net/Portals/0/Documents/2020-Standards-Crosswalk_Final_for-Professional-Programs.pdf

Commission on Teacher Credentialing, American Speech-Language-Hearing Association & Council on Academic Accreditation (2020). *Standards crosswalk*. https://www.ctc.ca.gov/educator-prep/accred-files/CTC-ASHA-Alignment-Matrix.doc

* Conderman, G., & Walker, D. A. (2015). Assessing dispositions in teacher preparation programs: Are candidates and faculty seeing the same thing? *The Teacher Educator*, *50*(3), 215–231. DOI: 10.1080/08878730.2015.1010053

Connolly, M., Milton, E., Davies, A. and Barrance, R. (2018). 'Turning heads: The impact of political reform on the professional role, identity and recruitment of head teachers in Wales'. *British Educational Research Journal.* https://doi.org/10.1002/berj.3450

Conroy, J. (2004). *Betwixt and between: The liminal imagination, education and democracy.* Peter Lang.

Conroy, J., Hulme, M., & Menter, I. (2013). Developing a clinical model for teacher education. *Journal of Education for Teaching*, *39*(5), 557–573.

Conroy, J. C., & Smith, R. (2017). The ethics of research excellence. *Journal of Philosophy of Education*, *51*(4), 693–708. https://doi.org/10.1111/1467-9752.12249

Cooksey, R. W. (1988). Social judgement theory in education: Current and potential applications. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgement the SJT view* (pp. 273-315). Elsevier.

Cooksey, R. W. (1996). The methodology of social judgment theory. *Thinking and Reasoning*, *2*(2), 141–174.

Council for the Accreditation of Educator Preparation. (n.d.). Glossary. https://caepnet.org/glossary

Council of Chief State School Officers. (2011, April). *Interstate Teacher Assessment and Support Consortium (InTASC) model core teaching standards: A resource for state dialogue.* https://ccsso.org/resource-library/intasc-model-core-teaching-standards-and-learning-progressions-teachers-10

Council of Chief State School Officers. (2013, April). *InTASC model: Core teaching standards and learning progressions for teachers 1.0: a resource for ongoing teacher development*. https://ccsso.org/sites/default/files/2017-12/2013_INTASC_Learning_Progressions_for_Teachers.pdf

Council of Chief State School Officers (2022). *Crosswalk of the Professional Standards for Educational Leaders to the Leadership Competencies for Learner-Centered, Personalized Education*. https://ccsso.org/resource-library/crosswalk-professional-standards-educational-leaders-leadership-competencies

Creswell, J. W., & Creswell, J. D. (2023). *Research design – international student edition: Qualitative, quantitative, and mixed methods approaches* (6th ed.). Sage.

Creswell, J. W., & Zhang, W. (2009). The application of mixed methods designs to trauma research. *Journal of Traumatic Stress: Official Publication of the International Society for Traumatic Stress Studies*, *22*(6), 612–621. https://doi.org/10.1002/jts.20479

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed). Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.

Darling-Hammond, L. (2017). Teacher education around the world: What can we Learn from international practice? *European Journal of Teacher Education*, *40*(3) 291–309. https://doi.org/10.1080/02619768.2017.1315399

Darling-Hammond, L. & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. John Wiley & Sons.

Darling-Hammond, L. & Lieberman, A. (2012). *Teacher education around the world: Changing policies and practices*. Routledge.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B. J., & Osher, D. (2019). Implications for educational practice of the science of learning and development. *Applied Developmental Science, 24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

Darling-Hammond, L., Schachner, A. C. W., Wojcikiewicz, S. K., & Flook, L. (2023). Educating teachers to enact the science of learning and development. *Applied Developmental Science*, *28*(1), 1–21. https://doi.org/10.1080/10888691.2022.2130506

Davis, B., & Sumara, D. (2008). Complexity as a theory of education. *Transnational Curriculum Inquiry*, *5*(2), 33–44. https://doi.org/10.14288/tci.v5i2.75

Davis, R. A., Conroy, J. C., & Clague, J. (2020). Schools as factories: The limits of a metaphor. *Journal of Philosophy of Education*, *54*(5), 1471–1488. https://doi.org/10.1111/1467-9752.12525

Department for Education. (2011). *Teachers' standards*. https://www.gov.uk/government/publications/teachers-standards

Department for Education. (2019). *Early Career Framework*. https://www.gov.uk/government/publications/early-career-framework

Department for Education. (2021, December). *Teachers' standards*. UK Government.

* Dewaele, J. M., Mercer, S., Talbot, K., & von Blanckenburg, M. (2021). Are EFL pre-service teachers' judgment of teaching competence swayed by the belief that the EFL teacher is a L1 or LX user of English? *European Journal of Applied Linguistics*, *9*(2), 259–282. DOI: 10.1515/eujal-2019-0030

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education.* Macmillan Publishing.

Dickson, B. (2011). Beginning teachers as enquirers: M-level work in initial teacher education. *European Journal of Teacher Education*, *34*,(3), 259–276.

Diem, S., & Young, M. D. (2015). Considering critical turns in research on educational policy. *International Journal of Educational Management*, *29*(7), 838–850.

Donaldson, G. (2010). *Teaching Scotland's future: Report of a review of teacher education in Scotland.* Scottish Government.

Donaldson, G. (2015). *Successful futures: Independent review of curriculum and assessment arrangements in Wales*. Welsh Government.

Early Childhood Personnel Center. (2020). *Cross walk of the Initial Practice-Based Professional Preparation Standards for Early Interventionists/Early Childhood Special Educators (2020) and the Professional Standards and Competencies for Early Childhood Educators (2020)*. https://ecpcta.org/wp-content/uploads/sites/2810/2020/10/Crosswalk-EI.ECSE-and-ECE-Standards-Final.pdf (Accessed 22 January 2024).

Education International & United Nations Education, Scientific and Cultural Organization. (2019). *Global framework of professional teaching standards*.

Education Scotland. (n.d.). *Curriculum for excellence*. https://education.gov.scot/media/wpsnskgv/all-experiencesoutcomes18.pdf

Education Scotland. (2018). *Self-evaluation framework for Initial Teacher Education*. Author.

Education Scotland. (2023a). *What is our role and status?* https://education.gov.scot/education-scotland/who-we-are/role-and-status/what-is-our-role-and-status

Education Scotland. (2023b, 6 June). *National Improvement Framework*. https://education.gov.scot/parentzone/curriculum-in-scotland/national-improvement-framework/

Education Scotland. (2023c, 6 September). *United Nations Convention on the Rights of the Child*. https://education.gov.scot/resources/united-nations-convention-on-the-rights-of-the-child/

Ell, F., Simpson, A., Mayer, D., McLean Davies, L., Clinton, J., & Dawson, G. (2019). Conceptualising the impact of initial teacher education. *The Australian Educational Researcher*, *46*, 177–200. https://doi.org/10.1007/s13384-018-0294-7

Ellis, V., and Childs, A. (2023). Introducing the crisis: The state, the market, the universities and teacher education in England. In V. Ellis (Ed.), *Teacher education in crisis: The state, the market and the universities in England.* Bloomsbury. https://www.bloomsburycollections.com/monograph?docid=b-9781350399693&st=Viv+Ellis

Estyn (2022). *Guidance for inspectors: What we inspect. Initial teacher education (ITE) from September 2022*.

Estyn (2023). *Guidance for inspectors: How we inspect. Initial teacher education (ITE) from October 2023*.

Faul, M. V., & Savage, L. (2023). Introduction to systems thinking in international education and development. In. M. Faul, & L. Savage (Eds.). *Systems thinking in international education and development: Unlocking learning for all?*. Edward Elgar Publishing. https://doi.org/10.4337/9781802205930

Furlong, J. (2015). *Teaching tomorrow's teachers: Options for the future of initial teacher education in Wales.* University of Oxford.

Furlong, J., Hagger, H. & Butcher, C. (2006). *Review of initial teacher training provision in Wales. Oxford University.*

General Teaching Council for Scotland. (n.d.-a). *Accredited initial teacher education programmes*. https://www.gtcs.org.uk/about-us/accreditation/accredited-initial-teacher-education-programmes

General Teaching Council Scotland. (n.d.-b). *Archive: 2012 professional standards.* https://www.gtcs.org.uk/professional-standards/archive-2012-professional-standards/

General Teaching Council Scotland. (n.d.-c). *Practitioner enquiry*. https://www.gtcs.org.uk/professional-update/practitioner-enquiry/

General Teaching Council Scotland. (2012). *Code of professionalism and conduct*. https://www.gtcs.org.uk/wp-content/uploads/2021/09/code-of-professionalism-and-conduct.pdf

General Teaching Council Scotland. (2018). *Gaelic language plan update 2018–2023: Ensuring high standards of learning and teaching in Gaelic.* https://www.gtcs.org.uk/wp-content/uploads/2021/10/gtc-scotland-gaelic-plan-2018-2023.pdf

General Teaching Council Scotland. (2021a). *GTC Scotland professional standards 2021*. https://www.gtcs.org.uk/wp-content/uploads/2021/09/professional-standards-side-by-side-comparison.pdf

General Teaching Council for Scotland. (2021b, August). *The standard for provisional registration: Mandatory requirements for registration with the General Teaching Council for Scotland*.

General Teaching Council Scotland. (2023, 29 October). *Comparison of professional standards 2012 and 2021* [PowerPoint slides]. https://www.gtcs.org.uk/professional-standards/archive-2012-professional-standards/

Gillies, D. (2018). The history of Scottish education since devolution. In T. G. K. Bryce, W. M. Humes, D. Gillies & A. Kennedy (Eds.), *Scottish education* (pp. 108–117). Edinburgh University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Glasgow Centre for Population Health (n.d.). *Understanding Glasgow: The Glasgow indicators project*. https://www.understandingglasgow.com/

* Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377–393. https://doi.org/10.1177/00224871177025

Goodman, C. M. (1987) The Delphi technique: A critique. *Journal of Advanced Nursing, 12*(6), 729–734. DOI: 10.1111/j.1365-2648.1987.tb01376.x

Goodwin, A. L. (2012). Teaching as a profession: Are we there yet? In C. Day (Ed.), *The Routledge international handbook of teacher and school development* (pp. 44–56). Routledge.

Goodwin, L. (2023). Teacher education for the 31st Century? Preparing teachers for unknown futures. In J. Madalinska-Michalak (Ed.), *Quality in teaching and teacher education: International perspectives from a changing world* (pp. 231–251). Brill.

Goodwin, A. L., & Kosnik, C. (2013). Quality teacher educators = quality teachers? Conceptualizing essential domains of knowledge for those who teach teachers. In P. Aubusson & S. Schuck (Eds). *Teacher development: Teacher education futures,* 17(3), 334–346. DOI: 10.1080/13664530.2013.813766

Green, R. A. (2014). The Delphi technique in educational research. *SAGE Open*. https://doi.org/10.1177/2158244014529773

Grossman, P. (2020). Making the complex work of teaching visible. *The Phi Delta Kappan*, 101(6), 8–13.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy* (W. Rehg, Trans.). MIT Press.

Haigh, M., & Ell, F. (2014). Consensus and dissensus in mentor teachers' judgments of readiness to teach. *Teaching and Teacher Education, 40*, 10–21. https://doi.org/10.1016/j.tate.2014.01.001

Haigh, M., Ell, F., & Mackisack, V. (2013). Judging teacher candidates' readiness to teach. *Teaching and Teacher Education*, *34*, 1–11. https://doi.org/10.1016/j.tate.2013.03.002

* Hamid, S. R. A., Hassan, S. S. S., & Ismail, N. A. H. (2012). Teaching quality and performance among experienced teachers in Malaysia. *Australian Journal of Teacher Education*, *37*(11), 85–103. DOI: 10.14221/ajte.2012v37n11.2

Hammersley, M. (2022). Emergent design. In U. Flick (Ed.), *The SAGE handbook of qualitative research design*. Sage. https://doi.org/10.4135/9781529770278

Hammond, K., Rohrbaugh, J., Mumpower, J., & Adelman, L. (1977). Social judgment theory: Applications in policy formation. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes in applied settings* (pp. 1–29). Academic Press.

Hand, R., & Rong, Y. (2014). Schools as clinics: Learning about practice in practice. P*eabody Journal of Education*, *89*(4), 453–465. https://doi.org/10.1080/0161956X.2014.938592

Harris, A., Jones, M., Southern, A. & Griffiths, J. (2022). *An Exploration of all-age schools in Wales.* Swansea University.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008–1015. DOI: 10.1046/j.1365-2648.2000.t01-1-01567.x

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Hattie, J., & Clinton, J. (2001). The assessment of teachers. *Teaching Education*, *12*(3), 279–300. DOI: 10.1080/10476210120096551

Hayward, L. (2023, 22 June). *It's our future: Report of the independent review of qualifications and assessment*. Scottish Government. https://www.gov.scot/publications/future-report-independent-review-qualifications-assessment/

Heasly, B. (2021). Edu-tensegrity: An expanded integration of 21st century education. *Discourse and Communication for Sustainable Education*, *12*(2), 76–95. DOI: 10.2478/dcse-2021-0018

Hegender, H. (2010). The assessment of teacher candidates' academic and professional knowledge in school-based teacher education. *Scandinavian Journal of Educational Research*, *54*(2), 151–171. https://doi.org/10.1080/00313831003637931

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education, 118*(4), 489-519.

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Sage Publications.

House, E. R., & Howe, K. R. (2000). Deliberative democratic evaluation in practice. In D. Stufflebeam, T. Kellaghan, & G. Maddaus (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 409–421). Springer.

House of Commons Education Committee. (2024, 8 May). *Teacher recruitment, training and retention. Second report of session 2023–24*. https://committees.parliament.uk/publications/44798/documents/222606/default/

Hovland, C. I., & Sherif, M. (1980). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Greenwood.

Hoy, D. (1994). *Critical theory and critical history*. In D. Hoy & T. McCarthy (Eds.), *Critical theory* (pp. 101–214). Blackwell.

* Hubball, H., & Dawson, S. (2014). Curriculum analytics: Application of social network analysis for improving strategic curriculum decision-making in a research-intensive university. *Teaching and Learning Inquiry, 2*(2), 59-74. https://doi.org/10.2979/teachlearninqu.2.2.59

Hylton, S. P., Joseph, J. D., Ward, T. J., & Gareis, C. R. (2022). Examining the validity of a student teaching evaluation instrument. *Teacher Educators' Journal*, *15*(1), 77–101.

Ingber, D. (1998). The architecture of life. *Scientific American*, *48*(57). DOI: 10.1038/scientificamerican0198-48

Institute of Chartered Accountants in England and Wales. (2018). *Artificial intelligence and the future of accountancy*. https://www.icaew.com/-/media/corporate/files/technical/technology/thought-leadership/artificial-intelligence.ashx

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, *28*(1), 91–105. https://doi.org/10.1080/02671522.2012.754229

* Johnston, P., Wilson, A., & Almerico, G. M. (2018). Meeting psychometric requirements for disposition assessment: Valid and reliable indicators of teacher dispositions. *Journal of Instructional Pedagogies*, *21*. https://files.eric.ed.gov/fulltext/EJ1194249.pdf

* Khan, G., Khan, A., Hussain, S., & Shaheen, N. (2017). Teacher evaluation: Global perspectives and lessons for Pakistan. *Dialogue (Pakistan)*, *12*(3). https://www.qurtuba.edu.pk/thedialogue/The%20Dialogue/12_3/Dialogue_July_September2017_333-346.pdf

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgement*. Harper Collins.

Keating, M. (2005). Higher education in Scotland and England after devolution. *Regional & Federal Studies*, *15*(4) 423–435. https://doi.org/10.1080/135975605002305

Kennedy, A., Carver, M., & Adams, P. (2023). *Measuring quality in initial teacher education: Final report*. Scottish Council of Deans of Education. https://www.mquite.scot/publications-and-presentations/

* Kennedy, A. S., & Lees, A. T. (2016). Preparing undergraduate pre-service teachers through direct and video-based performance feedback and tiered supports in Early Head Start. *Early Childhood Education Journal*, *44*, 369–379. https://doi.org/10.1007/s10643-015-0725-2

Kennedy, M. (1999). The role of preservice teacher education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 54–85). Jossey Bass.

Kerry, T. (1980). The demands made by RE on pupils' thinking. *British Journal of Religious Education*, *3*(2), 46–52. https://doi.org/10.1080/0141620800030203

* Khan, G., Khan, A., Hussain, S., & Shaheen, N. (2017). Teacher evaluation: Global perspectives and lessons for Pakistan. *Dialogue (Pakistan), 12*(3). https://www.qurtuba.edu.pk/thedialogue/The%20Dialogue/12_3/Dialogue_July_September2017_333-346.pdf

Kimmerer, R. W. (2015). *Braiding sweetgrass: Indigenous wisdom, scientific knowledge and the teachings of plants*. Penguin.

* Kingsley, L., & Romine, W. (2014). Measuring teaching best practice in the induction years: Development and validation of an item-level assessment. *European Journal of Educational Research*, *3*(2), 87–109.

Klassen, R. M., & Kim, L. E. (2019). Selecting teachers and prospective teachers: A meta-analysis, *Educational Research Review, 26*, 32-51. https://doi.org/10.1016/j.edurev.2018.12.003

Knight, B. (2017). The evolving codification of teachers' work: Policy, politics and the consequences of pursuing quality control in initial teacher education. *Teacher Education Advancement Network Journ*al, 9(1), 4–13.

Kornfeld, J., Grady, K., Marker, P. M., & Ruddell, M. R. (2007). Caught in the current: A self-study of state-mandated compliance in a teacher education program. *Teachers College Record*, *109*(2), 1902–1930. https://doi.org/10.1177/016146810710900802

Korthagen, F. A. J. (2004). In search of the essence of a good teacher: Towards a more holistic approach in teacher education. *Teaching and Teacher Education*, *20*(1), 77–97. https://doi.org/10.1016/j.tate.2003.10.002

Korthagen, F. (2017). Inconvenient truths about teacher learning: Towards professional development 3.0. *Teachers and Teaching, Theory and Practice*, *23*(4), 387–405. https://doi.org/10.1080/13540602.2016.1211523

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

la Velle, L. (2020). Teacher education: The transformation of transitions in learning to teach. *Journal of Education for Teaching: International Research and Pedagogy*, *46*(2), 141–144.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

* Lazarev, V., Newman, D., Nguyen, T., Lin, L., & Zacamy, J. (2017). The Texas Teacher Evaluation and Support System rubric: Properties and association with school characteristics. REL 2018-274. Regional Educational Laboratory Southwest. https://files.eric.ed.gov/fulltext/ED576984.pdf

Levine, A. (2006). Educating school teachers [Electronic version]. Retrieved 27 March 2024 from https://files.eric.ed.gov/fulltext/ED504144.pdf

Lofthouse, R. M. (2018). Re-imagining mentoring as a dynamic hub in the transformation of initial teacher education: The role of mentors and teacher educators. *International Journal of Mentoring and Coaching in Education, 7*(3), 248–260.

Lortie, D. C. (1975). *Schoolteachers: A sociological study*. University of Chicago Press.

* Lyness, S. A., Peterson, K., & Yates, K. (2021). Low inter-rater reliability of a high stakes performance assessment of teacher candidates. *Education Sciences*, *11*(10), 1–16. https://doi.org/10.3390/educsci11100648

* Maharaj, S. (2014). Administrators' views on teacher evaluation: Examining Ontario's teacher performance appraisal. *Canadian Journal of Educational Administration and Policy*, 152, 1–58.

Mahlmeister, L. (2015). Crosswalk: the Joint Commission and Centers for Medicare and Medicaid services pathway to patient safety and quality, *Journal of Perinatal and Neonatal Nursing,* 29(2), 107–115.

Martin, S. D., McQuitty, V., & Morgan, D. N. (2019). Complexity theory and teacher education, *Oxford Research Encyclopedias*. DOI: 10.1093/acrefore/9780190264093. 013.479

Marzano, R. J. (2017). *The new art and science of teaching*. Solution Tree.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Association for Supervision and Curriculum Development.

* Masuwai, A. M., & Saad, N. S. (2016). Evaluating the face and content validity of a Teaching and Learning Guiding Principles Instrument (TLGPI): A perspective study of Malaysian teacher educators. *Geografia*, *12*(3), 11–21. https://www.researchgate.net/publication/299265585

Matsumoto-Royo, K., & Ramírez-Montoya, M. S. (2021). Core practices in practice-based teacher education: A systematic literature review of its teaching and assessment process. *Studies in Educational Evaluation*, 70, 1–13. https://doi.org/10.1016/j.stueduc.2021.101047

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Sage.

McCool, M. (2024, August 14). Parents begin legal action over Glasgow teacher cuts. BBC Scotland News. https://www.bbc.co.uk/news/articles/c0e8gpdq7weo#:~:text=Glasgow%20City%20Council's%20budget%2C%20passed,over%20the%20next%20three%20years.

McEnaney, J. (2023, June 8). *Scottish Government announces withdrawal of vital funding for teachers*. The Herald. https://www.heraldscotland.com/politics/23574780.scottish-government-announces-withdrawal-vital-funding-teachers/

McKenna H. P. (1994). The Delphi technique: a worthwhile approach for nursing? *Journal of Advanced Nursing, 19,* 1221–1225.

McLean, D., Worth, J., & Smith, A. (2024, 20 March). *Teacher labour market in England*. National Foundation for Educational Research. https://www.nfer.ac.uk/publications/teacher-labour-market-in-england-annual-report-2024/

McLean Davies, L., Dickson, B., Rickards, F., Dinham, S., Conroy, J., & Davis, R. (2015). Teaching as a clinical profession: translational practices in initial teacher education – an international perspective. *Journal of Education for Teaching*, *41*(5), 514–528. DOI: 10.1080/02607476.2015.1105537

McMahon, M. (2021). *Literature review on professional standards for teaching*. General Teaching Council. https://www.gtcs.org.uk/wp-content/uploads/2021/09/literature-review-professional-standards-margery-mcmahon.pdf

Menter, I. (2016). Teacher education – making connections with curriculum, pedagogy and assessment. In D. Wyse, L. Hayward, & Pandya, J. (Eds.), *The SAGE handbook for curriculum, pedagogy and assessment* (pp. 1015–1028). Sage.

Merriam, S. (1998). *Qualitative research and case study applications in education*. Jossey-Bass Publishers.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). Jossey-Bass Publishers.

Milbourne, L., & Cushman, M. (2013). From the third sector to the big society: How changing UK government policies have eroded third sector trust. *Voluntas*, 24, 485–508. https://doi.org/10.1007/s11266-012-9302-0

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications.

Milton, E., Morgan, A., Davies, A. J., Connolly, M., Donnelly, D., & Ellis, I. (2020). Framing headship: a demand-side analysis of how the headteacher role is articulated in job descriptions. *International Journal of Leadership in Education*, *26*(2), 339–358. https://doi.org/10.1080/13603124.2020.1811898

* Mkhasibe, R. G., Maphalala, M. C., & Nzima, R. D. (2018). Perceptions of subject mentors of pre-service teachers' readiness to teach economics and management sciences in the development of South Africa. *Journal of Gender, Information and Development in Africa*, *7*(2), 241–259. https://www.researchgate.net/publication/330779459

* Montecinos, C., Rittershaussen, S., Cristina Solís, M., Contreras, I., & Contreras, C. (2010). Standards-based performance assessment for the evaluation of student teachers: A consequential validity study. *Asia-Pacific Journal of Teacher Education, 38*(4), 285–300. DOI: 10.1080/1359866X.2010.515941

Moon, D. (2012). Rhetoric and policy learning: On Rhodri Morgan's 'clear red water' and 'made in Wales' health policies. *Public Policy and Administration*, 28(3), 306–323. https://doi.org/10.1177/0952076712455821

Morgan, A. (2021, November). *Additional support for learning action plan: A progress report.* Scottish Government and COSLA. https://www.gov.scot/publications/additional-support-learning-action-plan-progress-report/documents/

Morgan, A., Davies, A.J., and Milton, E. (2024). Using discourse analysis to inform content analysis: A pragmatic, mixed methods approach to exploring how the headteacher role is articulated in job descriptions. In H. Kara, D. Mannay & A. Roy (Eds.), *Handbook of creative methodology*. Policy Press.

Morse, J. M. (1994). Emerging from the data: Cognitive processes of analysis in qualitative inquiry. In J. Morse (Ed.), *Critical issues in qualitative research* (pp. 23–43). Sage.

Moss, P., Girard, B., & Haniford, L. (2006). Validity in educational assessment. In Special issue on rethinking learning: What counts as learning and what learning counts. *Review of Research in Education, 30*, 109–162. DOI: 10.3102/0091732X030001109

Moss, P. A., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal*, *38*(1), 37–70. https://doi.org/10.3102/00028312038001037

Muir, K. (2022, 9 March). *Putting learners at the centre: Towards a future vision for Scottish education*. Scottish Government. https://www.gov.scot/publications/putting-learners-centre-towards-future-vision-scottish-education/

* Murley, L. D., Stobaugh, R., Jukes, P., & Tassell, J. (2014). Examining the reliability of a culminating teacher education assessment and discovering areas for reform. *Educational Renaissance*, *2*(2), 3–18. DOI: 10.33499/edren.v2i2.61

Murray-Harvey, R., Slee, P., Lawson, M. Silins, H., Banfield, G., & Russell, A. (2000). Under stress: The concerns and coping strategies of teacher education students. *European Journal of Teacher Education, 23*(1), 19-35.

Mutton, T., & Burns, K. (2024). Does initial teacher education (in England) have a future? *Journal of Education for Teaching*, *50*(2), 214–232. DOI: 10.1080/02607476.2024.2306829.

National Academy of Education. (2024). *Evaluating and improving teacher preparation programs*. https://naeducation.org/evaluating-and-improving-teacher-preparation-programs-project/

Nesterova, Y., & Anderson, S.K. (2024). Visions of peace: Exploring how Scottish youth understand and define peace. *Prospects*. https://doi.org/10.1007/s11125-024-09700-0

Oakeshott, M. (1971). Education: The engagement and its frustration. *Journal of Philosophy of Education*, *5*(1), 43–76. https://doi.org/10.1111/j.1467-9752.1971.tb00448.x

O'Keefe, D. J. (2015). *Persuasion: Theory and research* (3rd ed.). Sage.

O'Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education*, *39*(1), 4–16. http://www.jstor.org/stable/42001807

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376–407. https://doi.org/10.3102/0034654311413609

Organisation for Education Cooperation and Development. (2018). *OECD Initial teacher preparation study: Promising practices clinical practice approaches in initial teacher education in Australia*.

Organisation for Economic Co-operation and Development. (2020). *Achieving the new Curriculum for Wales*. https://doi.org/10.1787/4b483953-en

Organisation for Economic Co-operation and Development. (2021). *Scotland's Curriculum for Excellence: Into the future*. https://www.oecd.org/education/scotland-s-curriculum-for-excellence-bf624417-en.htm

Osher, D., Cantor, P., Berg, J., Steyer, L., & Rose, T. (2018). Drivers of human development: How relationships and context shape learning and development. *Applied Developmental Science, 24*(1), 6–36. https://doi.org/10.1080/10888691.2017.1398650

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – a web and mobile app for systematic reviews. *Systematic Reviews*, *5,* 210. https://doi.org/10.1186/s13643-016-0384-4

Oxford Learner's Dictionaries. (n.d.). *Judgement*. https://www.oxfordlearnersdictionaries.com/

Oxley, E., Nash, H. M., & Weighall, A. R. (2024). Consensus building using the Delphi method in educational research: A case study with educational professionals. *International Journal of Research & Method in Education*, 1–15. https://doi.org/10.1080/1743727X.2024.2317851

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*(71). https://doi.org/10.1136/bmj.n160

* Papanastasiou, E. C., Tatto, M. T., & Neophytou, L. (2012). Programme theory, programme documents and state standards in evaluating teacher education. *Assessment & Evaluation in Higher Education*, *37*(3), 305–320. DOI: 10.1080/02602938.2010.534760

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

* Parkes, K. A., & Powell, S. R. (2015). Is the edTPA the right choice for evaluating teacher readiness? *Arts Education Policy Review*, *116*(2), 103–113.
DOI: 10.1080/10632913.2014.944964

Porter, A., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.; pp. 259–297). AERA.

Power, S. (2016) The politics of education and the misrecognition of Wales, *Oxford Review of Education*, 42(3), 285–298.

* Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher*, *27*(2).
https://scholarworks.bgsu.edu/mwer/vol27/iss2/3

Pyrczak, F., & Oh, D. M. (2018). *Making sense of statistics: A conceptual overview* (7th ed.). Routledge.

Quality Assurance Agency [QAA] Scotland (2023, April). *Quality enhancement and standards review University of Glasgow: Review report*.
https://www.qaa.ac.uk/docs/qaa/reports/university-of-glasgow-qesr-23.pdf?sfvrsn=d80fad81_4

Qualifications Wales. (2023). *Made-for-Wales GCSE decisions*. Available at:
https://qualifications.wales/regulation-reform/reforming/qualified-for-the-future/made-for-wales-gcses/

* Rafiq, S., & Qaisar, S. (2021). Teachers perception about process of teacher evaluation: A case study of a private University of Lahore. *Gomal University Journal of Research*, *37*(3), 350–362. DOI: 10.51380/gujr-37-03-09

* Rafiq, S., Qaisar, S., & Butt, I. H. (2022). Analysis of tools used for teacher evaluation process at university level: A document analysis approach. *Gomal University Journal of Research*, *38*(2), 214–224. DOI: 10.51380/gujr-38-02-08

Rasooli, A., Rasegh, A., Zandi, H., & Firoozi, T. (2023). Teachers' conceptions of fairness in classroom assessment: An empirical study. *Journal of Teacher Education*, *74*(3), 260–273.
https://doi.org/10.1177/00224871221130742

Raths, J., & Lyman, F. (2003). Summative evaluation of student teacher: An enduring problem. *Journal of Teacher Education*, *54*(2), 201–216.
https://doi.org/10.1177/0022487103054003003

Raworth, K. (2017). *Doughnut economics: Seven ways to think like a 21st-Century economist.* Random House.

* Ritzhaupt, A. D., Ndoye, A., & Parker, M. A. (2010). Validation of the electronic Portfolio Student Perspective Instrument (EPSPI) conditions under a different integration initiative. *Journal of Computing in Teacher Education*, *26*(3), 111–119. DOI: 10.1080/10402454.2010.10784642

* Rizwan, S., & Masrur, R. (2018). Standard based three dimensional capacity development of in-service school teachers. *Bulletin of Education and Research*, *40*(3), 31–44. https://pu.edu.pk/images/journal/ier/PDF-FILES/2_40_3_18.pdf

* Roloff, J., Klusmann, U., Lüdtke, O., & Trautwein, U. (2020). The predictive validity of teachers' personality, cognitive and academic abilities at the end of high school on instructional quality in Germany: A longitudinal study. *AERA Open*, *6*(1), 2332858419897884.

Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *American Educator*, *36*(1), 12–39.

Sachs, J. (2005). Teacher professional standards: a policy strategy to control, regulate or enhance the teaching profession? In N. Bascia, A. Cumming, A. Datnow, K. Leithwood, & D. Livingstone (Eds.). *International handbook of educational policy.* Springer International Handbooks of Education, vol 13 (pp. 579–592). Springer. https://doi.org/10.1007/1-4020-3201-3_29

* Saltis, M. N., Giancaterino, B., & Pierce, C. (2020). Professional dispositions of teacher candidates: Measuring dispositions at a large teacher preparation university to meet national standards. *The Teacher Educator*, *56*(2), 117–131. DOI: 10.1080/08878730.2020.1817217

Sandholtz, J. H., & Shea, L.M. (2011). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education*, 63(1) 39–50.

* Sandoval, C., van Es, E. A., Campbell, S. L., & Santagata, R. (2020). Creating coherence in teacher preparation: Examining teacher candidates' conceptualizations and practices for equity. *Teacher Education Quarterly*, *47*(4), 8–32. https://www.jstor.org/stable/10.2307/26977528

Schiro, M. S. (2013). *Curriculum theory: Conflicting visions and enduring concerns*. Sage.

Schmalz, U., Spinler, S., & Ringbeck, J. (2021). Lessons learned from a two-round Delphi-based scenario study. *MethodsX*, *8*. DOI: 10.1016/j.mex.2020.101179

Schmoker, M. (2006). *Results now: How we can achieve unprecedented improvements in teaching and learning*. Association for Supervision and Curriculum Development.

Schmoker, M. (2023). Results now 2.0: *The untapped opportunities for swift, dramatic gains in achievement*. ASCD.

Scottish Government. (2022, 3 October). T*he number of schools of different faiths in Scotland: FOI release*. https://www.gov.scot/publications/foi-202200317073/

Scottish Government. (2023a, 15 October). *News: Centre of Teaching Excellence*. https://www.gov.scot/news/centre-of-teaching-excellence/

Scottish Government. (2023b, 5 December). *Programme for International Student Assessment (PISA 2022): Scotland's results – highlights*. https://www.gov.scot/publications/programme-international-student-assessment-pisa-2022-highlights-scotlands-results/

Scottish Qualification Authority. (n.d.). *About us*. https://www.sqa.org.uk/sqa/5656.html

Seidenberg, M. (2017) Language at the speed of sight: How we read, why so many can't, and what can be done about it. New York: BasicBooks.

* Shahzad, S., & Mehmood, N. (2019). Development of teaching effectiveness scale for university teachers. *Journal of Research in Social Sciences*, *7*(2), 1–14. DOI: 10.52015/jrss.7i2.74

Sibieta, L., & Fullard, J. (2021, July). *The evolution of cognitive skills during childhood across the UK.* Educational Policy Institute. https://epi.org.uk/wp-content/uploads/2021/07/EPI_UK-Comparisons-Cognitive-outcomes-1.pdf

Silverman, D. M., Hernandez, I. A., & Destin, M. (2023). Educators' beliefs about students' socioeconomic backgrounds as a pathway for supporting motivation. *Personality and Social Psychology Bulletin*, *49*(2), 215–232. https://doi.org/10.1177/01461672211061945

* Smalley, S., & Retallick, M. (2012). Agricultural education early field experience through the lens of the EFE model. *Journal of Agricultural Education*, *53*(2), 99–109. DOI: 10.5032/jae.2012.02099

Smith, H. (2013). A critique of the teaching standards in England (1984-2012): Discourses of equality and maintaining the status quo. *Journal of Education Policy*, *28*(4), 427–448.

Social Mobility Commission. (2021, 30 June). *Against the odds: Better outcomes for disadvantaged pupils*. UK Government. https://www.gov.uk/government/publications/against-the-odds

Spendlove, D. (2024) The state of exception: How policies created the crisis of ITE in England. In V. Elis (Ed.) *Teacher education in crisis* (pp. 43–62). Bloomsbury. https://doi.org/10.5040/9781350399693.ch-3

Stanford Medicine. (2024). *2.1: Competencies and objectives for medical student education*. https://med.stanford.edu/md/mdhandbook/section-2-general-standards/2-1--competencies-and-objectives-for-medical-student-education.html

Stewart D., & Shamdasani, P. (1980). *Focus groups: Theory & practice.* Vol. 20: Applied social research methods series. Sage.

Tabberer, R. (2013). *A review of initial teacher training in Wales.* Welsh Government.

* Tait-McCutcheon, S., & Knewstubb, B. (2018). Evaluating the alignment of self, peer and lecture assessment in an Aotearoa New Zealand pre-service teacher education course. *Assessment & Evaluation in Higher Education*, *43*(5), 772–785. https://doi.org/10.1080/02602938.2017.1408771

Tan, C. Y., Hong, X., Gao, L., & Song, Q. (2023). Meta-analytical insights on school SES effects. *Educational Review*, 1–29. https://doi.org/10.1080/00131911.2023.2184329

* Tanguay, C. L. (2020). High-stakes assessment in elementary education teacher preparation: Educators' perceptions and actions resulting in curriculum change. *Education Policy Analysis Archives*, *28*(53), 1–39. DOI: 10.15407/epaa.4840

Thomas, M., Rees, B., Evans, G.E., Thomas, N., Williams, C., Lewis, B., et al. (2020). Teach beyond boundaries: the conceptual framework and learning philosophy of an innovative initial teacher education programme in Wales, *Wales Journal of Education*, *22*(1), 114–140.

* Tigelaar, D. E. H., & van Tartwijk, J. (2010). The evaluation of prospective teachers in teacher education. *International Encyclopedia of Education* (3rd ed.), 511–517. DOI: 10.1016/B978-0-08-044894-7.00647-3

* Tillema, H. (2010). Formative assessment in teacher education and teacher professional development. *International Encyclopedia of Education* (3rd ed.), 563–571. DOI: 10.1016/B978-0-08-044894-7.01639-0

* Tobón, S., Juárez-Hernández, L. G., Herrera-Meza, S. R., & Núñez, C. (2021). Pedagogical practices: Design and validation of SOCME-10 rubric in teachers who have recently entered basic education. *Educational Psychology*, *27*(2), 155–165. https://doi.org/10.5093/psed2021a13

* Tracz, S., Torgerson, C., & Beare, P. (2017). The NCTQ selectivity standard and principal evaluation of teacher preparation. *The Teacher Educator*, *52*(1), 8–21. DOI: 10.1080/08878730.2016.1186766

UK Research and Innovation (UKRI). (2023). *Research Excellence Framework (REF).* https://2021.ref.ac.uk/about-the-ref/what-is-the-ref/index.html

United Nations. (n.d.). 4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. https://sdgs.un.org/goals/goal4

United Nations. (2022, 21 April). *Sustainable development goals: 4 quality education*. https://www.un.org/sustainabledevelopment/education/

United Nations Brundtland Commission. (1987). *Report of the World Commission on Environment and Development: Our common future*. http://www.un-documents.net/our-common-future.pdf

United Nations Educational, Scientific and Cultural Organization. (2016). *Education 2030: Incheon Declaration and Framework for Action for the implementation of Sustainable Development Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000245656

United Nations Educational, Scientific and Cultural Organization. (2021). *Reimagining our futures together: A new social contract for education.*

United Nations Educational, Scientific and Cultural Organization. (2024). *Global report on teachers: Addressing teacher shortages and transforming the profession*. https://unesdoc.unesco.org/ark:/48223/pf0000388832

Valentine, N., Durning, S., Shanahan, E. M., & Schuwirth, L. (2021). Fairness in human judgement in assessment: A hermeneutic literature review and conceptual framework.

Advances in Health Sciences Education, 26, 713 738. https://doi.org/10.1007/s10459-020-10002-1

* Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical, psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952. DOI: 10.1037/a0025125

Welsh Government. (2009). *Professional standards for teaching and leadership*.

Welsh Government. (2017a). *Criteria for the accreditation of initial teacher education programmes in Wales: Teaching tomorrow's teachers.*

Welsh Government. (2017b). *Education in Wales: Our national mission. Action plan 2017–21.*

Welsh Government (2019). *Professional standards for teaching and leadership.* https://hwb.gov.wales/api/storage/19bc948b-8a3f-41e0-944a-7bf2cadf7d18/professional-standards-for-teaching-and-leadership-interactive-pdf-for-pc.pdf

Welsh Government (2021). *Credit and qualifications framework for Wales.*

Wiliam, D. (2023, 31 January). Teacher quality: What it is, why it matters, and how to get more of it. *Impact*. https://my.chartered.college/impact_article/teacher-quality-what-it-is-why-it-matters-and-how-to-get-more-of-it/

Withers, J. (2023, May). *Fit for the future: Developing a post-school learning system to fuel economic transformation*. Scottish Government. https://www.gov.scot/publications/fit-future-developing-post-school-learning-system-fuel-economic-transformation/documents/

Woodhouse, L. D., Auld, M. E., Miner, K., Alley, K. B., Lysoby, L., & Livingood, W. C. (2010). Crosswalking public health and health education competencies: Implications for professional preparation. *Journal of Public Health Management and Practice*, *16*(3), 20–28.

World Bank Group. (2022, 30 August). *Teach primary: Helping countries to measure effective reaching practices*. https://www.worldbank.org/en/topic/education/brief/teach-helping-countries-track-and-improve-teaching-quality

Wyatt-Smith, C., Adie, L., & Harris, L. (2024). Supporting teacher judgement and decision-making: Using focused analysis to help teachers see students, learning, and quality in assessment data. *British Educational Research Journal*, *50*(3), 1420–1448.

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 35–52.

Wyatt-Smith, C. M., & Looney, A. (2016). Professional standards and the assessment work of teachers. In D. Wyes, L. Hayward, & J. Pandya (Eds.) *The SAGE handbook of curriculum pedagogy, and assessment*, Vol. 2 (pp. 805–820). Sage.

Yacek, D. W., & Jonas, M. E. (2023). Phronesis in teacher education: A critical reexamination. *European Journal of Teacher Education*, 1–17. https://doi.org/10.1080/02619768.2023.2253495

* Yahiji, K., Otaya, L. G., & Anwar, O. (2019). Assessment model of student field practice at faculty of Tarbiyah and teaching training in Indonesia: A reality and expectation.

International Journal of Instruction, *12*(1), 251–268. https://doi.org/10.29333/iji.2019.12117a

Yazan, B. (2015). Three approaches to case study methods in education: Yin, Merriam, and Stake. *The Qualitative Report*, *20*(2), 134–152. https://nsuworks.nova.edu/tqr/vol20/iss2/12

Yin, R. K. (2014). *Case study research: Design and method* (5th ed.). Sage.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

* Yinger, R. J., & Daniel, K. L. (2010). Accreditation and standards in teacher education. *International Encyclopedia of Education* (3rd ed.), 495–502. DOI: 10.1016/B978-0-08-044894-7.00643-6

Young, M.D., & Diem, S. (2018). Doing critical policy analysis in education research: An emerging paradigm. In C. Lochmiller (Ed.), *Complementary research methods for educational leadership and policy studies* (pp. 79–98). Palgrave Macmillan.

Zabek, F., Lyons, M.D., Alwani, N., Taylor, J.V., Brown-Meredith, E., Cruz, M.A. & Southall, V.H. (2023). Roles and functions of school mental health professionals within comprehensive school mental health systems. *School Mental Health*, 15, 1–18. https://doi.org/10.1007/s12310-022-09535-0

Zeichner, K. M., & Bier, M. (2015). Opportunities and pitfalls in the turn toward clinical experience in U.S. teacher education. In E. Hollins (Ed.), *Rethinking field experiences in preservice teacher preparation: Meeting new challenges for accountability*. Routledge.

Appendix 2.1

Video Task & Questionnaire

Informed Consent

Title of Project: Reliability and consistency in judging new teacher practices – why does it matter?

Name of Researcher: Sarah K Anderson, PhD- Senior Lecturer

You are being invited to take part in a research study. Before you decide to take part it is important for you to understand why the research is being done and what it will involve. Please read the following information carefully and discuss it with others if you wish. Ask the researcher/s if there is anything that is not clear or if you would like more information. Take some time to decide whether or not you wish to take part. Thank you for reading this.

The purpose of the study:

The purpose of this study is to better understand how judgements of teaching effectiveness are made in initial teacher education (ITE), to enable more accurate assessment of teaching capabilities, to reimagine the value and professional career trajectory of the 'teacher academic' as a reorientated role, and investigate impacting power dynamics amongst schools, local authorities, and ITEs. We ultimately aim to reframe efforts to produce high-quality teachers who deliver quality education for all. Participants will include teachers and university staff involved in judging ITE students' performance per teaching standards (i.e., school-based mentor teachers, associate tutors, and university staff). Participation is voluntary and will take approximately 30 minutes.

Why you have been asked to participate:

As an individual involved in teacher education, you have experience judging new teaching effectiveness; you have a valuable perspective of the practices of assessing novice teachers' skills during school placements.

What will happen if you agree to participate:

Participation in this study is voluntary. You will be asked to watch a 15 minute teaching video, and you will then be asked to judge the effectiveness of the teaching in each video according to levels of performance in the form of a questionnaire (i.e., proforma). You will also be asked to explain the process of how your judgement was made.

- All data collected in the project will be anonymised and participants referred to by pseudonyms.
- You may end participation at any time while watching the videos or completing the questionnaire.

- At the end of the questionnaire, you will be asked if you are willing to participate in a focus group to discuss results from the video observation and questionnaire. Once submitted, responses cannot be withdrawn as they are anonymous (with exception for those who choose to participate in a follow up focus group).
- Your decision to participate, or not, will have no effect on employment.
- Any personal information collected in the study will be destroyed once the project is complete. The project's end date (when we expect to have completed the full project and published the results) is 1 October 2024.
- You can ask for a written summary of results from the named researcher below.
- We will retain the anonymised research data for 10 years only with participant consent in line with University of Glasgow policy.
- Please note that confidentiality may not be guaranteed if researchers are contacted for results and due to the limited size of the participant sample.
- Confidentiality will be respected subject to legal constraints and professional guidelines.

What the data will be used for:

After analysis, results will be used to write several research articles and scholarly works that we hope will be of benefit to initial teacher education, future students, and ultimately to pupils learning. We also intend to share our findings with staff at the University of Glasgow and at conferences, both national and international.

This study has been funded through the national award of the Society for Educational Studies.

This project has been considered and approved by the College Research Ethics Committee. If you have any additional questions, please contact Dr Sarah K Anderson: sarah.anderson.3@glasgow.ac.uk. To pursue any complaint about the conduct of the research: contact Lead for Ethical Review, College of Social Sciences (socsci-ethics-lead@glasgow.ac.uk)

- I confirm that I have read and understood the Participant Information Sheet for the above study.
- I understand that my participation is voluntary.
- I acknowledge that participants will be referred to by pseudonym.

- I acknowledge that there will be no effect on my employment arising from my participation or non-participation in this research.

I agree that:

- All names and other material likely to identify individuals will be anonymised.
- The material will be treated as confidential and kept in secure storage at all times.
- The material will be retained in secure storage for use in future academic research.
- The material may be used in future publications, both print and online.
- I waive my copyright to any data collected as part of this project.
- Other authenticated researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.
- I acknowledge the provision of a Privacy Notice in relation to this research project.

I have read the information sheet, and I agree to take part in this research study.

Background Information

Instructions: There are three parts to this questionnaire. First you will be asked to provide demographic information (e.g., current role, years of experience, etc.). Second, you will watch a 15 minute teaching demonstration video and determine a level of performance in seven areas of teaching. Finally, you will be asked to determine your level of agreement with statements about judging teaching effectiveness.

Current role in initial teacher education (ITE) (please select all that apply)

Member of staff in the School of Education

Associate Tutor

School-based Mentor Teacher

School Experience Tutor

Which of the following best describes your role at your current school? (please select all that apply)

Head teacher/principal

Depute head teacher

Head of department/faculty

Assistant head of department/faculty

Subject leader

Chartered teacher

Classroom teacher post-threshold

Classroom teacher

Newly qualified teacher

Supply teacher

Teaching assistant

Other (please specify)

Gender

Female

Male

Non-binary / third gender

Prefer not to say

How many years of work experience in education do you have? (please do not include voluntary or internship experiences)

How many years of experience working in your current role?

Which of the following routes into teaching did you take?

- Undergraduate degree with teaching qualification
- Postgraduate certification in education
- No qualifications
- Other route (please specify)

Which of the following teaching qualification(s) do you hold (please select all that apply)?

- Nursery
- Primary
- Secondary
- Specialist
- None
- Other (please specify)

Where did you obtain your teaching qualification?

- Scotland
- England
- Wales
- Northern Ireland
- Other country outside the UK (please specify)

What is the highest level of formal education you have completed?

- Qualification below the bachelor level
- Bachelor degree or equivalent
- Post-graduate diploma or certification
- Master degree or equivalent
- Doctoral degree or equivalent

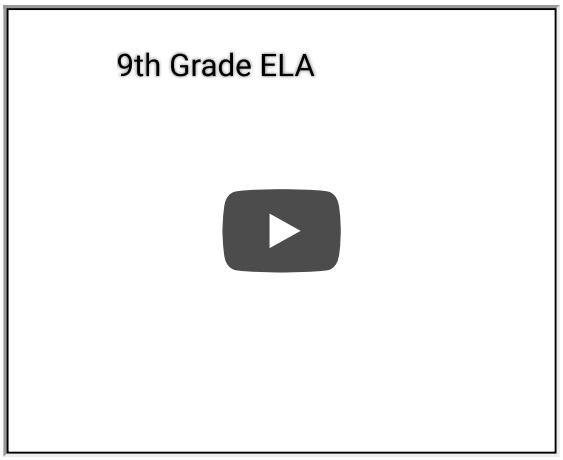
Teaching Vignette and Observation Video

Task Instructions: This task is a simulation of a student teacher classroom observation that would occur during a school-based experience in teacher preparation. Your task is to observe the student teacher and evaluate their teaching. You will watch a 15-minute video and then be asked to provide an overall judgement of the teaching effectiveness; you will then be asked to explain how and why you made that decision. You may take notes as you watch the video to reference in your explanation. You will determine a level of performance in seven areas:

- Learners
- Content
- Research
- Planning and preparation
- Instructional strategies
- Learning environment

- Assessment

Scenario: The video clip is of Emily Jones, a student teacher who is nearing the end of her final school placement of her 4-year Initial Teacher Education programme. The lesson you will watch to judge effective teaching is set in an English Language Arts class in America with 14-15 year old pupils.



To view in full screen, click on "Youtube" or click on the link https://youtu.be/Jyh3M8SCB3M

Video Questionnaire

1. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Learners: The teacher's practice shows understanding of learning and development and individual variations within and across the cognitive, linguistic, social, emotional, and physical areas. Regards the needs of all individuals and the class as a whole. Learning experiences are developmentally appropriate and intellectually challenging.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Learners**?

2. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Content: The teacher's practice demonstrates core knowledge and skills of the content area being taught. Learning experiences make the subject matter accessible and meaningful to learners to assure mastery of the content.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Content**.

3. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Research: The teacher's practice reflects core research and analytical methods that apply in teaching, including with regard to effective assessment of learners.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Research**.

4. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Planning & Preparation: The teacher's practice demonstrates that planning and preparation occurred which supports learners in reaching identified learning objectives.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Planning & Preparation**.

5. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Instructional Strategies: The teacher's practice includes an appropriate range of teaching activities which reflect and align with both the nature of the subject content being taught, and the learning, support, and development needs of the learners. Instruction facilitates engagement and integration of digital technologies.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Instructional Strategies**.

6. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Learning Environment: The teacher's practice demonstrates organisation and facilitation of learners' activities so they can participate constructively in a safe and secure environment and cooperative manner. The learning environment encourages positive social interaction, active engagement in learning, and self-motivation.	○	○	○	○	○

Please explain how you decided on the level of performance for the element of **Learning Environment**.

7. Assign a judgement of the teaching performance where (5) is Highly Effective and (1) is Unsatisfactory.

	5	4	3	2	1
Assessment: The teacher's practice demonstrates consistent, fair, valid, and reliable assessment of student learning using an appropriate range of methods to evaluate attainment of learning objectives.	○	○	○	○	○

How did you decide what level of performance was demonstrated for the element of **Assessment**?

8. What is your overall judgement of the teaching demonstrated in the video where (5) is Highly Effective and (1) is Unsatisfactory?

	5	4	3	2	1
Final overall rating	○	○	○	○	○

9. Which was the most difficult element to judge for the teaching in this video and why?

- Learners
- Content
- Research
- Planning & Preparation
- Instructional Strategies
- Learning Environment
- Assessment

Please explain why it was the most difficult element to judge.

10. Which was the easiest element to judge for the teaching in this video and why?

- Learners
- Content
- Research
- Planning & Preparation

Instructional Strategies

Learning Environment

Assessment

Please explain why it was the easiest element to judge.

Judging teaching effectiveness

Next you will be asked to respond to a series of questions about how judgements of teaching effectiveness are made.

11. When making judgement of teaching effectiveness, I...

start from a point of failure and look for instances to challenge that decision.

look for strengths first and then weigh these against identified weaknesses, reflecting on if the positives are more important than the negatives.

consider the teaching demonstrated against the learning outcomes based on teaching standards.

Other (please specify)

12. Rate your level of agreement or disagreement with the following statements about judging teaching effectiveness.

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Dis
a. It is important that judgements of teaching effectiveness are accurate.	○	○	○	○	○	
b. It is important that judgements of teaching effectiveness are consistent.	○	○	○	○	○	
c. It is important that different evaluators reach consensus.	○	○	○	○	○	
d. It is important that evaluators use evidence to make judgements.	○	○	○	○	○	

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Dis
e. It is important that professional judgement is used when judging teaching effectiveness.	○	○	○	○	○	

13. Rate your level of agreement or disagreement with the following statements about judging teaching effectiveness.

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Dis
a. Is is important that judgements about teaching effectiveness are made by more than one evaluator.	○	○	○	○	○	
b. It is important that potential sources of evaluator error are addressed.	○	○	○	○	○	

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Dis
c. It is important for the teacher to understand how judgements about their teaching effectiveness are made.	○	○	○	○	○	(
d. Judgements are always related to particular teachers at particular points in time and in particular situations.	○	○	○	○	○	(
e. It is important that judgements about teaching effectiveness are considered fair by stakeholders.	○	○	○	○	○	(

14. Rate your level of agreement or disagreement regarding factors which may influence how evaluators judge teaching.

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Di
a. Clarity of the judgement criteria	○	○	○	○	○	
b. Tension of using judgements for both professional growth and accountability	○	○	○	○	○	
c. Clarity of procedures for making judgements	○	○	○	○	○	
d. Individual understanding of effective teaching	○	○	○	○	○	
e. Contested nature of what defines effective teaching	○	○	○	○	○	
f. Professional teaching standards	○	○	○	○	○	
g. Power relationships between universities and schools in teacher education	○	○	○	○	○	

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Di
h. Personal intuition about what happens in a classroom	○	○	○	○	○	
i. Perceived levels of importance of different dimensions of teaching	○	○	○	○	○	
j. Complexity of the classroom environment in which judgements are made	○	○	○	○	○	

15. Rate your level of agreement or disagreement regarding factors which may influence how evaluators judge teaching.

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disa
a. Evaluator tendencies toward leniency or severity	○	○	○	○	○	(

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disa
b. Personal biases and beliefs of the evaluator	○	○	○	○	○	(
c. Experiences of the evaluator from observing other teachers	○	○	○	○	○	(
d. Prior interactions between the teacher and the evaluator	○	○	○	○	○	(
e. Holding a pre-observation discussion	○	○	○	○	○	(
f. Level of involvement of the individual being evaluated in the judgement process	○	○	○	○	○	(

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disa
g. Training of evaluators to use observation criteria for making judgements	○	○	○	○	○	(
h. Observation skills of the evaluator	○	○	○	○	○	(
i. Perceptual information (cues) available to the evaluator	○	○	○	○	○	(
j. Policies regarding evaluation of teaching effectiveness	○	○	○	○	○	(
k. Quality of the reasoning strategies used to make decisions	○	○	○	○	○	(

16. Why does it matter that judgements of teaching effectiveness are consistent and reliable?

Focus Group

Online focus groups are being arranged to discuss responses and initial results of the study; this is an essential part of the project to develop a deeper understanding of how judgements are made. Participation in a focus group is voluntary, 45-minute in length, will be conducted using a digital conferencing system (e.g., Zoom). The full participant information sheet is available here.

Are you willing to participate in a focus group to discuss your response and initial results of the study?

Yes

No

Please provide your name, email address, and phone number to arrange participation in the focus group.

Appendix 2.2

Ethical Approval

University of Glasgow

College of Social Sciences

21 November 2022

Dear Sarah K. Anderson

College of Social Sciences Research Ethics Committee

Project Title: Reliability and consistency in judging new teacher practices – why does it matter?

Application No: 400220088

The College Research Ethics Committee has reviewed your application and has agreed that there is no objection on ethical grounds to the proposed study. It is happy therefore to approve the project, subject to the following conditions:

- Start date of ethical approval: 12/12/2022
- Project end date: 01/10/2024
- Any outstanding permissions needed from third parties in order to recruit research participants or to access facilities or venues for research purposes must be obtained in writing and submitted to the CoSS Research Ethics Administrator before research commences: socsci-ethics@glasgow.ac.uk
- The research should be carried out only on the sites, and/or with the groups and using the methods defined in the application.
- The data should be held securely for a period of ten years after the completion of the research project, or for longer if specified by the research funder or sponsor, in accordance with the University's Code of Good Practice in Research: (https://www.gla.ac.uk/media/media_490311_en.pdf)
- Any proposed changes in the protocol should be submitted for reassessment as an amendment to the original application. The **Request for Amendments to an Approved Application** form should be used: https://www.gla.ac.uk/colleges/socialsciences/students/ethics/forms/staffandpostgraduateresearchstudents/

Yours sincerely,

Dr Susan A. Batchelor
College Ethics Lead

Susan A. Batchelor, Senior Lecturer
College of Social Sciences Ethics Lead
University of Glasgow
School of Social and Political Sciences &
Scottish Centre for Crime and Justice Research
Ivy Lodge, 63 Gibson Street, Glasgow G12 8LR.
0044+141-330-6167 socsci-ethics-lead@glasgow.ac.uk

Appendix 2.3

Case Study Protocol

University of Glasgow

School of Education

Protocol: SES National Award – Reliability and consistency in judging new teacher practices – why does it matter?

University of Glasgow

School of Education

A. Overview of the Study

1. Role of Protocol: This protocol serves to guide the research team with general procedures and plans to be followed. The protocol was created to assist the team in anticipating problems as well as keep the team targeted on the topic. It also is a way of increasing reliability of the research and guide data collection for a single case and across multiple cases. The protocol includes five main sections: overview, methods, data collection procedures, data collection questions, and a guide for the final report (Yin, 2018, pp. 84–94).

2. Goals of the Study: Explore the nature of judgements regarding initial teacher education (ITE) students' performance per normed teaching standards.

3. Purpose of the Study: The project seeks to enable more accurate judgements to positively affect teacher capacities, to reimagine the value and professional career trajectory of the 'teacher academic' as a reorientated role, and investigate impacting power dynamics amongst schools, local authorities, and ITEs. We seek to uncover the decision-making process used by individuals who judge teacher candidates' readiness to teach, a detailed investigation of what the judges specifically look for in order to make their decisions. We ultimately aim to reframe efforts to produce high-quality teachers who deliver an 'inclusive and quality education for all' (United Nations, 2022).

Contribution

- To better understand judgement processes to improve judgement-making of teaching effectiveness
- Directly influence the practices of assessing and enhancing novice teachers' skills in clinical school placements with the ultimate goal of enhancing pupil outcomes
- Expand opportunities for dialogue across systems through a renewed sharing of practices, policies, and professional standards
- To meet the shared responsibility of training high-quality future educators in a sustainable model, foster networked improvement communities, and inform perspectives beyond Great Britain
- Build partnerships with teachers, researchers, and university staff at the University of Glasgow (UofG), Leeds Beckett University and Aberystwyth University as well as across educational leaders in Scotland, England, and Wales
- Map the power relationships between schools, local authorities, and ITEs
- Development of shared medium- and long-term goals between ITEs and their school partners
- Shaping the processes and practices of partnered clinical education experiences in ITE
- Stimulate discussion around policy and accreditation guidelines for high-quality clinical partnerships and practice
- Consistency in the quality of teacher candidates finishing ITE
- Initiate reorientation programmes promoting mentoring as a collaborative process and they recommend that clarifications of expectations of aims of partnership (between higher education institutions [HEIs] and schools) and expectations regarding roles of mentors, students, HEIs be communicated and shared
- Collaborative support of initial teacher development as an expression of public solidarity for the future of teacher education
- Support research that incorporates co-construction, teachers as reflexive practitioners and knowledge producers

4. Research Questions:

1. What is the nature of shared judgement, consensus, and dissensus of observed teaching effectiveness amongst university staff, associate tutors, and school-based mentor teachers from partner ITE programmes?

2. How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?
3. How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

5. Propositions: On the basis of the literature and programme evidences, we expect the findings of the research to align with the following propositions.

R1: What is the nature of shared judgement, consensus, and dissensus of observed teaching effectiveness amongst university staff, associate tutors, and school-based mentor teachers from partner ITE programmes? **Proposition 1:** Link tutors and school-based mentor teachers – we expect to see dissensus, although in practice this is a negotiated consensus through the tripartite conversation; University staff and school-based mentors: we expect staff to make more clinical, criterion-based judgements, whereas mentor teachers will consider a more holistic viewpoint in making their judgement decisions

R2: How might enhanced reliability of professional judgement foster greater collaboration between schools and universities? **Proposition 2**: In order to have greater reliability we need collaboration. We anticipate there to be more precise actions and clearer formative assessment and actions that support student teacher development. We anticipated a better aligned and shared understanding of the standards.

R3: How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics? **Proposition 3:** We expect to hear a dialogue of joint accountability which in practice still places the university in a more powerful role.

6. Background Information: At a moment in history when professional disassociation of educators has been prolific, this project seeks a coming together towards a common understanding of effective teaching and assessment in ITE. The changing shape of teacher education requires a richer understanding of the nature of judging new teaching effectiveness. The transformative aspects of this research relate to the degree to which established norms are challenged in three key aspects: how classroom-based mentor teachers judge ITE students' performance per normed teaching standards, who institutions rely on to judge teaching effectiveness (i.e., school-based mentor teachers, associate tutors, and university staff), and how ITEs use concomitant judgements of teaching effectiveness amongst a context of power dynamics.

This project aims to be transformative for participants involved in the project through: sharing best practices, peer observation, developing new knowledge, building networked relationships, deepening commitments to reliability in judgements (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2019), reconciling tensions and dilemmas given multiple positions on problems both old and new, and taking responsibility for decisions where the resulting actions impact broader benefit of others.

The project also seeks to alter the role of the associate tutor, which has largely been lessened in academia, to create new value of the 'teacher academic' as the boundary spanner in teacher education – the connector of university to school and theory to practice. Willingness for ITEs to reconsider their role as intermediary organizations to coordinate relationships amongst stakeholders and lend social capital to teacher educators positioned within the central space is an opportunity for improvement. The project is an occasion for ITEs to demonstrate adaptability and change approaches to shared judgement based on emerging insights.

These transformational aspects aim to adjust the way ITEs engage in the joint process of teacher training with classroom teachers and associate tutors in a spirit of co-agency. Teacher participation in the project is an opportunity to empower teachers to assume a more active, responsible, and effective role in the ITE process. As la Velle stated (2020), 'Learning to teach is transformative, complex and life-long', and classroom teachers are at the centre of that process for novice teachers, re-emphasizing the unique and valued role that teachers hold in the social contract of education (UNESCO, 2021). We recognize that change is rarely linear or smooth, and learning to teach is itself a transformational process.

7. Conceptual Framework: We consider the nature of judgements regarding ITE students' performance per normed teaching standards as socially constructed and fundamentally situated; therefore judgements must be understood in context. Social judgement theory (SJT), which emphasizes careful identification and analysis of the context of judgement, aptly supports and informs the project design (Cooksey, 1988, 1996; Hammond et al.,1977; Hovland & Sherif, 1980). Social judgement theory highlights the indicators and guidelines used by judges, making it a fitting framework from which to investigate the decisions associate tutors, university staff, and mentor teachers make in multifaceted and dynamic learning situations in each of the three ITE contexts. The theory recognizes that professional judgement is a distinctly cognitive act as well as a socially positioned practice (Allal, 2013). Judgement of new teacher performance will be dependent on what evaluators *think* about effective teaching and the level of performance of required knowledge, skills, and dispositions of normed teaching standards they find acceptable. Additionally, teaching standards themselves are socially constructed within a larger social, economic, and political narrative of teacher education (Cochran-Smith, 2003), and attempts to understand how ITE students are judged requires consideration of underlying constructs.

The nature of judgements regarding ITE students' performance in this project will be investigated through the eight stages suggested by SJT (Cooksey, 1996).

1. **Conceptualizing the judgement problem:** the decisions that teacher educators (university and school based) make about teacher candidates' teaching effectiveness are highly variable and idiosyncratic; understanding decision-making as it unfolds in a currently performed task as well as establishing values or perspectives on future judgements yet to be made.

2. **Understanding context:** understanding of the conditions and circumstances (including criterion measures and power dynamics) under which judgements are made – the decision task environment; identifying the kinds of cue information found useful by experienced judges and any themes or criterion measures (e.g., visual and auditory cues from pupils and the teacher, criterion measures which can facilitate comparisons designed to highlight judgement activities).

3. **Identifying evidences and dimensions for judgement:** focusing down to establish a smaller set of the most potentially relevant cues when making judgements (what occurs on a teaching video and could reasonably be included in the observation questionnaire – themes from the common teaching standards); evaluation of suitable or unsuitable and then a level of suitability.

4. **Determining a sample of indicator profiles:** selecting videos of teaching representative of a classroom-based observation that would be carried out during teacher training to elicit judgements from the participants; this is a *representative design* (one 15–20 minute video with accompanying, standardized, contextual vignette); utilizing a formal situational sample ('it is possible to accomplish substantive situational sampling, particularly for applied judgement research, by randomly or systematically sampling actual cases that have been judged in the recent past'; Cooksey, 1996).

5. **Sampling participating judges:** the sample of judges must reflect the various roles of individuals who conduct observations of teaching effectiveness (i.e., link tutors, associate tutors, university staff, and mentor teachers); purposive sampling due to roles; to be aware of experiential background in the task and how it might impact on the participants' capacity to cope with the observation task requirements, we are collecting demographic information related to experience.

6. **Obtaining judgements:** judgements will be obtained through video stimuli – this adds the advantage of realism to the video scenario being judged.

7. **Capturing individuals' judgement policies:** descriptive statistics, multiple regression methods – logistical regression procedures, qualitative thematic analysis.

8. **Compare policies:** systemic influences on judgement may arise from potentially controllable factors (e.g., distractions, memory); levels of agreement; any predictability in the ways judgements are made; different weighting of cues.

This staged framework will enable us to capture, question, and compare the nature of judgement decisions and strategies used by participants as they determine readiness to teach and frame the wider conversation about the shared responsibility of training high-quality future educators.

While the suitability of SJT to this project is clear, critiques are essential to address potential weaknesses in project design and to inform methodological choices. There is a concern that SJT is too simplistic to take into account the myriad effects of variables on judgement – for example, interpretations of evidences, the quality of an argument, an individual's position on/involvement in a particular issue, as well as source credibility (O'Keefe, 2015). Additionally, variability in human nature amongst those involved in teacher education is a factor in this socially influenced process. To address these concerns, questionnaire-based data collection will include open-ended responses exploring justification of judgements, and focus groups will occur to confirm responses, consider individual positions, and explore how decisions are reasoned. Demographic information will also be collected to describe the participant positionality (e.g., years of teaching experience, degrees/certificates, educational roles held). Furthermore, the Delphi panel process builds in reciprocal attempts at understanding and allows for distanciation from respective roles possibly shaped by power dynamics, further enhancing the opportunity for deeper insights to emerge.

SJT draws attention to the possibility that judges with nearly the same position on an issue might still have a different valuation, and the importance of variation in evaluator involvement. It also draws forward consideration of the value of dissensus and the potential role of a pluralistic approach (Moss & Schutz, 2001); situated amongst impacting power dynamics, we have the prospect for deeper understanding when learning from differences. Through the lens of SJT, we aim to better understand judgement processes and enhance the reliability of how judgements are made.

8. Relevant Readings: The following readings informed the decision to conduct the study and informed the protocol design process.

- Baumfield, V. M., Conroy, J. C., Davis, R. A., & Lundie, D. C. (2012). The Delphi method: Gathering expert opinion in religious education. *British Journal of Religious Education, 34*(1), 5–19.
- Council of Chief State School Officers. (2013). *InTASC model core teaching standards and learning progressions for teachers 1.0*.
- Department for Education. (2021). *Teachers' standards: Guidance for school leaders, school staff and governing bodies.*
- Education International & UNESCO. (2019). *Global framework of professional teaching standards*.
- Elmore, R. F. (2007). Professional networks and school improvement: The medical rounds model, applied to K-12 education, provides a community of practice among superintendents committed to better instruction. *School Administrator, 64*(4).
- General Teaching Council for Scotland. (2021). *The standard for provisional registration*.
- Hattie, J. (2022). Visible learning. https://visible-learning.org/
- Welsh Government. (2019). *Professional standards for teaching and leadership*.

B. Methods

1. Case Study Research: The project will use a comparative, embedded, and descriptive multiple-case study design. A mixed methods approach will guide data collection and analysis and a cross-case synthesis will be conducted (Yin, 2018). A combination of the case study approaches of Merriam & Tisdell (2016) and Yin (2018) were used to guide the study design; this occurred because the design rigour of Yin and the constructivist-education epistemological approach of Merriam & Tisdell complement each other in a way that meets the need of research considering judgement (Yazan, 2015).

2. Methodology:

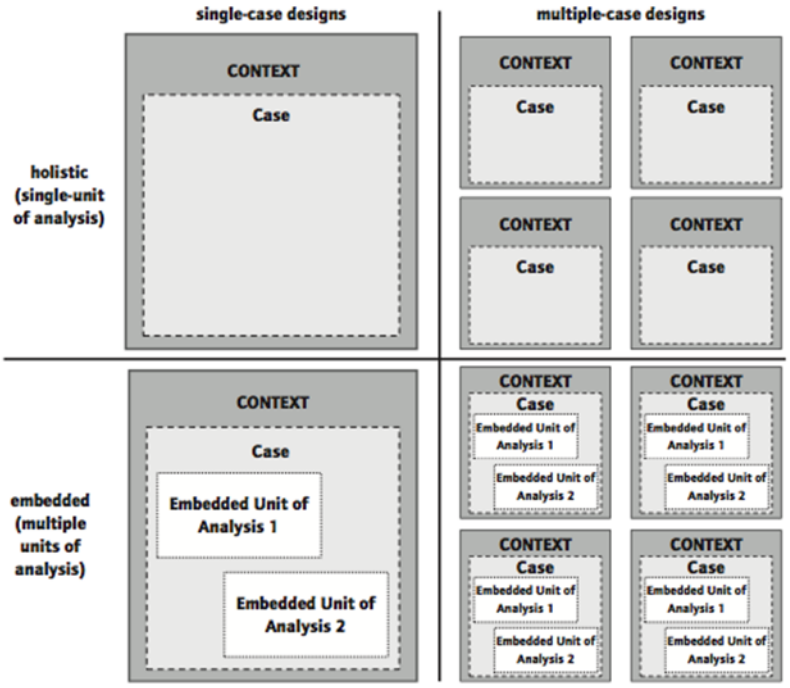
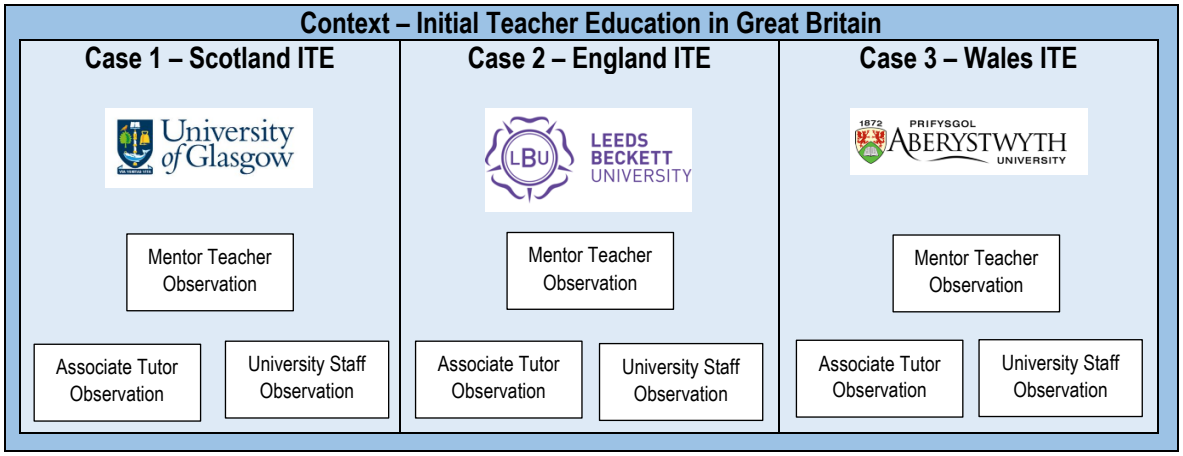


Figure 2.4 Basic Types of Designs for Case Studies
SOURCE: COSMOS Corporation.

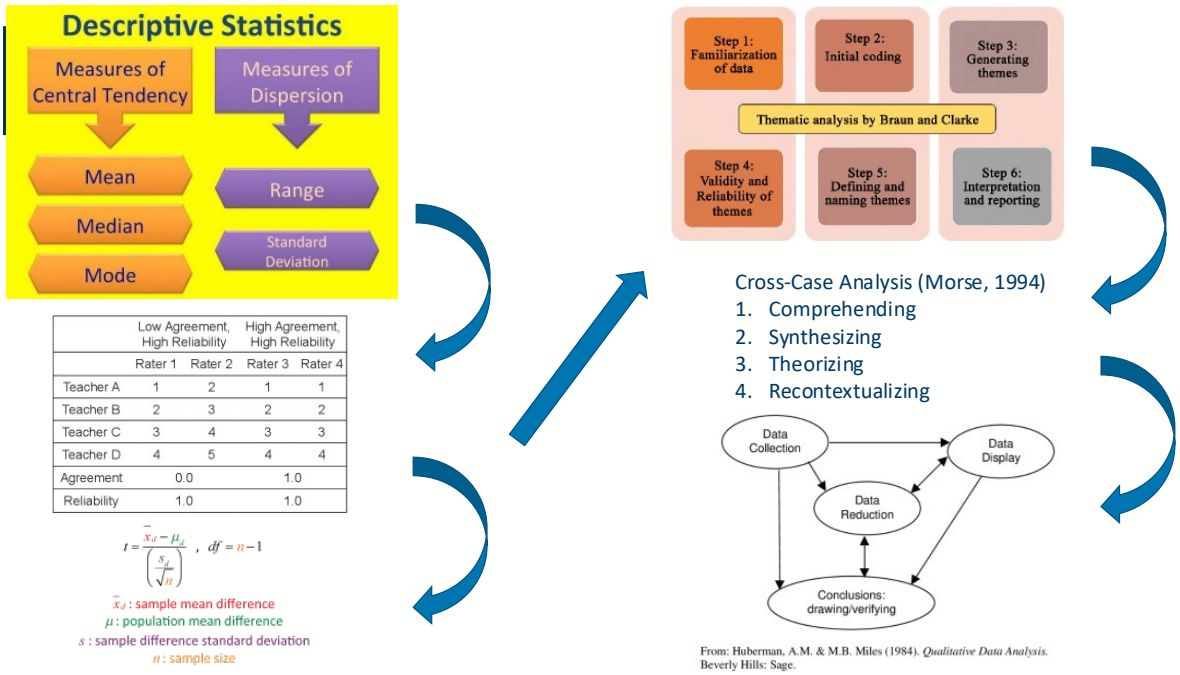


3. Participants: Participants will be selected through purposeful sampling, as the groups in question demonstrate a perspective within a defined context and with enough information for in-depth exploration (Merriam & Tisdell, 2016). Participants include full-time university staff, associate tutors, and school-based mentor teachers from each ITE programme, and the goal is to include 30 participants in each role at each location (n = 270).

4. Analysis Protocol:

University of Glasgow

School of Education



	Low Agreement, High Reliability		High Agreement, High Reliability	
	Rater 1	Rater 2	Rater 3	Rater 4
Teacher A	1	2	1	1
Teacher B	2	3	2	2
Teacher C	3	4	3	3
Teacher D	4	5	4	4
Agreement	0.0		1.0	
Reliability	1.0		1.0	

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}, \quad df = n - 1$$

$\bar{x}_d$: sample mean difference
μ : population mean difference
s : sample difference standard deviation
n : sample size

From: Huberman, A.M. & M.B. Miles (1984). *Qualitative Data Analysis.* Beverly Hills: Sage.

5. Single-Case Analysis: Quantitative data will include scaled observation ratings of teaching effectiveness. Descriptive statistics will be used to analyse the data set and examine variance across perspectives (e.g., frequencies, response agreement percentages, means, and standard deviations; Pyrczak & Oh, 2018). Inter-rater reliability will be calculated including percent agreement and a trend analysis. Comparative analyses will examine patterns of consensus and dissensus. A paired *t* test will determine if there is a significant difference between group means.

Within each case, qualitative data will be analysed using the constant comparative method (Glaser & Strauss, 1967) to construct codes, categories, subcategories, or themes. Computer-assisted qualitative data analysis software will be utilized to begin the coding process. Guidelines on thematic analysis (Braun & Clark, 2006) will be used to ensure reliability.

6. Cross-Case Analysis: Cross-case data analysis will occur using the four-stage framework of Morse (1994) – comprehending, synthesizing, theorizing, and recontextualizing – integrated with the analysis strategies of Miles and Huberman (1994) – broad coding, pattern coding, memoing, distilling and ordering, testing executive summary statements, and developing propositions (Houghton et al., 2015, p. 10).

Table 1 Stages of data analysis

Stages of analysis Morse (1994)	Analysis strategies (Miles and Huberman 1994)	Purpose
1 Comprehending	Broad coding	General accounting scheme that is not specific to content but points to the general domains in which codes can be developed inductively.
2 Synthesising	Pattern coding Memoing	Explanatory, inferential codes to create more meaningful analysis. 'One of the most useful and powerful sense-making tools at hand' (Miles and Huberman 1994).
3 Theorising	Distilling and ordering. Testing executive summary statements	Memos tie together different pieces of data into a recognisable group of concepts. 'Building towards a more integrated understanding of events, processes and interactions in the case' (Miles and Huberman 1994).
4 Recontextualising	Developing propositions	Formalise and systemise into a coherent set of explanations.

7. Trustworthiness/Credibility: To ensure the quality of the case study as educational research, a number of design strategies will be used and referred to during the study.

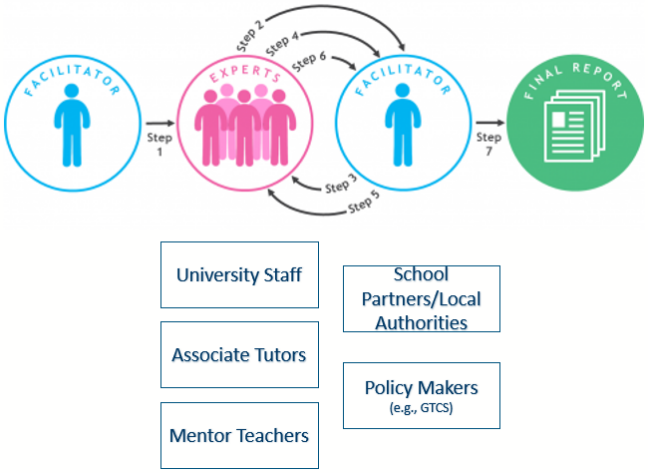
- adherence to a case study protocol developed from best practices in educational research (reliability)
- open-ended questions – justification for ratings
- maintenance of a case study database by a researcher who does not interact with participants or analyse data
- use of replication logic at each ITE
- focus group data collection by a researcher neutral to the ITE training process
- coding prior to analysis to the extent possible for anonymity of participants
- member checking

C. Data Collection Procedures

1. Data Collection Plan: Data collection for focus groups for each ITE will be conducted by a member of the research team or research assistant who is not employed by said ITE. A plan for each of the sources of data being collected at each ITE is outlined in the following timeline and procedures.

a. Video observation and questionnaire: we aimed to provide a common, authentic task for participants to respond to, one which result in a detailed account of judges decisions. The task needed to be applicable in all the judges settings and allow for contextual factors to be expressed. A structured but open-ended task was deemed appropriate.
    i. Analysis: thematic coding (Braun & Clare, 2006) will be used to interpret the data, providing the necessary overview of how judgements are made and their relative strength and importance. We will look at how important each theme is by looking at the frequency of occurrence and the way it was prioritized. This is important in understanding the weight of different evidences, which is a common source of disagreement amongst judges. Acceptable degree of agreement reach for the codes.
b. Focus groups
c. Delphi panel
    - Researchers define a problem and develop related questions.
    - Researchers select a panel of diverse experts (whose anonymity is generally protected).

- Researchers distribute the questionnaire to the panel.
- Researchers analyse and summarize the data and develop follow-up questions.
- Repeat step 4 as required.
- As consensus begins forming and issues are clarified, repeat step 4 as required.
- Researchers invite panellists to revise or review consensus and specify reasons for dissenting opinion.
- Repeat steps 4–7 as required.
- Researchers summarize consensus and provide feedback to the panel.
- Researchers publish the final consensus statement.



2. Data Recording and Storage: All data will be stored in a password-protected digital filing system via the UofG OneDrive. Files with data will be stored on the desktops of the researchers that are protected by the UofG (SSD) and in a joint Microsoft Teams folder, and cannot be accessed by anyone outside of the team. Focus group transcripts will be identified by a code rather than the subjects' name. The code list and transcripts will be stored separately. Data will be retained for 10 years after analysis is complete. Digital data will be stored on devices and data storage platforms (e.g., Office 365) that require password protection and dual authentication.

3. Confidentiality and Privacy: All data will be identified by a code and pseudonym rather than the participants' name or other identifying information. The code list and transcripts will be stored separately. Findings will be summarized without use of actual names. Participant demographic information and data will be coded/stored using a naming convention system.

ITE programme	UG, LBU, AU
Role	MT, AT, S
Participant	P1
Video observation	VO
Focus group	FG
Delphi panel	DP

4. Preparation: Because of the continuous interaction between theoretical concerns and data collection in this case study, a number of preparatory considerations will be addressed in order to execute the research design well.

d. *Research Training:* All research team members will have completed GDPR and Information Security requirements at their respective institutions.
e. *Ethics Review:* The project will be reviewed by ethics committees at each partner institution.

f. *Bracketing:* To avoid substantiating preconceived positions about participants, impact on P-12 learning, or impact of the EPP, bias of the researchers will be reduced through bracketing (Creswell, 2007). At a research team meeting, members of the team will discuss values, biases, or experiences about the topic that could influence how they collect, analyse, or report the data. Researchers will hold each other accountable for potential bias in their analyses and an auditing of preliminary findings will occur.

5. Limitations: Although the case study design is particularly situated for investigating a complex educational phenomenon and advancing the knowledge base, the following limitations have been identified:

g. Saturation of data for rich, thick description is limited due to purposeful selection of the participants and institutions.
h. The three cases in this study are only a portion of the ITE staff and school-based mentor teachers from each institution and are not necessarily representative or generalizable outwith the institutions.
i. Although efforts have been made to address bias, it remains an inherent issue in case study research.
j. This study only addresses a part of the whole decision-making process, a decision with a significant moral and ethical dimension embedded in long-established and complex processes and settings.
k. Limitations on seeing the entire sociocultural context.
l. Tripartite conversation is an essential feature of all three programmes but is not replicated in this study (future research – have all three evaluate together – consensus).
m. There is no critical examination of the context in this initial investigation of the decision-making process.
n. Choice of SJT over others – different conceptions of judging readiness would give rise to different ways of investigating.

D. Data Collection

1. What is the nature of shared judgement, consensus, and dissensus of observed teaching effectiveness amongst university staff, associate tutors, and school-based mentor teachers from partner ITE programmes?
2. How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?
3. How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

	Data source
Document review	National and International Teaching Standards
Video observation rating and justification	UofG mentor teachers
	UofG associate tutors
	UofG staff
Focus groups	UofG mentor teachers
	UofG associate tutors
	UofG staff
Delphi panel	UofG expert panel
	LBU mentor teachers

Video observation rating and justification	LBU link tutors
	LBU staff
Focus groups	LBU mentor teachers
	LBU link tutors
	LBU staff
Delphi panel	LBU expert panel
Video observation rating and justification	AU mentor teachers
	AU associate tutors
	AU staff
Focus groups	AU mentor teachers
	AU associate tutors
	AU staff
Delphi panel	AU expert panel
Symposium	Representatives of all three cases and different levels of power (approx. 50)

E. Guide for Final Reports

1. **Audiences:**
    a. SES Board Members
    b. ITE programmes
    c. National and international readership
    d. Policymakers
2. **General Format:**
    - Introduction
    - Research Questions
    - Conceptual Framework
    - Methods
    - Findings
    - Discussion
    - Conclusion
3. **Outputs:**
    a. **Reports** – to SES
    b. **Articles**
        i. article to the British Journal of Educational Studies
        ii. articles to journals at a national level for each respective nation
        iii. article to a journal of international standing
    c. **Presentations**
        - SES
        - international presentation(s)
        - national presentation(s)
    d. **Events**
        - collaborative event developed through CollectivED: The Centre for Mentoring, Coaching & Professional Learning at Leeds Beckett University

- symposium on transformative practices and reframing of teacher educator roles at the UofG Advanced Research Centre

References

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20–34.

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education*, *54*(1), 3–5.

Cooksey, R. W. (1988). Social judgement theory in education: Current and potential applications. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgement: The SJT view* (pp. 273–315). Elsevier.

Cooksey, R. W. (1996). The methodology of social judgment theory. *Thinking and Reasoning*, *2*(2), 141–174.

Hammond, K., Rohrbaugh, J., Mumpower, J., & Adelman, L. (1977). Social judgment theory: applications in policy formation. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes in applied settings* (pp.1-29). Academic Press.

Hovland, C. I., & Sherif, M. (1980). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Greenwood.

la Velle, L. (2020). Teacher education: The transformation of transitions in learning to teach. *Journal of Education for Teaching: International Research and Pedagogy*, *46*(2), 141–144.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research and case study applications in education*. Jossey-Bass.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications.

Morse, J. M. (1994). Emerging from the data: Cognitive processes of analysis in qualitative inquiry. In J. Morse (Ed.), *Critical issues in qualitative research* (pp. 23–43). Sage.

Moss, P., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal*, *38*(1), 37–70.

O'Keefe, D. J. (2015). *Persuasion: Theory and research* (3rd ed.). Sage.

United Nations. (2022, April 21). *Sustainable development goals: 4 quality education*. https://www.un.org/sustainabledevelopment/education/

United Nations Educational, Scientific and Cultural Organization. (2019). *Teaching and learning transformative engagement*.

United Nations Educational, Scientific and Cultural Organization. (2021). *Reimagining our futures together: A new social contract for education*.

Yazan, B. (2015). Three approaches to case study methods in education: Yin, Merriam and Stake. *The Qualitative Report, 20*(2), 134–152.

University of Glasgow

School of Education

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

Appendix 3.1

References for Studies in the Systematic Review

Appendix A3.1: Studies Included in the Systematic Review

1. Hylton, S. P., Joseph, J. D., Ward, T. J., & Gareis, C. R. (2022). Examining the validity of a student teaching evaluation instrument. *Teacher Educators' Journal*, *15*(1), 77–101.
2. Dewaele, J. M., Mercer, S., Talbot, K., & von Blanckenburg, M. (2021). Are EFL pre-service teachers' judgment of teaching competence swayed by the belief that the EFL teacher is a L1 or LX user of English? *European Journal of Applied Linguistics*, *9*(2), 259–282. DOI: 10.1515/eujal-2019-0030
3. Tobón, S., Juárez-Hernández, L. G., Herrera-Meza, S. R., & Núñez, C. (2021). Pedagogical practices: Design and validation of SOCME-10 rubric in teachers who have recently entered basic education. *Educational Psychology*, *27*(2), 155–165. https://doi.org/10.5093/psed2021a13
4. Tanguay, C. L. (2020). High-stakes assessment in elementary education teacher preparation: Educators' perceptions and actions resulting in curriculum change. *Education Policy Analysis Archives*, *28*(53), 1–39. DOI: 10.15407/epaa.4840
5. Sandoval, C., van Es, E. A., Campbell, S. L., & Santagata, R. (2020). Creating coherence in teacher preparation: Examining teacher candidates' conceptualizations and practices for equity. *Teacher Education Quarterly*, *47*(4), 8–32. https://www.jstor.org/stable/10.2307/26977528
6. Roloff, J., Klusmann, U., Lüdtke, O., & Trautwein, U. (2020). The predictive validity of teachers' personality, cognitive and academic abilities at the end of high school on instructional quality in Germany: A longitudinal study. *AERA Open*, *6*(1), 2332858419897884.
7. Shahzad, S., & Mehmood, N. (2019). Development of teaching effectiveness scale for university teachers. *Journal of Research in Social Sciences*, *7*(2), 1–14. DOI: 10.52015/jrss.7i2.74
8. Yahiji, K., Otaya, L. G., & Anwar, O. (2019). Assessment model of student field practice at faculty of Tarbiyah and teaching training in Indonesia: A reality and expectation. *International Journal of Instruction*, *12*(1), 251–268. https://doi.org/10.29333/iji.2019.12117a
9. Mkhasibe, R. G., Maphalala, M. C., & Nzima, R. D. (2018). Perceptions of subject mentors of pre-service teachers' readiness to teach economics and management sciences in the development of South Africa. *Journal of Gender, Information and Development in Africa*, *7*(2), 241–259. https://www.researchgate.net/publication/330779459
10. Basit, I., & Khurshid, F. (2018). Satisfaction of prospective teachers and teacher educators about the quality of teacher education programs. *Journal of Research in Social Sciences*, *6*(2), 168–188.
11. Ata, A., & Kozan, K. (2018). Factor analytic insights into micro-teaching performance of teacher candidates. *International Online Journal of Education and Teaching*, *5*(1), 169–178. http://iojet.org/index.php/IOJET/article/view/264/225
12. Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377–393. https://doi.org/10.1177/00224871177025

13. Kennedy, A. S., & Lees, A. T. (2016). Preparing undergraduate pre-service teachers through direct and video-based performance feedback and tiered supports in Early Head Start. *Early Childhood Education Journal*, *44*, 369–379. https://doi.org/10.1007/s10643-015-0725-2
14. Masuwai, A. M., & Saad, N. S. (2016). Evaluating the face and content validity of a Teaching and Learning Guiding Principles Instrument (TLGPI): A perspective study of Malaysian teacher educators. *Geografia*, *12*(3), 11–21. https://www.researchgate.net/publication/299265585
15. Brown, E. L., Suh, J., Parsons, S. A., Parker, A. K., & Ramirez, E. M. (2015). Documenting teacher candidates' professional growth through performance evaluation. *Journal of Research in Education*, *25*(1), 35–47.
16. Maharaj, S. (2014). Administrators' views on teacher evaluation: Examining Ontario's teacher performance appraisal. *Canadian Journal of Educational Administration and Policy*, 152, 1–58.
17. Kingsley, L., & Romine, W. (2014). Measuring teaching best practice in the induction years: Development and validation of an item-level assessment. *European Journal of Educational Research*, *3*(2), 87–109.
18. Hamid, S. R. A., Hassan, S. S. S., & Ismail, N. A. H. (2012). Teaching quality and performance among experienced teachers in Malaysia. *Australian Journal of Teacher Education*, *37*(11), 85–103. DOI: 10.14221/ajte.2012v37n11.2
19. Smalley, S., & Retallick, M. (2012). Agricultural education early field experience through the lens of the EFE model. *Journal of Agricultural Education*, *53*(2), 99–109. DOI: 10.5032/jae.2012.02099
20. Ritzhaupt, A. D., Ndoye, A., & Parker, M. A. (2010). Validation of the electronic Portfolio Student Perspective Instrument (EPSPI) conditions under a different integration initiative. *Journal of Computing in Teacher Education*, *26*(3), 111–119. DOI: 10.1080/10402454.2010.10784642
21. Beare, P., Torgerson, C., Marshall, J., Tracz, S., & Chiero, R. (2014). Examination for bias in principal ratings of teachers' preparation. *The Teacher Educator*, *49*(1), 75–88. DOI: 10.1080/08878730.2013.848005
22. Behizadeh, N., & Neely, A. (2018). Testing injustice: Examining the consequential validity of edTPA. *Equity & Excellence in Education*, *51*(3–4), 242–264. DOI: 10.1080/10665684.2019.1568927
23. Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment*, *23*(4), 229–249. http://dx.doi.org/10.1080/10627197.2018.1513788
24. Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools. REL 2014–024. Regional Educational Laboratory Mid-Atlantic. https://files.eric.ed.gov/fulltext/ED545232.pdf
25. Conderman, G., & Walker, D. A. (2015). Assessing dispositions in teacher preparation programs: Are candidates and faculty seeing the same thing? *The Teacher Educator*, *50*(3), 215–231. DOI: 10.1080/08878730.2015.1010053

26. Choi, H. S., Benson, N. F., & Shudak, N. J. (2016). Assessment of teacher candidate dispositions: Evidence of reliability and validity. *Teacher Education Quarterly*, *43*(3), 71–89.
27. Johnston, P., Wilson, A., & Almerico, G. M. (2018). Meeting psychometric requirements for disposition assessment: Valid and reliable indicators of teacher dispositions. *Journal of Instructional Pedagogies*, *21*. https://files.eric.ed.gov/fulltext/EJ1194249.pdf
28. Lazarev, V., Newman, D., Nguyen, T., Lin, L., & Zacamy, J. (2017). The Texas Teacher Evaluation and Support System rubric: Properties and association with school characteristics. REL 2018-274. Regional Educational Laboratory Southwest. https://files.eric.ed.gov/fulltext/ED576984.pdf
29. Lyness, S. A., Peterson, K., & Yates, K. (2021). Low inter-rater reliability of a high stakes performance assessment of teacher candidates. *Education Sciences*, *11*(10), 1–16. https://doi.org/10.3390/educsci11100648
30. Montecinos, C., Rittershaussen, S., Cristina Solís, M., Contreras, I., & Contreras, C. (2010). Standards-based performance assessment for the evaluation of student teachers: A consequential validity study. *Asia-Pacific Journal of Teacher Education, 38*(4), 285–300. DOI: 10.1080/1359866X.2010.515941
31. Murley, L. D., Stobaugh, R., Jukes, P., & Tassell, J. (2014). Examining the reliability of a culminating teacher education assessment and discovering areas for reform. *Educational Renaissance*, *2*(2), 3–18. DOI: 10.33499/edren.v2i2.61
32. Papanastasiou, E. C., Tatto, M. T., & Neophytou, L. (2012). Programme theory, programme documents and state standards in evaluating teacher education. *Assessment & Evaluation in Higher Education*, *37*(3), 305–320. DOI: 10.1080/02602938.2010.534760
33. Parkes, K. A., & Powell, S. R. (2015). Is the edTPA the right choice for evaluating teacher readiness? *Arts Education Policy Review*, *116*(2), 103–113. DOI: 10.1080/10632913.2014.944964
34. Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher*, *27*(2). https://scholarworks.bgsu.edu/mwer/vol27/iss2/3
35. Saltis, M. N., Giancaterino, B., & Pierce, C. (2020). Professional dispositions of teacher candidates: Measuring dispositions at a large teacher preparation university to meet national standards. *The Teacher Educator*, *56*(2), 117–131. DOI: 10.1080/08878730.2020.1817217
36. Tait-McCutcheon, S., & Knewstubb, B. (2018). Evaluating the alignment of self, peer and lecture assessment in an Aotearoa New Zealand pre-service teacher education course. *Assessment & Evaluation in Higher Education*, *43*(5), 772–785. https://doi.org/10.1080/02602938.2017.1408771
37. Tracz, S., Torgerson, C., & Beare, P. (2017). The NCTQ selectivity standard and principal evaluation of teacher preparation. *The Teacher Educator*, *52*(1), 8–21. DOI: 10.1080/08878730.2016.1186766
38. Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical, psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952. DOI: 10.1037/a0025125

39. Tillema, H. (2010). Formative assessment in teacher education and teacher professional development. *International Encyclopedia of Education* (3rd ed.), 563–571. DOI: 10.1016/B978-0-08-044894-7.01639-0
40. Yinger, R. J., & Daniel, K. L. (2010). Accreditation and standards in teacher education. *International Encyclopedia of Education* (3rd ed.), 495–502. DOI: 10.1016/B978-0-08-044894-7.00643-6
41. Tigelaar, D. E. H., & van Tartwijk, J. (2010). The evaluation of prospective teachers in teacher education. *International Encyclopedia of Education* (3rd ed.), 511–517. DOI: 10.1016/B978-0-08-044894-7.00647-3
42. Rafiq, S., Qaisar, S., & Butt, I. H. (2022). Analysis of tools used for teacher evaluation process at university level: A document analysis approach. *Gomal University Journal of Research*, *38*(2), 214–224. DOI: 10.51380/gujr-38-02-08
43. Rafiq, S., & Qaisar, S. (2021). Teachers perception about process of teacher evaluation: A case study of a private University of Lahore. *Gomal University Journal of Research*, *37*(3), 350–362. DOI: 10.51380/gujr-37-03-09
44. Khan, G., Khan, A., Hussain, S., & Shaheen, N. (2017). Teacher evaluation: Global perspectives and lessons for Pakistan. *Dialogue (Pakistan)*, *12*(3). https://www.qurtuba.edu.pk/thedialogue/The%20Dialogue/12_3/Dialogue_July_September2017_333-346.pdf
45. Rizwan, S., & Masrur, R. (2018). Standard based three dimensional capacity development of in-service school teachers. *Bulletin of Education and Research*, *40*(3), 31–44. https://pu.edu.pk/images/journal/ier/PDF-FILES/2_40_3_18.pdf

Appendix 4.1

Crosswalk 5 Comparisons

Appendix A4.1: Professional Teaching Standards Crosswalk				
UNESCO Global Framework	**SCOTLAND**	**ENGLAND**	**WALES**	**InTASC**
All Teachers	**Standards for Provisional Registration (SPR)**	**Teachers' Standards**	**Professional Standards for Teaching and Leadership (QTS)**	**All Teachers**
I. Teaching Knowledge & Understanding II. Teaching Practice III. Teaching Relations	1. Being a Teacher in Scotland 2. Professional Knowledge & Understanding 3. Professional Skills and Abilities	I. Teaching II. Personal & professional conduct	I. Pedagogy (P) II. Professional learning (PL) III. Collaboration (C) IV. Innovation (I) V. Leadership (L)	A. The Learner & Learning B. Content Knowledge C. Instructional Practices D. Professional Responsibilities
1. How students learn, and the particular learning, social, and development needs of their students (Domain 1)	*3.2.2 Engage learner participation* • value all learners and their participation, actively engaging children and young people in decision-making about their education • demonstrate care and commitment to working with every learner, embracing diversity to ensure that every learner feels welcome, included and ready to learn • demonstrate knowledge and understanding of wellbeing indicators and childhood development • recognise that childhood experiences impact on the learning and wellbeing of children and young people and actively respond in appropriate ways, seeking advice and collaborating as required; and • utilise strategies to nurture caring and supportive and purposeful relationships with learners and celebrate success	*2. Promote good progress and outcomes by pupils* • be accountable for pupils' attainment, progress and outcomes • be aware of pupils' capabilities and their prior knowledge, and plan teaching to build on these • guide pupils to reflect on the progress they have made and their emerging needs • demonstrate knowledge and understanding of how pupils learn and how this impacts on teaching • encourage pupils to take a responsible and conscientious attitude to their own work and study	P1. The teacher develops and demonstrates up-to-date theoretical knowledge and understanding as well as practical insight into how children and young people develop and learn. P4. The teacher demonstrates knowledge, understanding and experience of high expectations and effective practice in meeting the needs of all learners, whatever their different needs. P14. The teacher provides appropriate levels of challenge and expectations for the range of student abilities and characteristics, motivating learners to achieve.	Standard #1: Learner Development - The teacher understands how learners grow and develop, recognizing that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas, and designs and implements developmentally appropriate and challenging learning experiences. Standard #2: Learning Differences - The teacher uses understanding of individual differences and diverse cultures and communities to ensure inclusive learning environments that enable each learner to meet high standards.
2. The content and related methodologies of the subject matter or content being taught (Domain 1)	*2.1.3 Have knowledge and understanding of Curriculum Design* • principles of curriculum design and how these can be applied in context	*3. Demonstrate good subject and curriculum knowledge* • have a secure knowledge of the relevant subject(s) and curriculum areas, foster and maintain pupils'	P8. The teacher demonstrates secure knowledge of all relevant subject content and knowledge and understanding of the appropriate pedagogies. P9. The teacher demonstrates a knowledge and understanding of relevant pedagogies and	Standard #4: Content Knowledge - The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) he or she teaches and creates learning experiences that make these aspects of the discipline

	• theory and practical skills required in curricular areas as set out in current national and local guidelines • processes used to develop the curriculum • curriculum content and its relevance to the education of every learner • interdisciplinary learning between curricular areas e.g. literacy, numeracy and health and wellbeing, Learning for Sustainability and digital literacy • the skills and competencies that comprise teacher digital literacy and know how to embed digital technologies to enhance teaching and learning • the need to take account of learners with additional support needs	interest in the subject, and address misunderstandings • demonstrate a critical understanding of developments in the subject and curriculum areas, and promote the value of scholarship • demonstrate an understanding of and take responsibility for promoting high standards of literacy, articulacy and the correct use of standard English, whatever the teacher's specialist subject • if teaching early reading, demonstrate a clear understanding of systematic synthetic phonics • if teaching early mathematics, demonstrate a clear understanding of appropriate teaching strategies	disciplines within and across subject content, areas of learning and cross-curricular themes, and plans appropriately. P13 The teacher knows, understands and engages with the principles of curriculum design and innovation, with development of cross-curricular themes relevant to areas of learning and justifies decisions. I28. The teacher models an increasing repertoire of teaching techniques, as expertise emerges and flourishes, in order to inform and enhance the development of others.	accessible and meaningful for learners to assure mastery of the content. Standard #5: Application of Content - The teacher understands how to connect concepts and use differing perspectives to engage learners in critical thinking, creativity, and collaborative problem solving related to authentic local and global issues.
3. Core research and analytical methods that apply in teaching, including with regard to student assessment (Domain 1)	*2.1.2 Have knowledge and understanding of Research and Engagement in Practitioner Enquiry* • how to access and apply relevant findings from educational research to develop an enquiring stance • how to engage appropriately in the ethical investigation of practice. *3.3.1 Engage critically with literature, research and policy* • identify and source appropriate literature, research and policy • engage critically with research to challenge and inform professional practice and question and challenge educational assumptions, beliefs and values of self and system		P3. The range of purposes and practices for assessment is understood and articulated. I29. Research on cognitive, social, emotional and physical development has a positive impact upon pedagogy. The teacher can demonstrate how professional discernment and critical analysis are brought to bear in shaping developing practice. PL21. The teacher has an informed understanding of the contribution of research, including small-scale action research, to the development of practice.	
4. Planning and preparation to meet the learning objectives held for students (Domain 2)	*3.1.1 Plan effectively to meet learners' needs* • plan coherent, progressive and engaging teaching programmes which address the needs of learners • plan learning in accordance with current curriculum guidance including Gaelic medium education where appropriate	*4. Plan and teach well-structured lessons* • impart knowledge and develop understanding through effective use of lesson time • promote a love of learning and children's intellectual curiosity	P7. The teacher demonstrates a knowledge and understanding of the needs of all learners in planning, preparation and teaching, ensuring that the four purposes are the drivers for learners' experiences. P18. In planning, the teacher demonstrates awareness of the importance of encouraging	Standard #7: Planning for Instruction The teacher plans instruction that supports every student in meeting rigorous learning goals by drawing upon knowledge of content areas, curriculum, cross-disciplinary skills, and pedagogy, as well as

	• identify the potential barriers to learning and plan differentiated and appropriately challenging learning experiences to ensure learning is accessible for every learner; • communicate appropriately with every learner, modelling and promoting competence and confidence in literacy, numeracy, health and wellbeing and digital literacy; • ensure teaching builds confidence and promotes the progress of every learner	• set homework and plan other out-of-class activities to consolidate and extend the knowledge and understanding pupils have acquired • reflect systematically on the effectiveness of lessons and approaches to teaching • contribute to the design and provision of an engaging curriculum within the relevant subject area(s)	learners' reflection and evaluation around behaviours and outlooks for learning.	knowledge of learners and the community context.

5. An appropriate range of teaching activities that reflect and align with both the nature of the subject content being taught, and the learning, support, and development needs of the students (Domain 2)	*2.1.1 Have knowledge and understanding of Pedagogical Theories and Professional Practice*
	
	• pedagogical and learning theories and draw on these appropriately to inform
	  ○ curriculum design and content where appropriate taking account of Gaelic medium
	  ○ classroom organisation, learning environment and structures
	  ○ planning, learning and teaching and assessment
	  ○ interdisciplinary learning;
	  ○ outdoor learning, including direct experience of nature and other learning within and beyond school boundaries additional support needs
	• the stages of learners' cognitive, mental, social, emotional, physical, and psychological development and their influence on learning and wellbeing;
	• digital technologies to support learning
		5. Adapt teaching to respond to the strengths and needs of all pupils
		
		• know when and how to differentiate appropriately, using approaches which enable pupils to be taught effectively
		• have a secure understanding of how a range of factors can inhibit pupils' ability to learn, and how best to overcome these
		• demonstrate an awareness of the physical, social and intellectual development of children, and know how to adapt teaching to support pupils' education at different stages of development
		• have a clear understanding of the needs of all pupils, including:
		  ○ those with special educational needs;
		  ○ those of high ability;
		  ○ those with English as an additional language;
		  ○ those with disabilities
		• and be able to use and evaluate distinctive teaching approaches to engage and support them
			P10.The teacher understands the selection, use and justification of a range of imaginative teaching approaches for the benefit of each learner.
			
			P11.The teacher demonstrates an understanding of the use of real life, authentic contexts for learning being provided as a natural part of the learning experience. This extends the learner's cultural, linguistic, religious and socio-economic experience and illustrates applications of concepts and abstracts in practice.
				Standard #8: Instructional Strategies The teacher understands and uses a variety of instructional strategies to encourage learners to develop deep understanding of content areas and their connections, and to build skills to apply knowledge in meaningful ways.

6. Organisation and facilitation of students' activities so that students are able to participate constructively, in a safe and	*3.1.2 Utilise pedagogical approaches and resources*	*7. Manage behaviour effectively to ensure a good and safe learning Environment*	P2. The teacher understands the importance and demonstrates the effective establishment and on-going management of the learning environment, in promoting positive learning habits and behaviours that meet the four	Standard #3: Learning Environments - The teacher works with others to create environments that support individual and collaborative

cooperative manner (Domain 2)	• create meaningful contexts for learners through a range of different learning environments • employ teaching strategies and resources, including digital approaches, to meet the needs and abilities of every learner • use self-evaluation and professional learning to support and improve practice • use a variety of questioning techniques and a range of digital and traditional approaches to enhance learning and teaching; • create opportunities for learning to be transformative in terms of challenging assumptions and expanding world views. *3.2.1 Appropriately organise and manage learning* • create a safe, caring and purposeful learning environment which is welcoming and inclusive, well managed and well organised; • plan and organise effectively to facilitate whole-class lessons, group and individual work and promote independent learning; • use a range of opportunities that stimulate and reflect ongoing learning in varied and dynamic learning environments; • enable learners to make use of well-chosen resources, including digital technologies, to enhance learning, teaching and assessment, as appropriate; • create opportunities for learning to be transformative in terms of challenging assumptions and expanding world views; • evaluate the impact of the learning environment on every learner and learning and to challenge assumptions,	• have clear rules and routines for behaviour in classrooms, and take responsibility for promoting good and courteous behaviour both in classrooms and around the school, in accordance with the school's behaviour policy • have high expectations of behaviour, and establish a framework for discipline with a range of strategies, using praise, sanctions and rewards consistently and fairly • manage classes effectively, using approaches which are appropriate to pupils' needs in order to involve and motivate them • maintain good relationships with pupils, exercise appropriate authority, and act decisively when necessary	purposes and are understood by learners in that context. P17. The teacher promotes and secures learners' self-motivation and self-direction in their learning. L32. Contractual, pastoral, health and safety, legal and professional responsibilities are known and understood by the teacher.	learning, and that encourage positive social interaction, active engagement in learning, and self-motivation.

	surface bias and adapt provision, as appropriate			
7. Assessment and analysis of student learning that informs the further preparation for, and implementation of required teaching and learning activity (Domain 2)	*2.1.4 Have knowledge and understanding of Planning for Assessment, Teaching and Learning* • how to plan for effective assessment, teaching and learning across different contexts • approaches to assessment, recording and reporting as an integral part of learning and teaching • national assessment requirements and requirements of other relevant awarding and accrediting bodies • how to use feedback to engage learners in dialogue about their progress and next steps *3.1.4 Employ assessment, evaluate progress, recording and reporting as an integral part of the teaching process to support and enhance learning* • record, analyse and use assessment data to evaluate learning and teaching • use the results of assessment to identify development needs at class, group and individual level • use a range of differentiated assessment strategies that ensures support and challenge for all learners • use appropriate formative and summative assessment strategies to provide opportunities for challenge and growth appropriate to the needs of every learner and to meet the requirements of the curriculum and awarding and accrediting bodies; • contribute to clear, informative reports for parents/carers and the school which discuss progress in learning in a sensitive and constructive way	*6. Make accurate and productive use of assessment* • know and understand how to assess the relevant subject and curriculum areas, including statutory assessment requirements • make use of formative and summative assessment to secure pupils' progress • use relevant data to monitor progress, set targets, and plan subsequent lessons • give pupils regular feedback, both orally and through accurate marking, and encourage pupils to respond to the feedback	P12.The teacher demonstrates an understanding of how learning develops incrementally and tangentially, building on prior experience and learning, and plans for progress in learning based on this.	Standard #6: Assessment - The teacher understands and uses multiple methods of assessment to engage learners in their own growth, to monitor learner progress, and to guide the teacher's and learner's decision making.

8. Cooperative and collaborative professional processes that contribute to collegial development, and support student learning and development (Domain 3)	*3.1.3 Utilise partnerships for learning and wellbeing*
	
	• contribute to a rights-respecting culture where learners can meaningfully participate in decisions related to their learning, wellbeing, learning environment and their school
	• create and sustain effective working relationships with colleagues, parents/carers, families and the wider school community and partner agencies where appropriate, to support learning and wellbeing across the school
	• practise self-care and support the wellbeing of others, seeking support where necessary;
	• develop partnerships which:
	  ○ support decision-making that is compatible with a sustainable future in a just and equitable world
	  ○ connect learners to their dependence on the natural world and develop their sense of belonging to both the local and global community;
	  ○ connect relevance of learning to skills for life, learning and work
	
	3.2.3 Build positive, rights respecting relationships for learning
	
	• promote and develop positive and purposeful relationships with and between learners, colleagues, families and partners
	• use research-informed approaches to relationship building in a consistent way to build and sustain all professional relationships;
	• communicate appropriately with every learner, modelling and promoting competence and confidence in literacy and numeracy and health and wellbeing;
		Part two: personal and professional conduct
		
		• Teachers uphold public trust in the profession and maintain high standards of ethics and behaviour, within and outside school, by:
		  ○ treating pupils with dignity, building relationships rooted in mutual respect, and at all times observing proper boundaries appropriate to a teacher's professional position
		  ○ having regard for the need to safeguard pupils' well-being, in accordance with statutory provisions
		  ○ showing tolerance of and respect for the rights of others
		  ○ not undermining fundamental British values, including democracy, the rule of law, individual liberty and mutual respect, and tolerance of those with different faiths and beliefs
		  ○ ensuring that personal beliefs are not expressed in ways which exploit pupils' vulnerability or might lead them to break the law
		• Teachers must have proper and professional regard for the ethos, policies and practices of the school in which they teach, and maintain high standards in their own attendance and punctuality.
		• Teachers must have an understanding of, and always act within, the statutory frameworks which set out their professional duties and responsibilities.
			C24 The teacher actively seeks and engages with support from a range of formal and informal sources. This includes observation and team teaching, whilst demonstrating increasing levels of independence.
			
			C25.Organised and constructive work with a range of colleagues to enhance learners' experience is a consistent feature of the teacher's practice. Reflection on developing expertise is structured as a personal or a collaborative process, as appropriate.
			
			C26.The teacher develops high quality relationships with colleagues in order to have a positive impact upon learners' experiences within the school.
			
			I30. The teacher actively seeks support and advice from colleagues in developing innovative approaches within the classroom so that their impact can be evaluated, analysed and shared.
			
			L33. The teacher's understanding of, and commitment to, leading learning is demonstrated through collaborative experiences in schools and other contexts.
			
			L34. The teacher demonstrates an understanding of the nature of responsibilities within and across teams and of the contributions individuals make towards the school's ethos and the successful fulfilment of the school's vision.
				Standard #10: Leadership and Collaboration - The teacher seeks appropriate leadership roles and opportunities to take responsibility for student learning, to collaborate with learners, families, colleagues, other school professionals, and community members to ensure learner growth, and to advance the profession.

	• commit to and demonstrate equity and inclusion; • encourage learners to respect and care for themselves, others and the natural world.			
9. Communications with parents, caregivers, and members of the community, as appropriate, to support the learning objectives of students, including formal and informal reporting (Domain 3)	2.2.2 Have a knowledge and understanding of Learning Communities • the roles and responsibilities of teachers in establishing and sustaining positive and purposeful relationships across the learning community • the distinctive culture, context and ethos of the learning community including Gaelic medium ethos where appropriate; • the role of local, regional and national bodies in relation to the context	*8 Fulfil wider professional responsibilities* • make a positive contribution to the wider life and ethos of the school • develop effective professional relationships with colleagues, knowing how and when to draw on advice and specialist support • deploy support staff effectively • take responsibility for improving teaching through appropriate professional development, responding to advice and feedback from colleagues • communicate effectively with parents with regard to pupils' achievements and well-being	L31.The teacher demonstrates professional attitudes and behaviours, developing positive relationships with learners, parents/carers and colleagues, which illustrate a personal commitment to the fundamental principles of equity and of maximising the potential of all learners. P5. The teacher produces appropriate, timely and accurate records and reports and gives feedback to facilitate a deeper understanding of learning and enhance the learning experience. P6. The importance of positive involvement of parents/carers and other partners is understood and opportunities are taken to observe and evaluate processes. P16. In planning and delivery, the teacher demonstrates an awareness of the importance of encouraging learners to reflect upon their own learning.	
10. Continuous professional development to maintain currency of their professional knowledge and practice (Domain 3)	*1.2 Professional Commitment* Making a professional commitment to learning and learners that is compatible with the aspiration of achieving a sustainable and equitable world embodies what it is to be a teacher in Scotland. This means teachers commit to living the professional values and engage in lifelong learning, reflection, enquiry, leadership of learning and collaborative practice as key aspects of their professionalism. This commitment to professional learning and growth, to the growth of learners, and to helping support that of colleagues, is demonstrated through engagement with all aspects of professional practice. It is demonstrated by working		PL20.The teacher demonstrates an increasingly confident understanding of the theories and research about assessment, pedagogy, child and adolescent development and learning relevant to planning and day-to-day practice. PL22. The Professional Learning Passport influences the ongoing critical reflection and learning of the teacher and is developmental in prompting further professional growth. PL23. There is a commitment to incremental development of personal skills in the use of the Welsh language.	Standard #9: Professional Learning and Ethical Practice - The teacher engages in ongoing professional learning and uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (learners, families, other professionals, and the community), and adapts practice to meet the needs of each learner.

	collegially, in English or Gaelic medium with all members of our learning communities with enthusiasm, adaptability, critical thinking and associated constructive, professional dialogue. *1.3 Standard for Provisional Registration* Professional Values and Professional Commitment are at the core of the Standard for Provisional Registration. They are integral to, and demonstrated through, all our professional relationships and practices. They are about doing well by ourselves, others and the world in which we live. The personal and professional qualities of sustainability and social justice, integrity, trust and respect and professional commitment are crucial if we are to inspire and prepare learners for success in our complex, interdependent and rapidly changing world. *2.2.1 Have knowledge and understanding of Education Systems* • the principal national and international influences on Scottish education • current, relevant legislation, policies and guidance in relation to the teacher's role • pastoral and legal responsibilities, for example, in relation to equality, diversity, additional support needs, child protection, and wellbeing • frameworks, systems and processes to support and enhance teacher professionalism • biases and their impact on people and practices and challenge these *3.3.2 Engage in reflective practice to develop and advance career-long professional learning and expertise*		C27. There are examples of improvement in outcomes for learners following the teacher's seeking and adoption of advice.	

	• reflect and engage critically in self-evaluation using the relevant professional standard; • adopt an enquiring, reflective and critical approach to professional practice; • enhance learning and teaching by taking account of feedback from others including children and young people and actively engage in professional learning to support school improvement; and • maintain a reflective record of evidence of impact of professional learning on self and learners			
Not assigned	*1.1 Professional Values (social justice, trust and respect, and integrity)* Social Justice: • promoting health and wellbeing of self, colleagues and the children and young people in my care • building and fostering positive relationships in the learning community which are respectful of individuals. • embracing global educational and social values of sustainability, equality, equity, and justice and recognising children's rights • respecting the rights of all learners as outlined in the united nations convention on the rights of the child (UNCRC) and their entitlement to be included in decisions regarding their learning experiences and have all aspects of their wellbeing developed and supported • demonstrating a commitment to engaging learners in real world issues to enhance learning experiences and outcomes, and to encourage learning our way to a better future • committing to social justice through fair, transparent, inclusive, and sustainable policies and practices in relation to protected characteristics, (age,		P15. The teacher demonstrates a willingness to seek, listen to and take account of the views of learners in order to engage and encourage them as active participants in their own learning. P19. The teacher raises awareness of how high-quality learning experiences and performance outcomes lead to improved learning and a heightened sense of well-being.	

	disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion and belief, sex, sexual orientation) and intersectionality • valuing, as well as respecting, social, ecological, cultural, religious, and racial diversity and promoting the principles and practices of sustainable development and local and global citizenship for all learners. • demonstrating a commitment to motivating, and including all learners, understanding the influence of gender, social, cultural, racial, ethnic, religious and economic backgrounds on experiences of learning, taking account of specific learning needs and seeking to reduce barriers to learning • demonstrating a commitment to supporting learners who are experiencing or who have experienced trauma, children and young people from a care experienced background and understanding responsibilities as a corporate parent • understanding and challenging discrimination in all its forms, particularly that which is defined by the Equality Act 2010 Trust and Respect: • promoting and engendering a rights respecting culture and the ethical use of authority associated with one's professional roles • acting and behaving in ways that develop a culture of trust and respect for self, others and the natural world • understanding, acknowledging, and respecting the contribution of others in positively influencing the lives of learners • understanding health and wellbeing and the importance of positive and			

	purposeful relationships to provide and ensure a safe and secure environment for all learners and colleagues within a caring and compassionate ethos • respecting individual difference and supporting learners' understanding of themselves, others and their contribution to the development and sustainability of a diverse and inclusive society Integrity: • demonstrating kindness, honesty, courage, and wisdom • being truthful and trustworthy • critically examining professional beliefs, values and attitudes of self and others in the context of collegiate working • challenging assumptions, biases and professional practice, where appropriate			

Appendix 5.1

School Experience End of Placement Report 2021–2022

Please indicate Institution☑

END OF PLACEMENT REPORT FORM

Please provide evidence of the student's progress to date. The student should be assessed against the Standard for Provisional Registration but with consideration given to the stage that they are at in their ITE programme.

Please provide an overall grade for each of the eight sections using the following descriptors as a guideline:

- **S** – Satisfactory: student is making expected progress towards this Standard.
- **U** – Unsatisfactory: student is not making the expected progress towards this Standard.

EIGHT grades should be entered on this form. Indicate the appropriate grade by circling or deleting as appropriate: **do not use split grades**. The comments in all sections should support the grades allocated. **If progress is Unsatisfactory, this should be clearly communicated to the student and substantiating evidence provided in the report.**

Student Name	**School**	**Date**
ITE Programme	**Number of days absent** *(school to complete)*	
Names of persons completing the report	**Designations**	
I confirm that the content of the Report has been discussed with the student Yes / *No (delete as appropriate)* *Date:*	*If 'No' please indicate why this was not possible*	

University of Glasgow | School of Education ☐

University of Strathclyde Glasgow ☐

SCHOOL OF EDUCATION

Please indicate Institution☑

Standard for Provisional Registration

1 Being a Teacher in Scotland

1.1 Professional Values

Social justice is the view that everyone deserves equal economic, political and social rights and opportunities now and in the future.

Trust and Respect are expectations of positive actions that support authentic relationship building and show care for the needs and feelings of the people involved and respect for our natural world and its limited resources.

Integrity is the practice of being honest and showing a consistent and uncompromising adherence to strong moral and ethical principles and values

Progress at this stage of the course

☐ S

☐ U

1.2 Professional Commitment

Living the Professional Values, professional learning, reflection, enquiry, leadership of learning, collaborative practice, and understanding the needs of all learners.

Progress at this stage of the course

☐ S

☐ U

University of Glasgow | School of Education ☐

University of Strathclyde Glasgow ☐

SCHOOL OF EDUCATION

Please indicate Institution☑

	1.3 **Engaging with the Standard for Provisional Registration** Actively embracing and promoting principles and practices of sustainability Seeking all opportunities to be leaders of learning. Leading learning for, and with, all learners with whom there is engagement. Work with and support the development of colleagues and other partners. Engaging with the SPR and showing a commitment to working towards it. Progress at this stage of the course ☐ S ☐ U

University of Glasgow | School of Education ☐

University of Strathclyde Glasgow ☐

SCHOOL OF EDUCATION

Please indicate Institution☑

2 Professional Knowledge and Understanding	
	2.1 **Curriculum and Pedagogy** Pedagogical Theories and Professional Practice Research and Engagement in Practitioner Enquiry Curriculum Design Planning for Assessment, Teaching and Learning Progress at this stage of the course ☐ S ☐ U
	2.2 **Professional responsibilities** Education Systems Learning Communities Progress at this stage of the course ☐ S ☐ U

3 Professional Skills and Abilities

End of Placement Report

University of Glasgow | School of Education ☐

University of Strathclyde Glasgow ☐

SCHOOL OF EDUCATION

Please indicate Institution☑

	3.1 **Curriculum and Pedagogy** Plan effectively to meet learners' needs Utilise pedagogical approaches and resources Utilise partnerships for learning and wellbeing Employ assessment, evaluate progress, recording and reporting as an integral part of the teaching process to support and enhance learning Progress at this stage of the course ☐ S ☐ U
	3.2 **The Learning Context** Appropriately organise and manage learning Engage learner participation Build positive, rights respecting relationships for learning Progress at this stage of the course ☐ S ☐ U
	3.3 **Professional Learning** Engage critically with literature, research and policy Engage in reflective practice to develop and advance career-long professional learning and expertise Progress at this stage of the course ☐ S ☐ U

Areas for development and next steps

End of Placement Report

University of Glasgow | School of Education ☐

University of Strathclyde Glasgow ☐

SCHOOL OF EDUCATION

Please indicate Institution☑

Student		**Date**	
School Staff		**Date**	
Tutor		**Date**	

Appendix 5.2

Formative Assessed Visit Form

Formative Assessed Visit Observation Proforma

Student Teacher: _____

Course: _____

School Mentor: _____ Date _____

School Experience Tutor: _____ Class _____

Aspects of SPR being observed throughout the FAV (please see detailed descriptors at the end of the proforma)

2.1.3 Have knowledge and understanding of Curriculum Design
2.1.4 Have knowledge and understanding of Planning for Assessment, Teaching and Learning
3.1.1 Plan effectively to meet learners' needs
3.1.2 Utilise pedagogical approaches and resources
3.1.4 Employ assessment, evaluate progress, recording and reporting as an integral part of the teaching process to support and enhance learning
3.2.1 Appropriately organise and manage learning
3.2.2 Engage learner participation
3.2.3 Build positive, rights-respecting relationships for learning
3.3.2 Engage in reflective practice to develop and advance career-long professional learning and expertise

Becoming a Teacher – Evidence of Skills and Abilities	**Comments**
Planning	
Presentation and Communication	
Applying Curriculum Knowledge	
Lesson Structure	
Learning Activities and Engagement	

Questioning	
Classroom Management and Organisation	
Assessment	
Professional Reflection	

Strengths observed:

Next steps:

Any additional comments:

Aspects of the Standard for Provisional Registration Referenced

2.1.3 Have knowledge and understanding of Curriculum Design

As a student teacher you are required to demonstrate knowledge and understanding of:

- theory and practical skills required in curricular areas as set out in current national and local guidelines
- curriculum content and its relevance to the education of every learner
- interdisciplinary learning between curricular areas e.g. literacy, numeracy and health and wellbeing, Learning for Sustainability and digital literacy
- the skills and competencies that comprise teacher digital literacy and know-how to embed digital technologies to enhance teaching and learning
- the need to take account of learners with additional support needs.

2.1.4 Have knowledge and understanding of Planning for Assessment, Teaching and Learning

As a student teacher you are required to demonstrate knowledge and understanding of:

- how to plan for effective assessment, teaching and learning across different contexts
- how to use feedback to engage learners in dialogue about their progress and next steps.

3.1.1 Plan effectively to meet learners' needs

As a student teacher to demonstrate your professional skills and abilities you are required to:

- plan coherent, progressive and engaging teaching (programmes) which address the needs of learners
- plan learning in accordance with current curriculum guidance
- identify the potential barriers to learning and plan differentiated and appropriately challenging learning experiences to ensure learning is accessible for every learner
- communicate appropriately with every learner, modelling and promoting competence and confidence
- ensure teaching builds confidence and promotes the progress of every learner.

3.1.2 Utilise pedagogical approaches and resources

As a student teacher to demonstrate your professional skills and abilities you are required to:

- create meaningful contexts for learners through a range of different learning environments
- employ teaching strategies and resources, including digital approaches, to meet the needs and abilities of every learner
- use self-evaluation and professional learning to support and improve practice
- use a variety of questioning techniques and a range of digital and traditional approaches to enhance learning and teaching
- create opportunities for learning to be transformative in terms of challenging assumptions and expanding world views.

3.1.4 Employ assessment, evaluate progress, recording and reporting as an integral part of the teaching process to support and enhance learning

As a student teacher to demonstrate your professional skills and abilities you are required to:

- use the results of assessment to identify development needs at class, group and individual level
- use a range of differentiated assessment strategies that ensures support and challenge for all learners
- use appropriate formative and summative assessment strategies to provide opportunities for challenge and growth appropriate to the needs of every learner and to meet the requirements of the curriculum and awarding and accrediting bodies.

3.2.1 Appropriately organise and manage learning

As a student teacher to demonstrate your professional skills and abilities you are required to:

- Create a safe, caring and purposeful learning environment which is welcoming and inclusive, well managed and well organized
- Plan and organise effectively to facilitate whole-class lessons, group and individual work and promote independent learning
- Use a range of opportunities that stimulate and reflect ongoing learning in varied and dynamic learning environments
- Enable learners to make use of well-chosen resources, including digital technologies, to enhance learning, teaching and assessment (*as appropriate*)
- Create opportunities for learning to be transformative in terms of challenging assumptions and expanding world views
- Evaluate the impact of the learning environment on every learner and learning and to challenge assumptions, surface bias and adapt provision (*as appropriate*).

3.2.2 Engage learner participation

As a student teacher to demonstrate your professional skills and abilities you are required to:

- Value all learners and their participation, actively engaging children and young people in decision-making (about their education)
- Demonstrate care and commitment to working with every learner, embracing diversity to ensure that every learner feels welcome, included and ready to learn
- Utilise strategies to nurture caring and supportive and purposeful relationships with learners and celebrate success.

3.2.3 Build positive, rights-respecting relationships for learning

As a student teacher to demonstrate your professional skills and abilities you are required to:

- Promote and develop positive and purposeful relationships with and between learners
- Use research-informed approaches to relationship building in a consistent way to build and sustain all professional relationships
- Demonstrate equity and inclusion.

3.3.2 Engage in reflective practice to develop and advance career-long professional learning and expertise

As a student teacher to demonstrate your professional skills and abilities you are required to:

- reflect and engage critically in self-evaluation using the relevant professional standard
- adopt an enquiring, reflective and critical approach to professional practice
- enhance learning and teaching by taking account of feedback from others including children and young people and actively engage in professional learning to support school improvement
- maintain a reflective record of evidence of impact of professional learning on self and learners.

Appendix 5.3

Personal and Professional Development Plan

University of Glasgow | School of Education

Name:

PGDE

Personal and Professional Development Plan

RECORD OF SCHOOL PLACEMENTS

Placement	School	Class (primary)	Significant area of learning
Placement 1			
Placement 2			
Placement 3			

Placement 1

In discussion with your class/principal teacher, identify areas for particular focus to undertake in the second placement. These areas will have been identified from observations made by your class teacher(s), other school staff and your tutor. **Record your targets towards the end of the placement**. Your tutor may ask to see the record of your progress on these targets during Placement 2.

Target 1:

Target 2:

Target 3:

Student _____ Date _____

Class/Principal Teacher (Placement One) _____ Date _____

Tutor _____ Date _____

Placement 2

In discussion with your class/principal teacher, identify areas for particular focus to undertake in the next School Placement. These areas will have been identified from observations made by your class teacher(s), other school staff and your tutor. **Record your targets towards the end of the placement**. Your tutor may ask to see the record of your progress on these targets during Placement 3.

Target 1:

Target 2:

Target 3:

Student _____ Date _____

Class/Principal Teacher (Placement Two) _____ Date _____

Tutor _____ Date _____

Placement 3

In discussion with your class/principal teacher, identify areas for particular focus to undertake in the **Induction Year**. These areas will have been identified from observations made by your class teacher, other school staff and your tutor during Placement 3. **Record your targets towards the end of the placement**.

Target 1:

Target 2:

Target 3:

Student _____ Date _____

Class/Principal Teacher (Placement Three) _____ Date _____

Tutor _____ Date _____

Appendix 5.4

Recruitment Script

SES Project Recruitment Script

We are excited to invite you to participate in our research study to understand how teaching effectiveness is evaluated. You are chosen as a candidate participant due to your previous role as either school-based mentor teachers or associate tutors, or university staff in judging ITE students' performance per teaching standards while on school placement experiences.

The aim of our project is to improve the accuracy of assessing teaching capabilities in initial teacher education (ITE) in Scotland, Wales and England by examining how judgements of teaching effectiveness are made. We are confident that with your opinions and insights, we can achieve this goal.

As a participant in the study, you will be asked to watch a 15-minute teaching video and then complete a questionnaire about the effectiveness of the teaching in the video. You will also have the opportunity to explain your thought process in making your judgement.

Your participation in this study is voluntary, and it will not affect your employment in any way. We will not collect any identifying information from you unless you choose to participate in a follow-up online focus group session. The questionnaire should take approximately 30 minutes to complete in total including watching the video.

We would be grateful if you could complete the questionnaire by **Thursday, 1 June** end of day. Your contribution to this study is greatly appreciated, and we believe that your say will be invaluable in improving the evaluation of teaching capabilities in initial teacher education.

Here's the link to the questionnaire:

Many thanks for participating in the project. If you have any questions, please contact Sarah K Anderson, PhD: sarah.anderson.3@glasgow.ac.uk

Kind regards,

Appendix 5.5

Case Study 1 Questionnaire: Expanded Results

Appendix A5.5: University of Glasgow Non-Consolidated Questionnaire Tables

Table A5.5.1

Combined – All Roles Together: Participants' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Statement				*n = 17*					
It is important that judgements of teaching effectiveness are accurate.	8	7	1	1	0	0	0	6.29	0.82
It is important that judgements of teaching effectiveness are consistent.	9	6	2	0	0	0	0	6.41	0.69
It is important that different evaluators reach consensus.	3	10	4	0	0	0	0	5.94	0.64
It is important that evaluators use evidence to make judgements.	13	4	0	0	0	0	0	6.76	0.42
It is important that professional judgement is used when judging teaching effectiveness.	10	6	1	0	0	0	0	6.53	0.60
It is important that judgements about teaching effectiveness are made by more than one evaluator.	6	6	4	1	0	0	0	6.00	0.97
It is important that potential sources of evaluator error are addressed.	10	5	1	1	0	0	0	6.41	0.84
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	14	3	0	0	0	0	0	6.82	0.38
Judgements are always related to particular teachers at particular points in time and in particular situations.	5	3	3	4	1	1	0	5.24	1.51
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	10	5	0	1	1	0	0	6.29	1.13

Note. Questionnaire: Q12–13.

Table A5.5.2

Teacher Educators' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree	Mean	SD
	(7)	(6)	(5)	(4)	(3)	(2)	(1)		
Statement				*n = 5*					
It is important that judgements of teaching effectiveness are accurate.	2	1	1	1	0	0	0	5.80	1.17
It is important that judgements of teaching effectiveness are consistent.	2	2	1	0	0	0	0	6.20	0.75
It is important that different evaluators reach consensus.	0	4	1	0	0	0	0	5.80	0.40
It is important that evaluators use evidence to make judgements.	4	1	0	0	0	0	0	6.80	0.40
It is important that professional judgement is used when judging teaching effectiveness.	3	2	0	0	0	0	0	6.60	0.49
It is important that judgements about teaching effectiveness are made by more than one evaluator.	2	1	2	0	0	0	0	6.00	0.89
It is important that potential sources of evaluator error are addressed.	4	1	0	0	0	0	0	6.80	0.40
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	4	1	0	0	0	0	0	6.80	0.40
Judgements are always related to particular teachers at particular points in time and in particular situations.	1	0	2	2	0	0	0	5.00	1.10
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	4	1	0	0	0	0	0	6.80	0.40

Note. Questionnaire: Q12–13.

Table A5.5.3

Associate Tutors' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Statement				*(n = 4)*					
It is important that judgements of teaching effectiveness are accurate.	2	2	0	0	0	0	0	6.50	0.5
It is important that judgements of teaching effectiveness are consistent.	4	0	0	0	0	0	0	7.00	0.00
It is important that different evaluators reach consensus.	3	1	0	0	0	0	0	6.75	0.43
It is important that evaluators use evidence to make judgements.	3	1	0	0	0	0	0	6.75	0.43
It is important that professional judgement is used when judging teaching effectiveness.	2	2	0	0	0	0	0	6.50	0.50
It is important that judgements about teaching effectiveness are made by more than one evaluator.	1	1	1	1	0	0	0	5.50	1.12
It is important that potential sources of evaluator error are addressed.	2	2	0	0	0	0	0	6.50	0.50
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	3	1	0	0	0	0	0	6.75	0.43
Judgements are always related to particular teachers at particular points in time and in particular situations.	3	1	0	0	0	0	0	6.75	0.43
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	3	1	0	0	0	0	0	6.75	0.43

Note. Questionnaire: Q12–13.

Table A5.5.4

Mentor Teacher's Level of Agreement With Statements Related to Judging Teaching Effectiveness

Statement	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
				(n = 8)					
It is important that judgements of teaching effectiveness are accurate.	4	4	0	0	0	0	0	6.50	0.50
It is important that judgements of teaching effectiveness are consistent.	3	4	1	0	0	0	0	6.25	0.66
It is important that different evaluators reach consensus.	0	5	3	0	0	0	0	5.63	0.48
It is important that evaluators use evidence to make judgements.	6	2	0	0	0	0	0	6.75	0.43
It is important that professional judgement is used when judging teaching effectiveness.	5	2	1	0	0	0	0	6.50	0.71
It is important that judgements about teaching effectiveness are made by more than one evaluator.	3	4	1	0	0	0	0	6.25	0.66
It is important that potential sources of evaluator error are addressed.	4	2	1	1	0	0	0	6.13	1.05
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	7	1	0	0	0	0	0	6.88	0.33
Judgements are always related to particular teachers at particular points in time and in particular situations.	1	2	1	2	1	1	0	4.63	1.58
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	3	3	0	1	1	0	0	5.75	1.39

Note. Questionnaire: Q12–13.

Table A5.5.5

Combined – All Roles Together: Participants' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …				*(n = 17)*					
Clarity of the judgement criteria	9	6	1	1	0	0	0	6.35	0.84
Tension of using judgements for both professional growth and accountability	2	8	3	4	0	0	0	5.47	0.98
Clarity of procedures for making judgements	9	7	0	1	0	0	0	6.41	0.77
Individual understanding of effective teaching	9	6	2	0	0	0	0	6.41	0.69
Contested nature of what defines effective teaching	4	6	5	2	0	0	0	5.71	0.96
Professional teaching standards	11	4	1	1	0	0	0	6.47	0.85
Power relationships between universities and schools in teacher education	0	5	3	6	1	1	1	4.41	1.42
Personal intuition about what happens in a classroom	3	6	4	1	2	1	0	5.24	1.44
Perceived levels of importance of different dimensions of teaching	2	4	6	3	0	2	0	4.94	1.39
Complexity of the classroom environment in which judgements are made	6	6	4	1	0	0	0	6.00	0.91
Evaluator's tendencies toward leniency or severity	2	6	6	2	0	1	0	5.29	1.18
Personal biases and beliefs of the evaluator	4	5	4	1	1	1	1	5.18	1.72
Experiences of the evaluator from observing other teachers	1	9	5	1	0	1	0	5.41	1.09
Prior interactions between the teacher and the evaluator	1	5	4	1	5	1	0	4.59	1.46
Holding a pre-observation discussion	3	7	4	3	0	0	0	5.59	0.97
Level of involvement of the individual being evaluated in the judgement process	2	7	7	1	0	0	0	5.59	0.77
Training of evaluators to use observation criteria for making judgements	8	5	4	0	0	0	0	6.24	0.81

Observation skills of the evaluator	7	8	2	0	0	0	0	6.29	0.67
Perceptual information (cues) available to the evaluator	3	10	1	3	0	0	0	5.76	0.94
Policies regarding evaluation of teaching effectiveness	1	13	1	1	1	0	0	5.71	0.89
Quality of the reasoning strategies used to make decisions	3	10	3	1	0	0	0	5.88	0.76

Note. Questionnaire: Q14–15.

Table A5.5.6

Teacher Educators' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …				*(n = 5)*					
Clarity of the judgement criteria	3	2	0	0	0	0	0	6.60	0.49
Tension of using judgements for both professional growth and accountability	0	3	0	2	0	0	0	5.20	0.98
Clarity of procedures for making judgements	2	3	0	0	0	0	0	6.40	0.49
Individual understanding of effective teaching	2	3	0	0	0	0	0	6.40	0.49
Contested nature of what defines effective teaching	2	0	1	2	0	0	0	5.40	1.36
Professional teaching standards	3	2	0	0	0	0	0	6.60	0.49
Power relationships between universities and schools in teacher education	0	0	1	4	0	0	0	4.20	0.40
Personal intuition about what happens in a classroom	0	2	2	1	0	0	0	5.20	0.75
Perceived levels of importance of different dimensions of teaching	1	1	1	2	0	0	0	5.20	1.17
Complexity of the classroom environment in which judgements are made	1	3	1	0	0	0	0	6.00	0.63
Evaluator's tendencies toward leniency or severity	1	3	1	0	0	0	0	6.00	0.63
Personal biases and beliefs of the evaluator	2	2	1	0	0	0	0	6.20	0.75
Experiences of the evaluator from observing other teachers	1	4	0	0	0	0	0	6.20	0.40
Prior interactions between the teacher and the evaluator	1	2	0	1	1	0	0	5.20	1.47
Holding a pre-observation discussion	1	2	1	1	0	0	0	5.60	1.02
Level of involvement of the individual being evaluated in the judgement process	2	2	1	0	0	0	0	6.20	0.75
Training of evaluators to use observation criteria for making judgements	3	0	2	0	0	0	0	6.20	0.98

Observation skills of the evaluator	3	2	0	0	0	0	0	6.60	0.49
Perceptual information (cues) available to the evaluator	1	4	0	0	0	0	0	6.20	0.40
Policies regarding evaluation of teaching effectiveness	1	4	0	0	0	0	0	6.20	0.40
Quality of the reasoning strategies used to make decisions	2	3	0	0	0	0	0	6.40	0.49

Note. Questionnaire: Q14–15.

Table A5.5.7

Associate Tutor's Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …				*(n = 4)*					
Clarity of the judgement criteria	3	1	0	0	0	0	0	6.75	0.43
Tension of using judgements for both professional growth and accountability	1	2	0	1	0	0	0	5.75	1.09
Clarity of procedures for making judgements	3	0	0	1	0	0	0	6.25	1.30
Individual understanding of effective teaching	3	1	0	0	0	0	0	6.75	0.43
Contested nature of what defines effective teaching	0	3	1	0	0	0	0	5.75	0.43
Professional teaching standards	3	0	0	1	0	0	0	6.25	1.30
Power relationships between universities and schools in teacher education	0	2	0	1	1	0	0	4.75	1.30
Personal intuition about what happens in a classroom	1	2	0	0	1	0	0	5.50	1.50
Perceived levels of importance of different dimensions of teaching	0	2	1	1	0	0	0	5.25	0.83
Complexity of the classroom environment in which judgements are made	2	1	1	0	0	0	0	6.25	0.83
Evaluator's tendencies toward leniency or severity	0	0	2	2	0	0	0	4.50	0.50
Personal biases and beliefs of the evaluator	0	1	1	1	0	0	1	4.00	1.87
Experiences of the evaluator from observing other teachers	0	1	2	1	0	0	0	5.00	0.71
Prior interactions between the teacher and the evaluator	0	0	2	0	2	0	0	4.00	1.00
Holding a pre-observation discussion	0	2	1	1	0	0	0	5.25	0.83
Level of involvement of the individual being evaluated in the judgement process	0	1	2	1	0	0	0	5.00	0.71
Training of evaluators to use observation criteria for making judgements	2	0	2	0	0	0	0	6.00	1.00

Observation skills of the evaluator	2	0	2	0	0	0	0	6.00	1.00
Perceptual information (cues) available to the evaluator	0	2	0	2	0	0	0	5.00	1.00
Policies regarding evaluation of teaching effectiveness	0	3	0	1	0	0	0	5.50	0.87
Quality of the reasoning strategies used to make decisions	0	3	1	0	0	0	0	5.75	0.43

Note. Questionnaire: Q14–15.

Table A5.5.8

Mentor Teachers' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …				*(n = 8)*					
Clarity of the judgement criteria	3	3	1	1	0	0	0	6.00	1.00
Tension of using judgements for both professional growth and accountability	1	3	3	1	0	0	0	5.50	0.87
Clarity of procedures for making judgements	4	4	0	0	0	0	0	6.50	0.50
Individual understanding of effective teaching	4	2	2	0	0	0	0	6.25	0.83
Contested nature of what defines effective teaching	2	3	3	0	0	0	0	5.88	0.78
Professional teaching standards	5	2	1	0	0	0	0	6.50	0.71
Power relationships between universities and schools in teacher education	0	3	2	1	0	1	1	4.38	1.80
Personal intuition about what happens in a classroom	2	2	2	0	1	1	0	5.13	1.69
Perceived levels of importance of different dimensions of teaching	1	1	4	0	0	2	0	4.63	1.65
Complexity of the classroom environment in which judgements are made	3	2	2	1	0	0	0	5.88	1.05
Evaluator's tendencies toward leniency or severity	1	3	3	0	0	1	0	5.25	1.39
Personal biases and beliefs of the evaluator	2	2	2	0	1	1	0	5.13	1.69
Experiences of the evaluator from observing other teachers	0	4	3	0	0	1	0	5.13	1.27
Prior interactions between the teacher and the evaluator	0	3	2	0	2	1	0	4.50	1.50
Holding a pre-observation discussion	2	3	2	1	0	0	0	5.75	0.97
Level of involvement of the individual being evaluated in the judgement process	0	4	4	0	0	0	0	5.50	0.50

Training of evaluators to use observation criteria for making judgements	3	5	0	0	0	0	0	6.38	0.48
Observation skills of the evaluator	2	6	0	0	0	0	0	6.25	0.19
Perceptual information (cues) available to the evaluator	2	4	1	1	0	0	0	5.88	0.93
Policies regarding evaluation of teaching effectiveness	0	6	1	0	1	0	0	5.50	1.00
Quality of the reasoning strategies used to make decisions	1	4	2	1	0	0	0	5.63	0.86

Note. Questionnaire: Q14–15.

Appendix 6.1

Case Study 2 Questionnaire: Expanded Analysis

Appendix A6.1: Leeds Becket University Non-Consolidated Questionnaire Tables

Table A6.1.1

Combined – All Roles Together: Participants' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Statement				*(n = 24)*					
It is important that judgements of teaching effectiveness are accurate.	15	9	0	0	0	0	0	6.63	0.48
It is important that judgements of teaching effectiveness are consistent.	16	8	0	0	0	0	0	6.67	0.47
It is important that different evaluators reach consensus.	6	10	6	2	0	0	0	5.83	0.90
It is important that evaluators use evidence to make judgements.	18	4	2	0	0	0	0	6.67	0.62
It is important that professional judgement is used when judging teaching effectiveness.	13	9	2	0	0	0	0	6.46	0.64
It is important that judgements about teaching effectiveness are made by more than one evaluator.	8	8	4	3	1	0	0	5.79	1.15
It is important that potential sources of evaluator error are addressed.	8	13	3	0	0	0	0	6.21	0.64
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	16	8	0	0	0	0	0	6.67	0.47
Judgements are always related to particular teachers at particular points in time and in particular situations.	8	5	8	0	2	1	0	5.58	1.38
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	15	5	3	1	0	0	0	6.42	0.86

Note. Questionnaire: Q12–13.

Table A6.1.2

Teacher Educators' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Statement				*(n = 13)*					
It is important that judgements of teaching effectiveness are accurate.	8	5	0	0	0	0	0	6.62	0.49
It is important that judgements of teaching effectiveness are consistent.	9	4	0	0	0	0	0	6.69	0.46
It is important that different evaluators reach consensus.	4	5	4	0	0	0	0	6.00	0.78
It is important that evaluators use evidence to make judgements.	9	2	2	0	0	0	0	6.54	0.75
It is important that professional judgement is used when judging teaching effectiveness.	6	5	2	0	0	0	0	6.31	0.72
It is important that judgements about teaching effectiveness are made by more than one evaluator.	3	4	3	2	1	0	0	5.46	1.22
It is important that potential sources of evaluator error are addressed.	3	9	1	0	0	0	0	6.15	0.53
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	8	5	0	0	0	0	0	6.62	0.49
Judgements are always related to particular teachers at particular points in time and in particular situations.	6	1	5	0	1	0	0	5.85	1.23
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	8	2	3	0	0	0	0	6.38	0.84

Note. Questionnaire: Q12–13.

Table A6.1.3

Associate Tutors' and School Experience Tutors' Level of Agreement With Statements Related to Judging Teaching Effectiveness

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree	Mean	SD
	(7)	(6)	(5)	(4)	(3)	(2)	(1)		
Statement					*(n = 7)*				
It is important that judgements of teaching effectiveness are accurate.	4	3	0	0	0	0	0	6.57	0.49
It is important that judgements of teaching effectiveness are consistent.	4	3	0	0	0	0	0	6.57	0.49
It is important that different evaluators reach consensus.	2	3	2	0	0	0	0	6.00	0.76
It is important that evaluators use evidence to make judgements.	5	2	0	0	0	0	0	6.71	0.45
It is important that professional judgement is used when judging teaching effectiveness.	3	4	0	0	0	0	0	6.43	0.49
It is important that judgements about teaching effectiveness are made by more than one evaluator.	2	3	1	1	0	0	0	5.86	0.99
It is important that potential sources of evaluator error are addressed.	2	3	2	0	0	0	0	6.00	0.76
It is important for the teacher to understand how judgements about their teaching effectiveness are made.	5	2	0	0	0	0	0	6.71	0.45
Judgements are always related to particular teachers at particular points in time and in particular situations.	2	2	3	0	0	0	0	5.86	0.83
It is important that judgements about teaching effectiveness are considered fair by stakeholders.	4	2	0	1	0	0	0	6.29	1.03

Note. Questionnaire: Q12–13.

Table A6.1.4

Combined – All Roles Together: Participants' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …					*(n = 24)*				
Clarity of the judgement criteria	10	13	0	0	1	0	0	6.29	0.84
Tension of using judgements for both professional growth and accountability	3	12	4	3	1	1	0	5.42	1.22
Clarity of procedures for making judgements	5	18	1	0	0	0	0	6.17	0.47
Individual understanding of effective teaching	11	10	3	0	0	0	0	6.33	0.69
Contested nature of what defines effective teaching	8	6	8	1	1	0	0	5.79	1.08
Professional teaching standards	9	13	1	1	0	0	0	6.25	0.72
Power relationships between universities and schools in teacher education	5	6	5	6	0	1	1	5.13	1.54
Personal intuition about what happens in a classroom	4	10	7	1	1	1	0	5.50	1.19
Perceived levels of importance of different dimensions of teaching	3	10	6	3	1	1	0	5.33	1.21
Complexity of the classroom environment in which judgements are made	7	10	3	3	1	0	0	5.79	1.12
Evaluator tendencies toward leniency or severity	3	10	6	2	2	0	1	5.25	1.39
Personal biases and beliefs of the evaluator	5	4	8	4	2	0	1	5.08	1.47
Experiences of the evaluator from observing other teachers	7	6	8	1	1	0	1	5.54	1.41
Prior interactions between the teacher and the evaluator	5	5	6	4	2	1	1	5.00	1.61
Holding a pre-observation discussion	6	8	5	2	2	0	1	5.42	1.50
Level of involvement of the individual being evaluated in the judgement process	4	9	4	6	1	0	0	5.38	1.15
Training of evaluators to use observation criteria for making judgements	6	12	4	1	0	0	1	5.79	1.26

Observation skills of the evaluator	9	10	3	1	0	0	1	5.96	1.31
Perceptual information (cues) available to the evaluator	4	12	5	2	0	0	1	5.58	1.26
Policies regarding evaluation of teaching effectiveness	3	7	6	6	1	0	1	5.04	1.37
Quality of the reasoning strategies used to make decisions	2	11	5	4	1	0	1	5.21	1.32

Note. Questionnaire: Q14–15.

Table A6.1.5

Teacher Educators' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …					*(n = 13)*				
Clarity of the judgement criteria	4	8	0	0	1	0	0	6.08	1.00
Tension of using judgements for both professional growth and accountability	2	6	3	1	1	0	0	5.54	1.08
Clarity of procedures for making judgements	3	9	1	0	0	0	0	6.15	0.53
Individual understanding of effective teaching	5	6	2	0	0	0	0	6.23	0.70
Contested nature of what defines effective teaching	3	4	6	0	0	0	0	5.77	0.80
Professional teaching standards	2	9	1	1	0	0	0	5.92	0.73
Power relationships between universities and schools in teacher education	3	3	2	4	0	1	0	5.15	1.46
Personal intuition about what happens in a classroom	2	5	5	0	0	1	0	5.46	1.22
Perceived levels of importance of different dimensions of teaching	3	6	3	0	0	1	0	5.69	1.26
Complexity of the classroom environment in which judgements are made	4	6	1	1	1	0	0	5.85	1.17
Evaluator tendencies toward leniency or severity	2	5	2	1	2	0	1	5.00	1.71
Personal biases and beliefs of the evaluator	4	2	3	1	2	0	1	5.08	1.82
Experiences of the evaluator from observing other teachers	5	4	3	0	0	0	1	5.77	1.58
Prior interactions between the teacher and the evaluator	2	3	3	2	2	0	1	4.77	1.67
Holding a pre-observation discussion	3	5	2	0	2	0	1	5.23	1.76
Level of involvement of the individual being evaluated in the judgement process	3	4	2	3	1	0	0	5.38	1.27
Training of evaluators to use observation criteria for making judgements	4	4	4	0	0	0	1	5.62	1.55

Observation skills of the evaluator	5	6	1	0	0	0	1	5.92	1.54
Perceptual information (cues) available to the evaluator	3	6	3	0	0	0	1	5.62	1.50
Policies regarding evaluation of teaching effectiveness	2	3	2	5	0	0	1	4.85	1.56
Quality of the reasoning strategies used to make decisions	2	4	3	2	1	0	1	5.00	1.62

Note. Questionnaire: Q14–15.

Table A6.1.6

Associate Tutors' and School Experience Tutors' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by…					*(n = 7)*				
Clarity of the judgement criteria	3	4	0	0	0	0	0	6.43	0.49
Tension of using judgements for both professional growth and accountability	1	3	1	1	0	1	0	5.14	1.55
Clarity of procedures for making judgements	2	5	0	0	0	0	0	6.29	0.45
Individual understanding of effective teaching	3	4	0	0	0	0	0	6.43	0.49
Contested nature of what defines effective teaching	4	2	1	0	0	0	0	6.43	0.73
Professional teaching standards	4	3	0	0	0	0	0	6.57	0.49
Power relationships between universities and schools in teacher education	0	2	2	2	0	0	1	4.43	1.59
Personal intuition about what happens in a classroom	1	3	1	1	1	0	0	5.29	1.28
Perceived levels of importance of different dimensions of teaching	0	3	2	1	1	0	0	5.00	1.07
Complexity of the classroom environment in which judgements are made	1	3	1	2	0	0	0	5.43	1.05
Evaluator tendencies toward leniency or severity	1	2	3	1	0	0	0	5.43	0.90
Personal biases and beliefs of the evaluator	1	0	3	3	0	0	0	4.86	0.99
Experiences of the evaluator from observing other teachers	1	0	4	1	1	0	0	4.86	1.12
Prior interactions between the teacher and the evaluator	1	1	3	1	0	1	0	4.86	1.46
Holding a pre-observation discussion	1	3	2	1	0	0	0	5.57	0.90
Level of involvement of the individual being evaluated in the judgement process	1	3	1	2	0	0	0	5.43	1.05

Training of evaluators to use observation criteria for making judgements	1	6	0	0	0	0	0	6.14	0.35
Observation skills of the evaluator	2	4	1	0	0	0	0	6.14	0.64
Perceptual information (cues) available to the evaluator	0	5	2	0	0	0	0	5.71	0.45
Policies regarding evaluation of teaching effectiveness	1	3	2	0	1	0	0	5.43	1.18
Quality of the reasoning strategies used to make decisions	0	5	2	0	0	0	0	5.71	0.45

Note. Questionnaire: Q14–15.

Table A6.1.7

Mentor Teachers' Level of Agreement With Statements Related to Factors Influencing Judgement

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)	Mean	SD
Judgement-making is influenced by …					*(n = 4)*				
Clarity of the judgement criteria	3	1	0	0	0	0	0	6.75	0.43
Tension of using judgements for both professional growth and accountability	0	3	0	1	0	0	0	5.50	0.87
Clarity of procedures for making judgements	0	4	0	0	0	0	0	6.00	0.00
Individual understanding of effective teaching	3	0	1	0	0	0	0	6.50	0.87
Contested nature of what defines effective teaching	1	0	1	1	1	0	0	4.75	1.48
Professional teaching standards	3	1	0	0	0	0	0	6.75	0.43
Power relationships between universities and schools in teacher education	2	1	1	0	0	0	0	6.25	0.83
Personal intuition about what happens in a classroom	1	2	1	0	0	0	0	6.00	0.71
Perceived levels of importance of different dimensions of teaching	0	1	1	2	0	0	0	4.75	0.83
Complexity of the classroom environment in which judgements are made	2	1	1	0	0	0	0	6.25	0.83
Evaluator tendencies toward leniency or severity	0	3	1	0	0	0	0	5.75	0.43
Personal biases and beliefs of the evaluator	0	2	2	0	0	0	0	5.50	0.50
Experiences of the evaluator from observing other teachers	1	2	1	0	0	0	0	6.00	0.71
Prior interactions between the teacher and the evaluator	2	1	0	1	0	0	0	6.00	1.22
Holding a pre-observation discussion	2	0	1	1	0	0	0	5.75	1.30
Level of involvement of the individual being evaluated in the judgement process	0	2	1	1	0	0	0	5.25	0.83
Training of evaluators to use observation criteria for making judgements	1	2	0	1	0	0	0	5.75	1.09

Observation skills of the evaluator	2	0	1	1	0	0	0	5.75	1.30
Perceptual information (cues) available to the evaluator	1	1	0	2	0	0	0	5.25	1.30
Policies regarding evaluation of teaching effectiveness	0	1	2	1	0	0	0	5.00	0.71
Quality of the reasoning strategies used to make decisions	0	2	0	2	0	0	0	5.00	1.00

Note. Questionnaire: Q14–15.

Appendix 8.1

SES Delphi Panel Invitation

University of Glasgow | School of Education

‘Reliability and consistency in judging new teacher practices – why does it matter?’

Expert Delphi Panel

Expressions of interest are invited from leading decision-makers and academics in preservice teacher education for membership of an expert Delphi panel. The purpose of this Delphi panel is to offer an expert opinion to results from a project examining how judgements of teaching effectiveness are made and reliability and consistency in evaluations of teaching skills during school-based clinical experiences. The project involves partnership with teachers, researchers, and university staff of three initial teacher education (ITE) programmes in Scotland, Wales, and England and is funded by a generous award from the Society for Educational Studies. The changing shape of teacher education requires a richer understanding of the nature of judging new teaching effectiveness.

The project which began in August 2022 utilizes a multi-case analysis exploring the nature of judgements regarding ITE students’ performance per normed teaching standards in a comparative, embedded, and descriptive multiple-case study design. A mixed methods approach has guided data collection through an online video observation and associated questionnaire, focus groups, a cross-case synthesis, and now in the last phase, through a consensus building Delphi panel.

The significance of this research relates to the degree to which established norms are challenged in three key aspects: how classroom-based mentor teachers judge ITE students’ performance per normed teaching standards, who institutions rely on to judge teaching effectiveness (i.e., school-based mentor teachers, associate tutors, and university staff), and how ITEs use concomitant judgements of teaching effectiveness amongst a context of power dynamics. The project ultimately seeks to answer the following research questions:

1. What is the nature of shared judgement, consensus, and dissensus of observed teaching effectiveness amongst university staff, associate tutors, and school-based mentor teachers from partner ITE programmes?

2. How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

3. How are the roles of university-based and school-based teacher educators in judging teaching effectiveness in ITE shaped by power dynamics?

The eight (8) Delphi panel members will comprise leading international and national researchers and decision-makers in the field of teacher education. The panel will discuss in depth the recommendations for practice, policy and further research of the main project focused on judgements of teaching effectiveness.

The Delphi panel date has been selected to occur on Thursday, 16 May 2024. The time commitment is a total of six hours; this includes two hours for pre-reading of an executive summary of the findings as preparation (provided a week in advance), and one in-person Delphi panel discussion (approximately four hours in total – meals and breaks excluded). The Delphi panel will be hosted in Glasgow and will be guided by a member of the research team. Travel time to Glasgow should also be considered; all panel member costs will be covered by project funding. An initial panel summary will be considered at the end of the discussion for further comment. The final project report will be available by September 2024.

Appendix 8.2

Delphi Briefing Paper

University of Glasgow | School of Education

SES Society for Educational Studies

Delphi Panel Briefing Paper

'Reliability and consistency in judging new teacher practices – why does it matter?'

Firstly, thank you so much for agreeing to participate in our Delphi symposium. Originally created by The Brookings Institution as a response to Cold War challenges, the Delphi process is intended to assist in clarifying the central strategic questions at stake in any given social/political/cultural practice where the outcomes may, prima facie, be uncertain. While of particular value in moments of substantial stress and urgency it has nonetheless real salience where social practices are often considered indeterminate and contested (Baumfield et al., 2013). The modus operandi is to conduct a series of conversations that each have, as their conclusion, a more refined and focused account of the most important and or urgent questions that require practical redress. We invite you to offer some preliminary reflections on the current state of professional judgement with respect to student/early career teachers' competence relating to their practicum/classroom practice. For each question we would ask you to answer succinctly with no more than 3 observations on each question and no more than 3 sentences on each observation. If you could complete and return your responses by 10 May, we would be grateful. We will then use these observations to shape the scope of the opening panel conversation.

Myriad studies over many years have questioned the veracity and credibility of professional judgement with respect to early teachers' effectiveness and excellence. Report after report has been issued from a disparate group of national and transnational bodies including the OECD and UNESCO; the General Teaching Council Scotland (GTCS) and Council for the Accreditation of Educator Preparation (CAEP) delineating often very long lists of competences, capacities, abilities… that indicate the effectiveness of teacher candidates. These lists are further expanded by increasing numbers of sub-governmental/national bodies that oversee teacher preparation and development, including academy trusts, charter schools, semi-private consortia. A quick glance at the UK's Government (applies only to England) approved list of accredited providers is both fascinating and bewildering (see: https://www.gov.uk/government/publications/accredited-initial-teacher-training-itt-providers/list-of-providers-accredited-to-deliver-itt-from-september-2024).

1a. In your judgement, what are the advantages and disadvantages of having a wide range of providers drawing upon varied and various schemas for assessing effectiveness of beginning teachers?

1b. Do you consider that university teacher educators, associate/link tutors and school-based teacher educators (i.e., mentor teachers) draw upon the same criteria when making judgements? Please explain your response.

1c. Do you consider that there is a common understanding of what particular professional capacities mean in practice for different kinds of assessors? Please explain your response.

1d. In what ways, if any does it matter in theory and in practice if there is disagreement in observations of teacher effectiveness?

1e. How important is it that we encourage consistency? (Please explain your response)

1f. How important is it that we allow for breadth of opinion? (Please explain your response).

1g. How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

The nature of teacher competences has, for many years, been centrally determined by governments, professional accrediting bodies or non-departmental public bodies. They are often highly prescriptive but open to wide interpretation. Some practices common (particularly but not exclusive to issues of interpersonal relationships, pastoral welfare, classroom control and so forth) in previous generations would no longer be considered appropriate. The development of competence and related frameworks was intended to replace the assumed and/or tacit knowledge gained through experience often used to determine the quality of student/early career teachers. Tacit knowledge had often been a point of contention between different parties involved in student teacher practical assessment (Bullogh and Draper, 2004).

2a. What kinds /sources of evidence do you consider to be most important in coming to a judgement of teacher effectiveness? Note up to four.

2b. To what extent do you consider judgements based on tacit knowledge to be important in assessing student teacher quality?

2c. What do you consider the relationship between this tacit knowledge and centrally determined competences is?

2d. What do you consider the relationship between tacit knowledge and such centrally determined competences should be?

2e. How might the roles of university-based and school-based teacher educators in judging teaching effectiveness in initial teacher education be shaped by power dynamics?

Much has been said and written in recent months about the creative and disruptive potential of artificial intelligence (AI). It is, as yet unclear to what extent the much vaunted and heavily rhetorised claims will come to fruition, but it seems inconceivable that technology will not play an increasing role in the judgements of teacher quality and effectiveness.

3a. What role is AI likely to play in teacher assessment?

3b. What role should AI play in assessing early career teachers?

3c. What might be the advantages and disadvantages of using AI in making judgements of early career professionals?

We kindly ask that your observations be returned via email by 10 May to Sarah Anderson, project PI, at sarah.anderson.3@glasgow.ac.uk.

References

Baumfield, V.M., Conroy, J.C., Davis, R.A. and Lundie, D.C. (2011) The Delphi method: gathering expert opinion in religious education, *British Journal of Religious Education*.

Bullogh, R.V. and Draper, R.J. (2004) Making sense of a failed triad, *Journal of Teacher Education*, 55:5, 387–484.

Appendix 8.3

SES Delphi Panel Agenda 16 May 2024

University of Glasgow
School of Education

SES Society for Educational Studies

'Reliability and consistency in judging new teacher practices – why does it matter?'

May 15 Social

19:30	Pane social gathering

May 16 Delphi Panel Agenda

8:45-9:00	Arrival at meeting room
9:00-10:00	Breakfast, Introductions and Synopsis of Responses – Key Points
10:00-12:00	Discussion Period 1
12:00-13:00	Lunch
13:00-15:00	Discussion Period 2
15:00-15:30	Break (with tea/coffee and cakes)
15:30-17:00	Discussion Period 3 – Plenary with research team
17:30	Optional – Open House Event

Confirmed Expert Panel Members

1. [Redacted]
2. [Redacted]
3. [Redacted]
4. [Redacted]
5. [Redacted]
6. [Redacted]
7. [Redacted]
8. [Redacted]
9. [Redacted]

Research Team Onsite for Panel

- Sarah Anderson – University of Glasgow Senior Lecturer and Project PI
- Professor Jim Conroy – University of Glasgow and Project Co-I
- Pinky Jain – Leeds Beckett University
- Professor Rachel Lofthouse – Leeds Beckett University
- Sevda Ozsezer-Kurnuc – University of Glasgow, Research Associate

Addition Research Team Members

- Professor Andrew Davies – Swansea University
- Daryl Phillips - Aberystwyth University

Logistical Information

Nearest subway station: Hillhead

Meeting room: https://frontdoor.spa.gla.ac.uk/map/embedded.html#!/?to=1040456

Accommodations

Recommended accommodations near the University of Glasgow

- Glasgow Grosvenor Hotel
- Sandyford Lodge Hotel
- Alamo Guest House
- Acorn Hotel

Visit Scotland: https://www.visitscotland.com/places-to-go/glasgow/accommodation

Reimbursement of Expenses

The information below is required for non-staff reimbursement claims.

Beneficiary Account name:
Beneficiary Address:
Beneficiary Bank Name:
Sort Code(SWIFT/BIC):
Account Number (IBAN):

USA – digit ABA routing number
Canada – 5 Digit Transit code
Australia – 6 Digit BSB number

If there is an Intermediary bank, the SWIFT code and Account No. of that bank:

Amount to be claimed:

Bank verification - screenshot of bank statement showing account details (no transactional details) OR pre-printed bank account pay in slip OR screenshot of scored through cheque.

Receipts:

- Send to: School of Education Finance and Resources education-finance@glasgow.ac.uk
- Reference: SES 2022 National Award
- Copy to sarah.anderson.3@glasgow.ac.uk

Appendix 8.4

Delphi Synopsis

‘Reliability and consistency in judging new teacher practices – why does it matter?’

Delphi Symposium
16th May 2024

1a. In your judgement, what are the advantages and disadvantages of having a wide range of providers drawing upon varied and various schemas for assessing effectiveness of beginning teachers?

Advantages

- Breadth avoids parochialism
- Allows for comparison
- Allows for breadth of interpretation about the import of certain teaching/educational outcomes
- Broader levels of discernment
- Heterogeneity offers some reflection of the complexity of the demands
- Challenges consensus

Disadvantages

- Fit for particular circumstances/not the profession
- Too frequently default to personal preferences
- Too loose and the teacher student struggles to understand the expectations
- Unreliability of judgement
- Too many opinions to offer much discernment and with too little experience
- Third parties an unnecessary burden on the system
- Waste/inefficiencies/redundant competition

1b. Do you consider that university teacher educators, associate/link tutors and school-based teacher educators (i.e., mentor teachers) draw upon the same criteria when making judgements? Please explain your response.

- It requires maintenance, servicing, vigilance iterativity and collaboration
- The exigencies of the local/pressing determine evaluation
- College based – more general
- Even where there are generic frameworks local practice/interpretation differ much
- Assessment tools have produced greater consistency
- Different stakeholders have diff emphases – culture/exam/differentiation (more political)
- People bring their diff experiences so make diff judgements
- (Collaborations produce more consistency)
- Diff experiences bring different judgements
- University – ideal; school practical

1c. Do you consider that there is a common understanding of what particular professional capacities mean in practice for different kinds of assessors? Please explain your response.

- Yes – where structured opportunities/coherent training available
- Common understanding undermined by influence of the implicit
- Initial agreement often gives way to local practices – political/public and political forces
- Only where there is collaboration is there agreement
- Standards provide a basis for consistency
- Understand capabilities capacities in constantly changing worlds
- Lack of exemplification makes coherence challenging

1d. In what ways, if any does it matter in theory and in practice if there is disagreement in observations of teacher effectiveness?

- Only matters to the extent that we think teachers should demonstrate skills rather than capabilities
- Disagreement/productive tension is important to system health/disagreement – as indeed it feeds a consensual agreement
- Forcing consistency may undermine other educational values such as student need
- The lack of consistency allows different discourses to prevail – many professionally immature mentors may be a problem
- Different observers can be foregrounding different aspects of the teaching endeavour
- Why would disagreement matter?
- The complexity of children and context make some inconsistency inevitable and simulation can only take us so far
- Flexibility and accommodation to circumstance

How important is it that we encourage consistency? (Please explain your response)

- Very important at the level of standards but flexibility in implementation
- Very important but with clarity as to where authority resides
- Very important as incoherence can leak into the educational experience of children – can be chaotic
- As a gatekeeping strategy very important
- Consistency of opportunity to develop appropriate dispositions is very important
- Important yet developmental and recuperative
- As an instrument of support
- Forced consistency at odds with the conceit of teaching individuals
- The idea that there is one best way is misleading
- It is a challenge when colleagues challenge the need for some consistency – but with accommodation
- Consistency of measurement with regard to a skills framework more imp. than consistency of observation

1f. How important is it that we allow for breadth of opinion? (Please explain your response).

- Breadth of opinion conduces to being human
- Judging human practice is qualitative/subjective (ARE THESE THE SAME THING?)
- Without diversity of opinion there is a danger that certain concerns come to dominate (e.g. behaviour)
- Opinion for its own sake may just be that!
- Important but must be evidenced based
- As a condition of
- Important as teacher mentor's opinions may be no more than that!
- Necessary as there are so many sides to teaching
- We cannot escape complexity by pretending it doesn't exist

1g. How might enhanced reliability of professional judgement foster greater collaboration between schools and universities?

- Need to develop greater school-based mentor confidence
- This was the essence of the clinical model vouchsafed by both shared international research evidence and respected regulatory governance
- Too much acquiescence to university-based tutors
- The development of competence-based frameworks may be considered an antidote to tacit/presumed knowledge
- Greater inter-institutional trust and reliability as a consequence of it –construction not the other way around

1a-g. QUESTIONS

1. **MANY OF YOU ARGUED IN FAVOUR OF INDIVIDUAL JUDGEMENT GUIDED BY ADHERENCE TO AGREED COMPETENCE FRAMEWORKS – HOW SPECIFIC SHOULD WE BE? QUALITATIVE JUDGEMENTS MAY BE TOO POSITIONALLY FREIGHTED TO BE OF MUCH GOOD!**

2. **WHAT, IF ANYTHING, IS ACTUALLY LOST IN IMPOSING INSTRUMENTS/PRACTICES OF CONSISTENCY?**

3. **THERE SEEMED TO BE SOME AMBIVALENCE ABOUT THE ROLE OF TACIT KNOWLEDGE – AS A LOCAL PHENOMENON WORTHY BUT TOO OPINION LADEN/WHY SHOULD WE FAVOUR DIFFERENT EXPERIENCES OVER CONSISTENCY?**

4. **WHERE ARE THE LIMITS OF OUR FLEXIBILITY? HOW ARE WE TO DETERMINE THEM?**

5. **CAN WE ARRIVE AT A DIFINITIVE 'COMMAND' LIST OF COMPETENCES?**

6. **WOULD TEACHER EDUCATION BE IMPROVED IF WE WERE ABLE TO DRAW ON AN OECD (OR SIMILAR) MANDATED COMPETENCE FRAMEWORK THAT WAS INTERNATIONALLY CONSISTENT?**

2a. What kinds /sources of evidence do you consider to be most important in coming to a judgement of teacher effectiveness?

- Synoptic judgements of both performance and personality by competent/expert staff
- Learning response of the young people
- Progress of young people
- Motivation, engagement and organisation of the young people
- Good/sound subject and curriculum knowledge
- Openness to research in developing an understanding of what works within the classroom
- Behaviour management through effective planning and adaptation to learning needs
- Challenge for all pupils
- Evidence of Individual and whole class student engagement in learning and evidence of pupil growth
- Demonstrated capacity to sustain relationships with students and their communities which respects diversity and context
- Teacher professional expertise – judgement, knowledge, ability to build new knowledge
- Classroom management
- Student feedback
- Family feedback
- Teacher sanity
- Self-reflexiveness
- Self-awareness
- Practical knowledge
- Ethical awareness – esp. in relation to the other
- Disposition and engagement of students in the classroom with each other and with the educators in the classroom
- Students progressing in their learning, beyond just reading and math, in socialisation and in movement
- Observations of students integrating and working together
- Being able to meet grade-expected competencies
- Educators in the classroom

2b. To what extent do you consider judgements based on tacit knowledge to be important in assessing student teacher quality?

- Is phronesis a better term? While elusive is a part of the enterprise
- Ok as long as it isn't implicit
- Space for disagreement
- Multiple observations points is important
- Important but difficult to use systemically – needs articulation
- Caution in preferencing it
- School context v important which is why we need multiple placements
- Difficult to delineate/articulate but you know it is there when you see it
- Deeply embedded but objectivity only emerges in the surfacing and articulation

2c. What do you consider the relationship between this tacit knowledge and centrally determined competences is?

- Competence-based learning a response to post-War technocratic short-circuiting of tacit knowledge but ultimately there are questions about its import – reduction and attenuation
- Either integrated or marginalised if it negates the aims and values of the school system
- Competences provide a framing of the tacit
- Determines actions but stands outside the framing of competences – practice is its own thing
- Competence Guide; tacit knowledge may make the difference!!
- A part of the repertoire but needs to be subjected to scrutiny
- Context can shape the assumed tacit framing – hence the need for more than one placement
- Tacit knowledge can be influential but must be subjected to the common framings
- The worth of tacit knowledge emerges later

2d. What do you consider the relationship between tacit knowledge and such centrally determined competences should be?

- Dynamic equilibrium – eventually pushed into a narrative space
- Valorise tacit knowledge – needs to be there but contextualized in the broader framing
- University tutors broad conspectus should enable students to have their performance contextualized
- A secondary consideration subordinated to judgements made on the basis of agreed framing
- Agreed framing should have priority early and tacit knowledge later
- Rather than see competences as determinative of success and failure, can we ask why someone may appear to fail – more context driven

2e. How might the roles of university-based and school-based teacher educators in judging teaching effectiveness in initial teacher education be shaped by power dynamics?

- Inherent in a highly politicised space – diffusion of roles/responsibilities can disarm asymmetries in powers
- Come into play where there is an attempt to impose one's values/perspective on another
- Power brokers outside the primary responsibles (e.g. Unions)
- Lots of the determinants surfaced here require political decision-making
- Whether teachers or academics the power favours the older/more experienced
- Teacher ed as a site of continuous arm wrestling
- Many mentors defer their ultimate decision
- Gatekeeper bias
- Little evidence in a well-constructed partnership

2a-e. QUESTIONS

1. **SHOULD WE PAY MORE ATTENTION TO THE PERSONALITY OF THE TEACHER CANDIDATE?**
2. **IS TACIT KNOWLEDGE NECESSARY IN MAKING JUDGEMENTS ABOUT ECTS PROGRESS/ACHIEVEMENT?**
3. **CAN WE JUSTIFY JUDGEMENTS MADE ON THE BASIS OF TACIT KNOWLEDGE?**

3a. What role is AI likely to play in teacher assessment?

- Currently At Sea but some potential in emerging technologies
- AI as a teacher coach assimilating wide swathes of observation/research/analysis
- Synthesising perspective-laden examples
- Personal planning assistants – tailored provision but can this capture the human dynamic
- Will test for the effectiveness of lessons
- Classifying evidence as related to given skills
- Can show what is going on but not the quality of the interactions
- True judgement can't be made by a machine!
- Those adept in the deployment of AI will dictate the performance criteria

3b. What role should AI play in assessing early career teachers?

- A supporting role – analysing transactions
- Enhance decision-making/judgement
- Enhance reliability and consistency
- Planning assistant
- Building individual learning profile activities…
- Quantify and shape classroom management processes
- Clarify our language/criteria…
- Use it to become more productive/accurate
- Significant in collating and analysing performance data/portfolio data
- Probably no role – judgement of this kind is a distinctively human activity; it is a conversation

3c. What might be the advantages and disadvantages of using AI in making judgements of early career professionals?

Advantages

- Enhanced consistency
- Systematising the data
- Scan, collate, synthesise myriad sources and turn them into tools
- None!

Disadvantages

- The potential to undermine what Wittgenstein/Arendt would call 'The Forms of Life – relationality and embodiment
- Risk of human/cultural bias
- Risk of algorithmic amplification of bias
- Unmoored from context
- Lack of innovation/creativity in feedback
- A skills based tool may not be able to comprehend 'style'
- Lack of transparency/black box phenomenon
- Teacher ed prof judgement becomes redundant
- Context light
- AI doesn't make judgements

3a-c. QUESTIONS

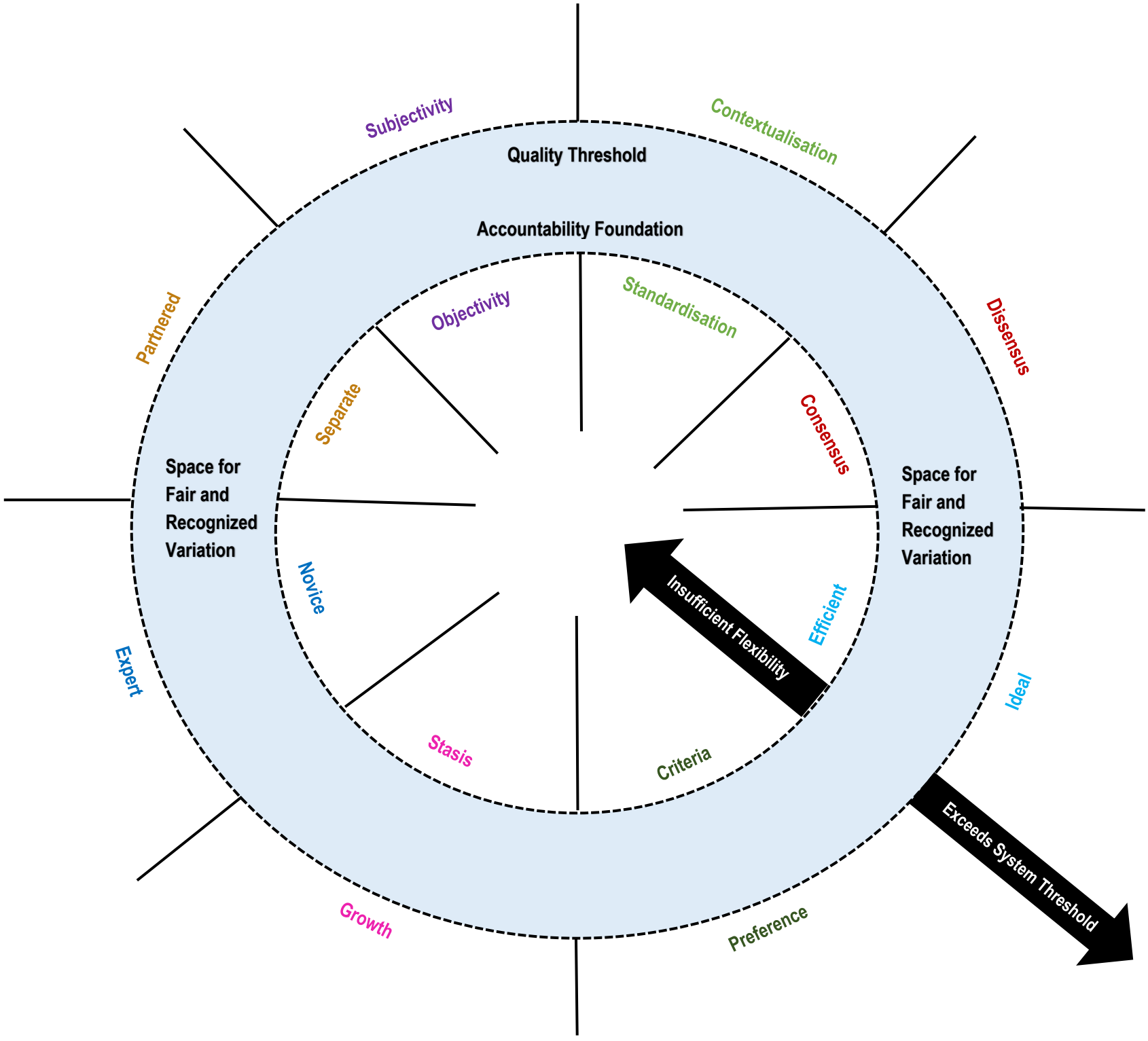
1. **GENERALLY NEGATIVE COMMENTS ON THE USE OF AI GIVEN THAT JUDGEMENT IS A HUMAN ACTIVITY – BUT TO WHAT EXTENT MIGHT THIS BE DENIAL?**

2. **WHAT HAPPENS IF/WHEN A VOICE-ACTIVATED AI VIDEO MONITORING SYSTEM CAN FEED AN AI 'BRAIN' CAN MAKE CONSISTENT JUDGEMENTS THAT, IN ALL RESPECTS, MIRRORS PRECISELY HUMAN FORM?**

3. **IN WHAT WAYS SHOULD/MIGHT WE VOUCHSAFE THE DISTINCTLY HUMAN LEXCION OF LIFE?**

Appendix 10.1

Duplexity Model


Subjectivity
Contextualisation
Quality Threshold
Accountability Foundation
Partnered
Dissensus
Objectivity
Standardisation
Separate
Consensus
Space for Fair and Recognized Variation
Space for Fair and Recognized Variation
Novice
Efficient
Insufficient Flexibility
Expert
Ideal
Stasis
Criteria
Exceeds System Threshold
Growth
Preference