British Journal of Educational Studies

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rbje20

Judging Student Teacher Effectiveness: A Systematic Review of Literature

Sarah K. Anderson, Sevda Ozsezer-Kurnuc & Pinky Jain

To cite this article: Sarah K. Anderson, Sevda Ozsezer-Kurnuc & Pinky Jain (16 Jul 2024): Judging Student Teacher Effectiveness: A Systematic Review of Literature, British Journal of Educational Studies, DOI: 10.1080/00071005.2024.2374070

To link to this article: https://doi.org/10.1080/00071005.2024.2374070

British Journal of Educational Studies
Vol. 00, No. 00, 2024, pp. 1–33

Judging Student Teacher Effectiveness: A Systematic Review of Literature

By SARAH K. ANDERSON, *School of Education, University of Glasgow, Glasgow, UK*, SEVDA OZSEZER-KURNUC, *School of Education, University of Glasgow, Glasgow, UK; Turkish Ministry of National Education* and PINKY JAIN, *Carnegie School of Education, Leeds Beckett University, Leeds, UK*

ABSTRACT: This paper reports on a systematic literature review to understand better methodologies and data collection tools used to judge student teaching effectiveness, ways in which validity and reliability are considered, the processes involved in assessing new teaching effectiveness within teacher education programmes, and how evaluation and results are used to judge readiness to teach. The accurate and consistent judgement of teaching competence during and at completion of preparation continues to be an area of increasing interest and concern. The PRISMA review process identified 45 key papers. An in-depth analysis underscored several crucial factors, such as the challenge of ensuring the reliability of judgements within dynamic educational environments and the need for broader understanding and applications of reliability and dependability when making judgements. The findings of this systematic literature review hold implications that merit consideration by teacher education programmes for processes to judge teaching effectiveness. The analysis also highlighted the intricacies inherent in evaluating teaching effectiveness, alongside ongoing discourse regarding the criteria and measures for judging competence of student teachers.

Keywords: educator preparationteacher education programmesjudgementreliabilityvalidityinitial teacher education

1. INTRODUCTION

Evaluation of student teachers' readiness to teach is a central component of high-quality teacher preparation, which is often assessed during practice placements in schools. The accurate and consistent judgement of teaching competence during preparation continues to be an area of increasing interest and concern (Asher, 2018; Haigh *et al*., 2013; Schmoker, 2023; Seidenberg, 2017), particularly during this era of high accountability and increased scrutiny of student teacher preparation. Darling-Hammond (2017) argued that there is a robust relationship between high-performing school education systems, the quality of student teachers, and robust intellectual and

ISSN 0007-1005 (print)/ISSN 1467-8527 (online)

https://doi.org/10.1080/00071005.2024.2374070
http://www.tandfonline.com

professional barriers to admission into the profession. Such 'appropriate' barriers include effective assessment of pre-service and early career teachers. Hattie (2023) stated that while teacher education programmes claim to have a comprehensive set of core attributes to determine competence, this core remains different across providers and systems in a state that Levine (2006, p. 109) labelled 'unruly' and 'disordered'. Raths and Lyman (2003) suggested that many student teachers manage to pass into the profession despite manifesting significant incompetence due to failures of professional agreement as to what constitutes a judgement of competence.

An inspection approach continues to take shape and dominate discourse in teacher education, in what has been referred to as a crisis in teacher education provision (Mutton and Burns, 2024). This has sparked discord amongst teachers and teacher educators alike concerning perceived disproportionate levels of accountability in the form of high-stakes observations and evaluations, and performative measures during preparation, including those used in decision-making for entry into the profession. Interestingly, in a review of 32 studies by Klassen and Kim (2019), findings revealed only small correlations of both academic and non-academic criteria during preparation as predictors of 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. Indeed, Haigh and Ell (2014) found that university and classroom-based teacher mentors take an 'idiosyncratic approach' (p. 19) to reaching decisions about teaching, and even where judges have a shared vision of quality teaching, significantly different findings often emerged. Such failures of agreement are not uncommon and can be attributed to a range of factors (i.e., contextual differences, time constraints, and asymmetric attributions of importance). There are implications from this variability and dissensus amongst judgement-making to be explored. Amongst complicating factors shared across the UK is the increasing reliance on adjunct and school-based supervisors and a perceived disconnect between theory and practice. 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).

Rather than see this as an enduring problem, this systematic literature review seeks to understand better methodologies and data collection tools used to judge teaching effectiveness of student teachers, ways in which validity and reliability are considered, the processes involved in assessing new teaching effectiveness within teacher education programmes (TEPs), and how evaluation and results are used to judge readiness to teach. The aim is to better inform practices in teacher education, the experience of evaluators and student teachers, and to contribute to the conversation around what teaching quality means. This article reports results of a substantial and systematic review on judging teaching effectiveness and factors impacting

reliability and validity. The review is part of a multi-phased study which also included a comparative policy analysis of teaching standards and a descriptive multiple-case study with mixed-methods data collected. In other scholarly works, we report on the outcomes of the larger study informed by this systematic review (Anderson *et al.*, 2023). In this review, we shed light on prevailing trends and identified gaps in the literature, guiding areas that demand further investigation.

2. Conceptual Framework

Social judgement theory (SJT) as a system-orientated perspective of understanding human judgement in specific ecological circumstances supported and informed the overall project inclusive of this review (Cooksey, 1996; Hammond *et al.*, 1977). As a framework, SJT guides enquiry through eight stages: conceptualise the judgement problem, understand the ecology, identify relevant cues and dimensions for judgement, sample cue profiles, sample judges, obtain judgements, capture policies, and compare policies (Cooksey, 1996). This systematic review was explicitly intended to address stage one, to conceptualise the nature of the problem (i.e., judging teaching effectiveness), and the evidence utilised for making those judgement decisions. Thus, the approach maintained recognition of the complexities of exploring judgements. This multi-phased project recognised that judgement is both a cognitive act and a socially positioned practice (Allal, 2013). Therefore, judgement of student teachers' performance will be dependent on a myriad of factors, including complex surroundings, normed teaching standards, and variations in decision-making and evaluation tools. This acknowledgement informed the selection of search terms, consideration of researcher positionality, and the iterative approach to analysis.

3. Methodology

To better conceptualise what SJT notes as the 'judgement problem', this review sets out to explore the most recent evidence related to nature of student teacher judgement in both UK and international contexts in addition to documenting methodological tendencies. A systematic literature review can be considered 'the most reliable and comprehensive summary about "what works" in a given field' (Van Der Knaap *et al.*, 2008, p. 49) and was thus chosen. The review was conducted using the PRISMA framework (Page *et al.*, 2021), while also following established methods of systematic review from Bryman (2016). The review process was guided by three key questions, which played a vital role in identifying search terms and establishing inclusion and exclusion criteria:

- What types of evidence are used to make judgements of teaching effectiveness?

- What is the nature of reliability and validity of evaluations/tools used to judge student teachers?
- How are results of judgements of teaching effectiveness used in the research literature?

Search Strategy and Inclusion Criteria

The search was conducted in January 2023 and utilised electronic searches across 19 education-focused databases using specified terms (see Appendix A in Anderson *et al.*, 2024). To be included in this review, texts had to meet the following inclusion criteria:

- peer-reviewed;
- English language and published within or outside of the UK;
- published from 2010 to 2023;
- address a relevant aspect regarding the nature of student teacher judgement related to:
  - how student teachers are evaluated in their teaching practice;
  - criteria used to judge teaching effectiveness;
  - validity and trustworthiness of evaluation instruments;
  - relationships between rater groups (i.e., classroom teachers, university staff);
  - related to rater's judgement of student teachers;
- demonstrate high-quality methods (i.e., presence of research question, alignment between methodology, analysis, findings, and conclusions);
- pertaining to initial teacher education.

A ten-year parameter was considered but expanded to account for impact of the Covid pandemic years. The review therefore did not include any pre-print or in press publications, non-peer reviewed articles, book chapters, books, or government documents. Searches were not limited by research methodology, encompassing a range of empirical and non-empirical studies, such as those which utilized secondary analysis of programmatic and administrative data.

Screening Literature

Studies were screened in a multi-step process (see Figure 1) to identify relevant literature. The second author conducted the database search and retrieved the studies, and the first author confirmed results. Studies were exported to the Rayyan software application to facilitate the screening processes (Ouzzani *et al.*, 2016); irrelevant papers were excluded by screening abstracts and titles. A total of 632 peer-reviewed articles were retrieved; 601 were identified as an initial sample with 31 duplicates

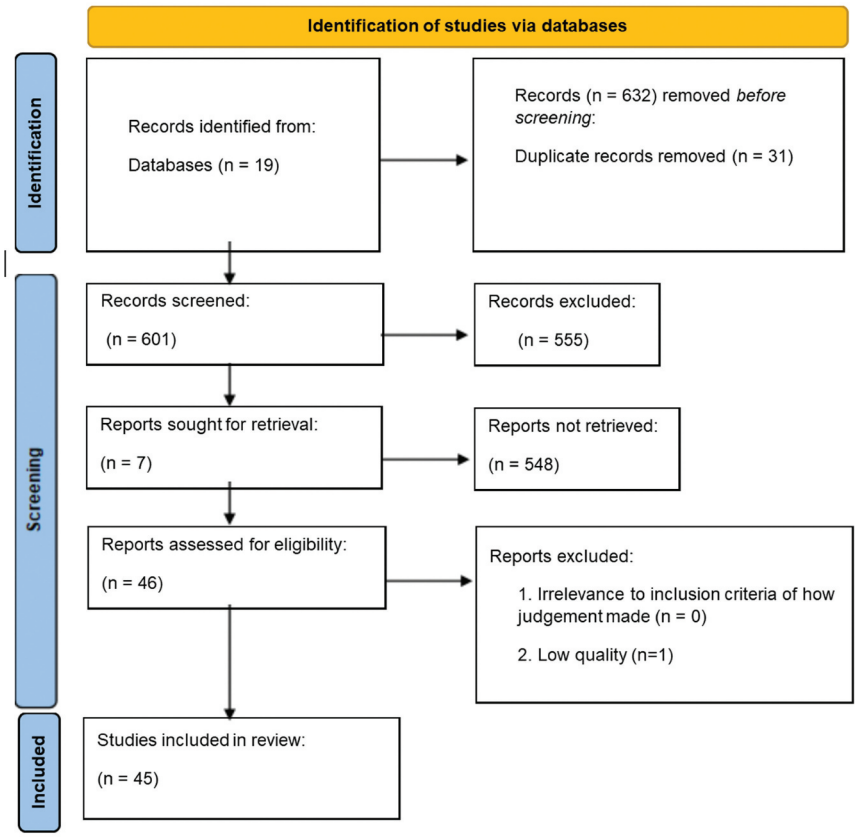

Figure 1. PRISMA flow diagram of screening (adapted from Page *et al*., 2021)

removed. Abstracts and titles of these were screened and reviewed by the second author; 555 were excluded based on unanimous agreement of first and second author. There was an initial disagreement regarding the inclusion of seven papers, which was successfully resolved through discussion. The process resulted in a high percentage agreement of 98.8% among the authors and Cohen's Kappa coefficient was calculated at 0.91, indicating 'almost perfect agreement'. Following this, 46 publications were deemed suitable for full-text screening and review for relevance and quality. One study was excluded due to lacking research purpose, questions, and clear methodology. Screening led to inclusion of 45 studies which underwent full summary and extraction. These studies are referred to in the text, Table 1, and Appendix B by number (see Appendix B in Anderson *et al.*, 2024 for full citations).

Table 1: A map of research focuses, methodology and identified evaluation tools in examined literature

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 1: Hylton et al. (2002)	Validation of a candidate evaluation instrument, developed and used by a TEP	V: construct validity*	Candidate evaluation with an authentic tool, Single TEP in a university, US	Empirical, Quantitative, Secondary	Pre-existing evaluation results (n = 1,486 mid and final ratings) by candidates, school-based and university-based teacher educators.
Study 2: Dewaele et al. (2021)	Influential factors of preservice 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**, Primary	Evaluation instrument (n = 266 candidates about a teacher), including a comment section for explanations.
Study 3: Tobón et al. (2021)	Validation of an emerging evaluation tool (SOCME-10) for new teacher population	V: construct validity*	Schoolteacher evaluation, No TEP, Mexico	Empirical, Mixed**, Primary	Questionnaires (n = 21 experienced schoolteachers* and 25 new teachers), a comment section. Evaluation instrument (n = 557 new teachers self-rating). **Expert.*

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 4: Tanguay (2020)	Perspectives of university teacher educators on edTPA as a standardised state mandate tool	V: face validity*	Candidate evaluation with an authentic tool, Single TEP in a university, US	Empirical, Qualitative, Primary and secondary	Interviews (n = 8 university-based teacher educators). Documents of candidates and programme, i.e., artifacts, program workshop materials.
Study 5: Sandoval et al. (2020)	a teacher education programme's alignment with equity outcomes, edTPA	V: consequential validity	Candidate evaluation with an authentic tool, Single TEP in a university, US	Empirical, Qualitative, Secondary	Documents of candidates (n = 53 course essays, n = 9 portfolios)
Study 6: Roloff et al. (2020)	Predictive validity of entry characteristics and grades from teacher education on teacher's future instructional quality	V: predictive validity	Early career teachers' evaluation in school context, 2 + TEPs across universities, Germany	Empirical, Quantitative, Primary and secondary	Evaluation instrument (n = 3,768 classroom student's ratings on 113 schoolteachers). Questionnaires (n = 113 teachers). Administrative records of teachers during teacher education.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 7: Shahzad and Mehmood (2019)	Development and validation of an emerging higher education lecture evaluation tool (TES) to be used by university students	V: construct validity*	Higher education lecturer evaluation, No TEP, Pakistan	Empirical, Mixed, Primary and secondary	Interviews (n = 10 higher education lecturers*), focus group interviews (n = 3 groups of graduates), and literature search. Questionnaires (n = 16). Evaluation instrument (n = 698 higher education students). **Expert.*
Study 8: Yahiji et al. (2019)	Examination of a assessment model used in field experience (validity, reliability, objectivity, practicality)	V: face validity	Candidate assessment, Single TEP in a university, Indonesia	Empirical, Mixed, Primary and secondary	Questionnaires and focus group interviews (n = 14 university-based teacher educators*, n = 14 school-based teacher educators*). Documents of candidates, i.e., assignments. **Expert judgers*

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 9: Mkhasibe, et al. (2018)	Comparing teacher mentors and university supervisors' perception of student teacher's readiness to teach	R: consistency and accuracy*	Candidate assessment, Single TEP in a university, South Africa	Empirical, Qualitative, Primary and secondary	Focus group interviews (n = 12 school-based teacher educators). Pre-existing evaluation reports (observation) (n = 3 university-based teacher educators).
Study 10: Basit and Khurshid (2018)	Satisfaction level of teacher educators and candidates with candidate assessment techniques	V: face validity	Candidate assessment, 2+ TEPs across universities, Pakistan	Empirical, Quantitative, Primary	Questionnaire (n = 300 university-based teacher educators, 890 candidates).
Study 11: Ata and Kozan (2018)	Construct validity and reliability of Intern Keys, i.e., interpretable factor structure	V: construct validity*	Candidate evaluation with an authentic tool, 2+ TEPs across universities, US	Empirical, Quantitative, Primary	Evaluation instrument (n = 116 university-based teacher educators)

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 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	Pre-existing evaluation results (n = 2,362 candidate portfolios). Administrative records of candidates, and for employed teachers (n = 277) student's achievement
Study 13: Kennedy and Lees (2016)	Candidate's growth through CLASS scores supported feedback and tiered support	V: consequential validity	Candidate evaluation with an authentic tool, Single TEP in a university, US	Empirical, Mixed, Primary and secondary	Focus group interviews and pre-existing evaluation results (n = 19 candidates)
Study 14: Masuwai and Saad (2016)	Face and content validity (representativeness, relevance) of an evaluation instrument (TLGPI)	V: content validity*	Teacher educator assessment, 2+ TEPs across universities, Malaysia	Empirical, Mixed**, Primary	Questionnaires (n = 9 university-based teacher educators*), with a comment section. * *Expert judgers.*
Study 15: Brown et al. (2015)	Documenting candidates' professional growth through PEI	V: consequential validity*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, Quantitative, Secondary	Pre-existing evaluation results (n = 97 candidates) by candidates, school-based and university-based teacher educators.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 16: Maharaj (2014)	School administrator's view of a teacher performance appraisal model (TPA), used for new and experienced teachers	V: face validity*	Schoolteacher evaluation, No TEP, Canada	Empirical, Mixed**, Primary	Questionnaires (n = 166 school principals), with a comment section.
Study 17: Kingsley and Romine (2014)	Construct validity, dimensionality, and reliability of I-LAST, a learning-oriented evaluation tool	V: construct validity*	Candidate assessment, Single TEP in a university, US	Empirical, Quantitative, Primary	Evaluation instrument (n = 46 school-based teacher educators for candidate's ratings). Questionaries (n = 3 university-based teacher educators, and n = 3 school-based teacher educators)
Study 18: Hamid et al. (2012)	Predictive effect of teacher cognitive ability and personality in performance	V: construct validity*	Schoolteacher evaluation, No TEP, Malaysia	Empirical, Quantitative, Primary	Evaluation instrument (n = 1366 schoolteachers self-rating).
Study 19: Smalley and Retallick (2012)	Evaluation practices in agricultural teacher education programmes	JM: Instrument implementation and result use*	Candidate assessment, 2+ TEPs across universities, US	Empirical, Quantitative, Primary	Questionnaires to coordinators (n = 66 agricultural education teacher preparation programmes)
Study 20: Ritzhaup et al. (2010)	Candidate's perspectives of e-portfolios, i.e., whether supports them	V: face validity	Candidate assessment, 2+ TEPs in a university, US	Empirical, Quantitative, Primary	Questionnaires (n = 224 candidates)

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 21: Beare et al. (2014)	If employment supervisors show bias based on new teachers' socioeconomic status and ethnicity, using SEPTPP.	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 by employment supervisors regarding new teachers' preparedness to teach, including yearlong ratings.
Study 22: Behizadeh and Neely (2018)	Consequential validity of edTPA, in a social justice oriented TEP	V: consequential validity*	Candidate assessment, Single TEP in a university, US	Empirical, Qualitative, Primary	Reflective commentary (n = 16 candidates)
Study 23: Bell et al. (2018)	Administrators' judgment accuracy: assessment of thinking and reasoning strategies	R: consistency and accuracy*	Schoolteacher evaluation, No TEP, US	Empirical, Mixed, Primary	Evaluation instrument (n = 35 school principal's rating on a teachers) and think-aloud exercises during rating.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 24: Chaplin et al. (2014)	Correlation between and amongst teacher effectiveness measures: RISE, 7Cs, VAM	R: consistency and accuracy*	Schoolteacher evaluation, No TEP, US	Empirical, Quantitative, Secondary	Pre-existing evaluation results (n = 329 teachers) by school principals and classroom students. Administrative records of student achievement of these teachers.
Study 25: Conderman and Walker (2015)	Examining similarities between candidate and instructor's concerns of candidate's dispositions	R: consistency and accuracy*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, Quantitative, Primary	Evaluation instrument to gather ratings from 248 candidates and 80 university-based teacher educators regarding candidate's disposition exhibitions.
Study 26: Choi et al. (2016)	Reliability and validity of a candidate dispositions rating form developed and used in a TEP (TEDRF)	R: internal consistency reliability*	Candidate evaluation with an authentic tool, Single TEP in a university, US	Empirical, Quantitative, Primary	Evaluation instrument (n = 147 candidates, mid and final ratings) by university-based and school-based teacher educators. Evaluation instrument for candidate's engagement with students.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 27: Johnston et al. (2018)	Advancing psychometric assessment of nine previously validated dispositional indicators (EDA)	V: construct validity*	Candidate assessment, 2+ TEPs across universities, US	Empirical, Quantitative, Primary	Interviews to quantify (n = 22, university-based* and school-based teacher educators* and candidates*). Questionnaires (n = 16), using Q-Sort procedure. **Expert*
Study 28: Lazarev et al. (2017)	T-TESS rubric's ability to distinguish teacher's teaching quality, and internal consistency	R: internal consistency reliability*	Schoolteacher evaluation, No TEP, US	Empirical, Quantitative, Secondary	Pre-existing evaluation results (n = 8,250 records) by qualified raters. Administrative records of 251 schools.
Study 29: Lyness et al. (2021)	inter-rater reliability of portfolios scored by 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, Primary and secondary	Evaluation instrument (n = 19 portfolios by 2 local raters). Interviews with 10 raters. Pre-existing evaluation results of double-scored portfolios as 'true scores'.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 30: Montecinos et al. (2010)	Consequential validity of a candidate evaluation called Samples of Teaching Performance (STP)	V: consequential validity*	Candidate assessment, 2+ TEPs across universities, Chile	Empirical, Mixed, Primary	Evaluation instrument (n = 24 reports by 2 school-based teacher educators). Questionnaires (n = 62 candidates, n = 40 school-based teacher educators) with a comment space. Focus groups (n = 47 candidates, n = 40 school-based teacher educators)
Study 31: Murley et al. (2014)	Inter-rater reliability between university course instructors and trained project participants; perspectives of scoring prompts, rubrics (TWS)	R: consistency and accuracy*	Candidate evaluation with an authentic tool, 2+ TEP in a university, US	Empirical, Mixed, Primary and secondary	Evaluation instrument and feedback form (n = 100 teacher work samples) by university-based and school-based teacher educators. Pre-existing evaluation results of work samples as 'true scores'.
Study 32: Papanastasiou et al. (2012)	Examining the coherence between the programme and state standards in a TEP	JM: Instrument development	Candidate assessment, Single TEP in a university, US.	Empirical, Qualitative, Secondary	Documents of programmes i.e., portfolio creation guidelines, standards, lesson plans.

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 33: Parkes and Powell (2015)	Commenting on problems and alternatives about edTPA	V: Predictive validity*	Candidate assessment, No TEP, US	Scholarly written, Qualitative, *N/A*	No data collection
Study 34: Pufpaff et al. (2015)	rater agreement pre and after digital training	R: Consistency and accuracy*	Candidate assessment, Single TEP in a university, US	Empirical, Mixed, Primary and secondary	Evaluation instrument and questionaries (n = 10 university-based teacher educators for candidate assignments). Pre-existing evaluation results of course instructors as 'true scores'.
Study 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 (n = 4,681 cases, three yearlong mid and end term) by candidates, school-based and university-based teacher educators.
Study 36: Tait-McCutcheon and 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, Primary	Evaluation instrument (n = 34 candidates), ratings from self, peer group, and university-based teacher educators. Interviews (n = 14 candidates)

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 37: Tracz et al. (2017)	Predictive relation between selectivity standards and principal supervisor ratings of teachers through SEPTPP	V: Predictive validity*	Early career teachers in school context, 2+ TEPs across universities, US	Empirical, Quantitative, Secondary	Pre-existing evaluation results (n = 11,723 graduates) by employer principals. Administrative records of graduated teachers, SAT (n = 289) and GPA (n = 3,420)
Study 38: Voss et al. (2011)	Developing and validating an instrument for assessing teachers' general pedagogical and psychological knowledge (PPK)	V: Construct validity*	Candidate assessment, 2+ TEPs across universities (and across federal states), Germany	Empirical, Mixed, Primary and secondary	Questionnaires (n = 20 university-based teacher educators* and schoolteachers). Evaluation instrument (n = 71 schoolteachers, n = 845 candidates for self-ratings, n = 620 school student's rating of 27 candidates). Literature search. **Expert judgers*
Study 39: Tillema (2010)	Utilising formative assessment for teacher professional development.	JM: Instrument implementation and result use	Schoolteacher evaluation, Context-free	Scholarly written, Qualitative, N/A	No data collection
Study 40: Yinger and Daniel (2010)	Standards and accreditation processes in teacher education	JM: Instrument development	Candidate assessment, Context-free	Scholarly written, Qualitative, N/A	No data collection

(Continued)

TABLE 1: (*Continued*)

Publication	Study description	Themes; *sub-themes*	Study context	Research methodology	Data collection tools and source of data
Study 41: Tigelaar and van Tartwijk (2010)	Prospective teacher evaluation methods such as portfolios, self-assessment	JM: Instrument implementation and result use*	Candidate assessment, Context-free	Scholarly written, Qualitative, N/A	No data collection
Study 42: Rafiq et al. (2002)	Examining public and private university lecture's evaluation proformas	JM: Instrument structure*	Higher education lecturer evaluation, No TEP, Pakistan	Empirical, Qualitative, Secondary	Documents of programme: 8 higher education lecturer evaluation proformas
Study 43: Rafiq and Qaisar (2021)	University lecturer's views about their evaluation in a private university	V: Face validity*	Higher education lecturer evaluation, No TEP, Pakistan	Empirical, Quantitative, Primary	Questionnaire (n = 150 higher education lecturers)
Study 44: Khan et al. (2017)	Reviewing teacher evaluation methods, student achievement-based assessment.	JM: Instrument implementation and result use*	Schoolteacher evaluation, No TEP, Pakistan	Scholarly written, Qualitative, N/A	No data collection
Study 45: Rizwan and Masrur (2018)	schoolteachers' instructional planning and strategy knowledge and skills	V: Consequential validity	Schoolteacher evaluation, No TEP, Pakistan	Empirical, Quantitative, Primary	Evaluation instrument (n = 345 schoolteacher self-ratings)

Note. RR: Reliability and validity of judgement EU: Evaluation Use MM: Methodology used to make judgement *Candidate evaluation with an authentic tool were examined

Data Extraction

Data was extracted from each paper into a summary frame developed specifically for this review which included citation, study aim(s), research question(s), research focus, evaluation context, methodology, and findings (see Anderson *et al.*, 2024). Next, data were organised into a chart which is included as Table 1; themes and subthemes were added after analysis.

Analysis

A total of 45 studies were deemed suitable for thematic analysis according to Braun and Clarke's (2006) Six-Step Framework. Analysis involved the following steps: familiarisation with data, creating initial codes, searching for themes, appraising themes, naming themes, and producing the report. This process identified initial recurring themes which were examined iteratively and collaboratively by first and second authors. Themes were cross-checked with each study's focus, which helped to validate and clarify alignment of decisions. The results of each study were summarised under the inductively identified themes and sub-themes (see third column of Table 1). Results were audited by the second author three times and one time by the third author to ensure accuracy. Employing this process ensured a rigorous and thorough account of the findings (King, 2004).

Ethical Considerations

Commensurate with a systematic literature review, ethical considerations relate primarily to transparency of processes for reproducibility and to researcher bias. Creation and adherence to the PRISMA protocols also increased trustworthiness of the research (Snyder, 2019). Making research public and open to critique was another method for establishing credibility (McDonagh, 2016). The review was presented at two conferences; sharing early interpretations and themes allowed analysis to be refined through questioning and critique. During the process, the third author served in the role of a 'critical friend' (Herr and Anderson, 2015) who engaged in debriefing conversations throughout the search, extraction, and analysis processes. The bracketing and reflective practices were used to address potential bias (Creswell, 2007). The use of bracketing involves a thorough, honest, and in-depth personal reflection throughout the research process. Brainstorming, repeated analysis, reintegration of meaning, and audits ensured that a high level of neutrality was achieved.

4. FINDINGS

The review identified 45 peer-reviewed articles that met established inclusion criteria. Table 1 provides a descriptive presentation of findings organised in response to the research questions.

Country and Context of Data Collection

The papers located data collection predominately in the USA (56%, $n = 25$) followed by Pakistan (13%, $n = 6$), Germany ($n = 2$), and Malaysia ($n = 2$). Other countries with one study each included Canada, Chile, Indonesia, Mexico, New Zealand, and South Africa. Three studies did not have a country-specific context. One study (#2) was carried out within the context of two countries involving student teachers in Germany and Austria. No research was identified from the four UK home nations. When the study was about teacher education, this took place within the context of university-based teacher preparation programmes. The majority of research involved examination within a single TEP ($n = 13$) or multiple TEPs across universities in the same country ($n = 10$), followed by studies from multiple programmes within a single university ($n = 6$). Only one study involved multiple TEPs from different states or countries.

Methods and Participants

Quantitative ($n = 20$) and mixed ($n = 14$) research methods were prominent, followed by studies relying solely on qualitative methods ($n = 11$). Of those 11 qualitative studies, only six were empirical studies, while five were scholarly written non-empirical studies. Non-empirical studies were identified as not based on systematic data collection and/or analysis. Instead, they relied on conventional literature and the authors' own experiences and scholarship. Thereby, a majority of studies were empirically driven ($n = 40$, empirically driven quantitative = 20, empirically driven mixed = 14, empirically driven qualitative = 6). A notable portion of studies ($n = 19$) relied solely on primary data collection, while others used only secondary evidence ($n = 11$). Ten studies employed a combination of both primary and secondary methods. Sources of data in empirical studies ($n = 40$) revealed that almost half drew upon data from university-based teacher educators ($n = 17$), followed by student teachers ($n = 13$) and school-based mentor teachers ($n = 10$). Data collected from these participants typically took the form of participant views, experiences, and assessments of student and teacher evaluation tools, or involved the ratings of student teachers and mentor teachers.

Studies which involved primary data employed a variety of data collection tools, including evaluation instruments, surveys, interviews, focus group discussions, think-aloud data, and feedback forms. Two almost equally favoured instruments were evaluation instruments ($n = 17$), encompassing fabricated (i.e., created for research purposes), emerging (i.e., in development stage) and authenticated versions (i.e., actively in use in TEPs) to elicit judgements and ratings on students and teachers as well as questionnaires ($n = 14$). Primary data collection involved obtaining direct verbal views and feedback ($n = 10$) through interviews ($n = 5$), engaging in focus group discussions ($n = 5$), and employing think-aloud data ($n = 1$). Written views and feedback ($n = 7$) were collected

through qualitative comments as a part of questionnaires ($n = 5$), feedback form ($n = 1$), and reflective commentary ($n = 1$). Studies drawing on secondary data used ready evaluation outputs (i.e., ratings assigned to a student teacher), documents (i.e., programme information), administrative data (i.e., student teacher's ethnicity), and existing literature. Evaluation outputs ($n = 12$) encompassed data, results, and reports from evaluations, followed by document reviews ($n = 5$) involving an examination of teacher education programme documents and course work.

Prevalence of Research Focus and Domains

Three main research foci were identified and are presented (see third column of Table 1): validity ($n = 25$), reliability ($n = 13$), and judgement making of teaching effectiveness ($n = 7$). Notably, some studies addressed multiple foci; in such cases, the primary focus was utilised for determining prevalence, and the secondary focus was recorded. Majority of studies focused on construct validity ($n = 8$), followed by face validity ($n = 6$), consequential validity ($n = 6$), predictive validity ($n = 4$) and content validity ($n = 1$). Studies mainly focused on reliability of judgement specifically aimed to identify consensus and consistencies in judgement amongst raters ($n = 9$). However, only few directly focused on internal consistency reliability ($n = 2$) and influences on rater reliability and how to improve it ($n = 2$). Amongst these studies, no study directly addressed the rater's reasonings, but three studies identified rater's reasonings with an intention to explain reliability (see #23, 29, 36). Studies 29 and 36 included interviews with raters for their reasoning in an examination of interrater reliability (IRR) and Study 23 employed a statistical analysis of independent sample t-tests to examine relationships between administrators' accuracy on scoring connected to reasoning strategies. Studies focused on judgement making of teaching effectiveness focused on instrument implementation and result use ($n = 4$), instrument development ($n = 2$), and instrument structure ($n = 1$).

Evaluation Instruments

Eleven instruments used to judge teaching effectiveness were identified (see Table 1 column four). Evaluation instruments developed by TEP faculty ($n = 5$) and the adopted instruments ($n = 5$) were the most prominent followed by instruments modified by TEPs ($n = 1$). Adopted evaluation tools were either a product of an educational research centre (such as edTPA) or of independent researchers (e.g., CLASS, STP). Amongst eleven tools, those grounded in professional standards ($n = 5$) were the most prominent, followed by those based on national or state standards ($n = 4$). By design, teaching evaluation-focused tools (i.e., instructional performance) were the most prominent ($n = 8$), followed by those

focused on teaching dispositions (n = 3). However, a closer look revealed that some instruments, aside from their main focus on teaching effectiveness, also contained elements of disposition (e.g., the competence tool in Study 1, STP in Study 30). During implemantation stage, evaluation results were predominatly used for, or to contribute to, summative decision-making; efforts to support growth with progress oriented formative feedback, tailored support and monitoring were rare. Notably, of these tools, two focused on dispositional growth and one focused on growth in teaching effectiveness. Assessment methods including a combination of self-assessment and observation (n = 4), and portfolio/student teacher work (n = 4) were the most common followed by observations alone (n = 2). One implementation distinctively incorporated peer assessment in conjunction with observation.

Reliability

Reliability signifies consistent or dependable results. Examination revealed a number of findings related to the nature of reliability of judgements of teaching effectiveness focused within four key areas: internal consistency reliability, inter-rater reliability, influences on rater reliability, and proposed ways to improve reliability.

Internal Consistency Reliability

Internal consistency reflects the extent to which items within an evaluation instrument measure the same construct, in this case, teaching effectiveness. Overall, researchers tend to prioritise internal consistency reliability testing more frequently than other reliability tests in the represented studies. Several studies revealed that consistencies and accuracies in assessments tend to be more prevalent in holistic scoring rather than analytic scoring across raters and time (#23, 26, 31, and 35), suggesting scores may be more reliable when utilised in a holistic manner (#31). In Study 28, the degree of consistency amongst items of the T-TESS rubric considered the statistical properties and the extent to which it differentiated student teachers on teaching quality. Further, in Study 31 it was 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 (#11, 23, 41).

Interrater Reliability (IRR)

Several studies involved examination of inter-rater reliability considering the consistency of the judgements of several raters. Some studies confirmed instances of interrater agreement and consistency (#23, 24, 25, 31, 34, 35, and

36). A majority of studies employed descriptive statistics using either exact or partial percentage agreement or comparing raters' scores with an identified 'true' score (#23, 31, 34, and 36). Studies which calculated IRR with advanced statistics techniques (i.e., Cohen Kappa) were less prevalent (#26 and 29). One study (#25) calculated a Chi-square test to examine the match between raters. One study considered 'similarity in mean scores' as evidence for IRR (#35), and another used qualitative interview data as evidence estimating IRR (#9). Measures such as Cohen's kappa and intra-class correlations were recommended for accurate reporting (#29). Several studies revealed that where there are consistencies in assessments, these might stem from raters' preferences and mindset in the form of clustered ratings around a limited range of available scoring options (#9, 35, 36).

Other studies revealed inconsistencies and disagreements between raters (#9, 23, 25, 26, 29, 31, and 35). Analysis identified two notable patterns emerging. Student teachers tended to rate themselves lower than peers and mentors (#36), yet their self-ratings, in comparison to university-based teacher educators, were either similar (#25 and 35) or lower (#36). The second pattern noted that mentor teacher ratings were almost always higher than both student teacher (#35) and university-based teacher educators (#9 and 35). The review also noted inconsistencies between school-based mentors and university-based teacher educators (#9 and 26). In Study 35, statistically significant similarities in overall mean scores between student teachers and supervising faculty regarding professional dispositions were found. The study also noted statistically significant higher rating from mentor teachers over university-based teacher educators not only at one point in time but also across time. This was deemed an important finding as mentors were noted as 'professional teachers in the field observing the actual teaching practices and dispositions of teacher candidates' (#35, p. 128).

A student teacher's active involvement in self-evaluating effectiveness created additional opportunities for growth, in particular when self-reported ratings were deliberated alongside mentors and university teacher educators demonstrating triangulation (#15 and 35). This fostered student teachers' autonomy, self-reflection, and self-monitor practices (#15 and 39). Study 13 further evidenced engagement with evaluation feedback as supportive. In Study 28, the use of triangulation was evident through focus group discussions to confirm questionnaire results.

Influences on Rater Reliability

Analysis revealed a widely shared objective in evaluating teaching to ensure that judgements are considered accurate and reliable. Studies exemplified inconsistencies and inaccuracies in judgement due to factors stemming from ratee, rater, tool characteristics, deployment of evaluation, and the choice of methods used to determine reliability and validity. Interestingly, a masking effect of quantitative data over qualitative data was evidenced in Study 2 which appeared

to suggest that utilisation of qualitative data could unearth biases in rater's judgement in some cases in contrast to quantitative data thus arguing for the use of multiple sources of evidence. Analysis also revealed that poor IRR might be related to measurement approaches. For example, it could stem from a restricted range in rating scale (#35) such as having only three options leaving little room for variability and actually leading to consistently high ratings (#29, 35). IRR was found to also potentially be related to having a binary mindset (i.e., satisfactory, unsatisfactory) even if the scales include a greater range to potentially avoid conflict with other ratees (#16 and 21). It was also found that some judgements were made using non-scoring criteria aspects such as experience from conducting other evaluations and own teaching experience (#23).

In Study 2, a student teacher's gender was found to be unrelated to evaluation of the teacher's skills, causing no bias. Societal groups the ratee belonged to (e.g., ethnicity) influenced judgement in two studies (#23 and 36). And while ethnicity significantly affected outcomes like edTPA pass rates (#12), it was not a significant factor in each context, for example in Study 21 which involved principle ratings. Nonetheless, there was a call to address bias against student teachers from minority backgrounds (#12 and 15). Also, raters themselves – who they are, their cognitive processes, social background, personal beliefs, preferences, or prejudices – are argued to influence judgement (#11 and 23). Study 23 showed that deviating from formally designated tasks led to subjective personal conclusions that may not accurately reflect the teacher's actual effectiveness.

Improving Reliability

The most widely taken action to improve reliability has been the standardisation of sources, scoring, and criteria (#33 and 41) with the fundamental idea to exclude contextual influences (#33). However, standardisation did not guarantee evaluators made objective and reliable judgements (#4) and standardised assessment tools were found to not always align within the context of specific subject areas (e.g., art teacher #33). Constructing measurable indicators for assessment was reported as important to mitigate potential subjectivity of judgements (#25), as some indicators were challenging to operationalise.

Training was one of the most frequent conclusions to achieve greater reliability and validity of judgements (#1, 26, 27 and 31) for all raters (#1, 4, 26 and 31) and explicitly for mentors (#9). Suggestions of Study 31 recommended more than one-time training was needed and also advised a quality control scoring session. Further studies specifically focused on the impact of training with pre- and post-training tests and concluded poor interrater agreement (#29) or little to no improvement in interrater agreement (#34). Findings from Study 34 concluded IRR improved post-training for some evaluations (i.e., research paper, case study) but decreased for others (i.e., digital portfolio).

Some empirical studies concluded training was not an effective solution (#23, 29, and 31). In Study 23, school administrators exhibited a variety of reasoning strategies to rationalise judgements. Interestingly, the findings indicated variation in reasoning strategies did not influence 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 with student teacher self-reporting has been suggested (#35 and 41) and could validate self-evaluations (#41). Other recommendations to improve reliability included multiple raters rather than a single rater (#24 and 35), employing a variety of assessment methods (#24), and assessing multiple times (#16 and 35). Portfolios of student teachers' work have also been suggested (# 20) and are widely adopted in many TEPs (#20 and 41), especially in the USA. However, the intention of being student-centred has been found to be flawed by use for organisational needs such as quality assurance (#20).

Validity

Several studies examined the ways in which judgements of teaching reflect real situations, are adequate to measure what they intend to measure, and if instruments fully represent what they aim to measure. In terms of instrument validity, face, content and construct validity emerged as the most frequently employed, while consequential validity was also evident. In Study 17, Classical Test Theory (CTT) was used, with authors arguing CTT as one of the most used tests in the field; the study also used the Rasch model to explore construct validity. In Study 7, an extensive validation process to develop a teaching effectiveness scale in higher education was carried out; efforts involved interviews, focus groups, subject matter experts, calculation of Content Validity Index (CVI), content validity ratio, and confirmatory factor analysis to ensure construct validity. Further studies also examined content validity through both Cohen's Kappa Index and CVI, supported by face validity (#14). Further, Study 38 involved in-service teacher relevance ratings stated as content-based evidence for instrument validity. Factor analysis was also evident in establishing construct validity (#1, 3, 7, 11, 18, and 26).

Content of evaluation tools was underpinned by a combination of standards and evidence. The standards encompassed state and/or national standards ($n = 3$), professional standards from associations ($n = 4$), and TEP institutional standards or frameworks ($n = 1$). Further evidence used to establish validity came from internally generated original evidence such as need analysis ($n = 4$), prior academic research ($n = 3$; i.e., effective teacher qualities), a framework for teaching, and recommendations of professional organisations. In Study 4, researchers suggested validity could be compromised through a shifting in purpose from evaluating constructs of effective teaching for student teacher growth to

gatekeeping purposes. This leads to a shift where the evaluation itself becomes the focus of instruction (#33) causing a 'teaching to the test' approach (#4 and 33) that could weaken validity.

Findings revealed insufficient evidence pertaining to success in student teacher evaluations to predict subsequent teaching success (i.e., predictive validity). Certain measures and indicators, both prior to individuals entering teacher education and during preparation, were found to be valuable in predicting the future effectiveness of teachers and others were not (#6 and 37). Some evaluation scores in certain subjects (i.e., reading edTPA) were shown to prevent ineffective teachers from entering the workforce (#12 and 24); others failed to predict teaching success in subjects such as mathematics (#12) and arts (#32). Research regarding edTPA, the most frequently used student teacher assessment tool identified in the review has yet to establish predictive validity. Using evaluation as a one-time gatekeeper was found to possibly screen out student teachers who would become effective teachers (#4).

A number of studies addressed concepts of face validity, examining if evaluations appear to measure what they are supposed to measure. Several studies included in this review revealed a notable sense of dissatisfaction with evaluations (# 4, 8, 10, and 16) indicating low confidence of student teachers and teacher educators to actively engage in evaluation processes. This was predominantly attributed to perceptions of evaluation tool's lack of validity and reliability (#4, 8, 10 17, and 41). Low engagement with evaluation measures was also attributed to cultural insensitivity evident in the tools (#4 and 33), high-stake consequences linked to results (#4 and 12), and unclear and impractical evaluation tools (#8). In one study, this also led to school administrators identifying their own criteria, moving away from agreed standards-based indicators of the evaluation tools to rely on their own (#23). Study 33 suggested a more radical approach, to grant autonomy to teacher educators and mentor teachers to choose contextually appropriate evaluation tools. Interestingly, a drive towards standardisation was found to potentially come at a cost, such as not considering programme values (#4), moving away from authentic, culturally responsive evaluation, or disregarding the real-time context of teacher–student relationships (#33).

5. Discussion

The intricacies inherent in evaluating teaching effectiveness, alongside ongoing discourse regarding criteria for judging competence, were illuminated by the outcomes of this review. Exploration of the methods of evaluating teaching effectiveness has underscored several crucial factors, such as the challenge of ensuring the reliability of judgements within dynamic educational environments, understanding and applications of reliability and dependability and

consequential factors. The findings of this systematic review hold implications that merit consideration by Teacher Education Programs (TEPs).

Features of Reliability and Validity in Judging Teaching Effectiveness

Examination of research domains in prior research showed a focus on reliability as the main research foci, followed by use of evaluations and interrogation per validity measure to create confidence in the judgement process. Investigations of content, construct and face validity were evident, yet very little regarding internal consistency. Most research was empirically driven and conducted in an individual, university-based TEP with little comparative analysis or understanding investigated across systems evident. The types of evidence used to make judgements of teaching effectiveness involved results from a variety of sources, with the primary source being observations conducted by university-based teacher educators followed by information from students themselves. Input from school-based mentor teachers was less common.

The individuals who conduct judgements included university-based supervisors, mentor teachers, a combination of these two, self-evaluation, and peer evaluation. Student teacher contributions were prominently visible in the context of formative evaluation; self-evaluations did not extend to the decision-making level. Discrepancy between how student teachers rate themselves and how they are assessed by peers, mentors, and university-based teacher educators may be less relevant or directly relevant to the overall consistency of judgements, depending on context.

Analysis of instruments further underscores the extensive diversity of indicators utilised to judge effective teaching. Teaching standards were often identified as the identified cues for judgement and to establish both content and construct validity. Standards provided the basis and rationale for judgements that are made, and these judgements are made by professionals who know what the standards are and contextualise application. It is noted, however, that during any lesson observation, some standards may not be observable or met bringing question to construct validity. Tools were designed mostly in the form of scale or rubric and were almost equally employed in a formative or summative manner. Most evaluation tools were used for diagnosing and measuring growth followed by informing decisions regarding eligibility and licensure to screen out those who may not be ready to teach. Engagement of those who make judgements of teaching effectiveness with evaluation has been influenced by the confidence in reliability and validity of judgements and tools, and not necessarily each measure or the indicators embedded in tools able to predict a student teacher's future teaching.

Considering Dependability

It is clear through the variety of ways researchers engage questions of validity and reliability that there is a constant interplay amongst how these are approached and how they are perceived by those engaged in the processes of evaluating teaching effectiveness. For example, an evaluation tool with high content and construct validity confirmed with advanced statistical modelling may not be used with fidelity and thus not yield reliable results. Additionally, IRR can be achieved regardless of accuracy of a measure; 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 of what is being measured. However, in teaching, the setting (Cooksey, 1996) (i.e., classroom environment, learner complexity, different schools, etc.) and subjects always change. It is within these concepts that social judgement theory (SJT) emerges to guide consideration of the diverse settings in which a demonstration of competency is judged. Neither ecological validity (i.e., the connections between judgement criteria and the cues used to make judgements) nor cue utilisation validity (i.e., the connection between the cues that are observed and the judges making decisions about student teachers) were really evident in the included studies except for Study #23.

Variability evidenced in these studies further reflects the complexity and uncertainty of the teaching endeavour and questions the desirability of standardisation and high-level objectivity. Perhaps, 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 professional judgement and personal insights of raters or individuality of each student teacher. Perhaps moving from the cannon of quantitative language to that of trustworthiness and dependability is more fitting to judge a phenomenon that defies uniformity. Within studies included in this review, it was put forward that multiple raters could help mitigate variance in understanding and implementation of evaluations (#24), and multiple ratings could yet 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 amongst the research examined. Findings overall indicate a needed alignment of evidence used to make to make judgements of teaching effectiveness and two critical aspects (Haigh *et al*., 2013) of why these occur in the first place: to confirm that student

teachers have the necessary personal qualities and relationships needed to assume independent responsibility of a classroom and that they can plan, teach, and assess for pupil learning.

Challenges in Complexity

As Cooksey (1996) noted and this systematic review re-confirmed, judgement-making appears to remain a best estimate of the right choice under specific constraints which always runs the risk of being in 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., a student teacher receiving licensure or not). Even the simplest teacher decisions can have multiple causal pathways (Opfer and Pedder, 2011). The degree of ambiguity and variation with which decision-makers are able to cope amongst an intertwined set of probabilistic relationships indeed varies from one setting, TEP, or education system to the next. What is considered important to investigate and establish validity and reliability remains just as variable according to the literature. An interesting deliberation emerges to reconsider predictive validity and to persist 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 judged, 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 reengage 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, socialisation, 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. There was a tension between high-stake consequential outcomes of judgements and educative uses of evaluation for growth which revealed itself 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 operationalised indicators, or if these can be relatively broad, such as the holistic ability to gracefully teach increasingly diverse learners. Socialisation 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 standardised ways of doing is what is desired, or if it

is more necessary for new teachers to be transformative and 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 sometimes 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.

For Further Exploration

This review brought forward multiple considerations for future research and substantially informed the larger aforementioned project. It would be of particular interest for future research to be situated within the UK given no studies in this review were found to be in this context. It may also be of interest to consider other facets of reliability and validity, such as *intra*-rater reliability or ecological validity. Only one study focused on raters' justifications of judgements of student teachers' readiness to teach (see #23); further exploration of summative decision-making and how reasoned opinions and preferences factor into professional judgements is needed. All studies in this review involved teacher preparation situated within university contexts. Given the increasing diversity of alternative teacher education programmes (e.g., Teach First, Teaching School Hubs, School Centred Initial Teacher Training), it would be interesting to explore facets of judging teaching effectiveness from this perspective. Additionally, the function of judgements, and the potential positive and negative outcomes of them, could be further explored in the space of consequential validity. This approach could enable a deeper understanding of whether the measurements conducted in TEPs effectively align with the requirements and conceptualisations of effective teaching.

6. Limitations

Although efforts have been made to address the constraints of the systematic review, limitations of this 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 studies presented. Results obtained through a review of literature are only as reliable as the methods adopted in the original primary research. Consequently, any inherent issues in research design remain and may have influenced results even given the quality of the original research as an inclusion criterion. 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 relevant work may have been found relevant if framed in a different way, therefore exploring research beyond the inclusion parameters may have identified further sources.

7. Conclusions

This review underscores ongoing discourse surrounding the validity and reliability of assessment measures and criteria employed in evaluating student teachers, with implications extending beyond mere predictive capacities. Particularly pertinent is the debate surrounding the high-stakes nature of judgement-based evaluations used in pivotal decisions of entry into teacher preparation and the teaching profession itself. Given the multifaceted nature of both purpose and process, there persists a compelling need for deliberation in selecting indicators of teaching effectiveness. In this regard, it is imperative that indicators reflect standards of the profession and conceptually contextualised while upholding a holistic perspective. 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.

8. Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Society for Educational Studies 2022 National Award.

9. Supplementary Material

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00071005.2024.2374070.

10. Orcid

Sarah K. Anderson http://orcid.org/0000-0001-9084-7413
Pinky Jain http://orcid.org/0000-0001-7233-3187

11. 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.

Anderson, S. K., Jain, P., and Ozsezer-Kurnuc, S. (2023, September 7–8) *Professional Judgement and Standard Frameworks: Exploring Duplexity in Assessment of*

Teachers' Practices. Society for Education Studies 2023 Colloquium (Oxford, UK). Available at: https://eprints.gla.ac.uk/306207/ (accessed 20 March 2024).

Anderson, S. K., Ozsezer-Kurnuc, S., and Jain, P. (2024) *Judging Student Teacher Effectiveness: A Systematic Literature Review* (Society for Educational Studies and University of Glasgow). Appendices A-B. Available at: https://eprints.gla.ac.uk/308586/ (accessed 28 March 2024).

Asher, L. (2018) How Ed Schools Became a Menace to Higher Education (The Chronicle of Higher Education). Available at: https://www.chronicle.com/article/how-ed-schools-became-a-menace/.

Biesta, G. (2015) What is education for? On good education, teacher judgement, and educational professionalism, *European Journal of Education*, 50 (1), 75–87. doi: 10.1111/ejed.12109.

Biesta, G. (2020) *Educational Research: An Unorthodox Introduction* (London, Bloomsbury).

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology, *Qualitative Research in Psychology*, 3 (2), 77–101. doi: 10.1191/1478088706qp063oa.

Bryman, A. (2016) *Social Research Methods* (5th edn) (Oxford, UK, Oxford University Press).

Cochran-Smith, M., Ell, F., Ludlow, L., Grudnoff, L., and Aitken, G. (2014) The challenge and promise of complexity theory for teacher education research, *Teachers College Record*, 116 (4), 1–38. doi: 10.1177/016146811411600407.

Conroy, J., Hulme, M., and Menter, I. (2013) Developing a clinical model for teacher education, *Journal of Education for Teaching*, 39 (5), 557–573. doi: 10.1080/02607476.2013.836339.

Cooksey, R. W. (1996) The methodology of social judgment theory, *Thinking and Reasoning*, 2 (2), 141–174. doi: 10.1080/135467896394483.

Creswell, J. W. (2007) *Qualitative Inquiry & Research Design: Choosing Among Five Traditions* (2nd edn) (Thousand Oaks, CA, Sage).

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. doi: 10.1080/02619768.2017.1315399.

Haigh, M. and Ell, F. (2014) Consensus and dissensus in mentor teachers' judgments of readiness to teach, *Teaching and Teacher Education*, 40, 10–21. doi: 10.1016/j.tate.2014.01.001.

Haigh, M., Ell, F., and Mackisack, V. (2013) Judging teacher candidates' readiness to teach, *Teaching and Teacher Education*, 34, 1–11. doi: 10.1016/j.tate.2013.03.002.

Hammond, K., Rohrbaugh, J., Mumpower, J., and Adelman, L. (1977) Social judgment theory: applications in policy formation. In M. Kaplan and S. Schwartz (Eds) *Human Judgment and Decision Processes in Applied Settings* (New York, NY, Academic Press), 1–29. doi: 10.1016/B978-0-12-397240-8.50008-2.

Hattie, J. (2023) *Visible Learning: The Sequel – A Synthesis of Over 2,100 Meta-Analyses Relating to Achievement* (New York, Routledge).

Herr, K. and Anderson, G. L. (2015) *The Action Research Dissertation: A Guide for Students and Faculty* (2nd edn) (Thousand Oaks, CA, Sage).

King, N. (2004) Using templates in the thematic analysis of text. In C. Cassell and G. Symon (Eds) *Essential Guide to Qualitative Methods in Organizational Research* (Thousand Oaks, CA, Sage), 256–271. doi: 10.4135/9781446280119.n21.

Klassen, R. M. and Kim, L. E. (2019) Selecting teachers and prospective teachers: a meta-analysis, *Educational Research Review*, 26, 32–51. doi: 10.1016/j.edurev.2018.12.003.

Levine, A. (2006) *Educating School Teachers* [Electronic version]. Available at: https://files.eric.ed.gov/fulltext/ED504144.pdf (accessed 27 March 2024).

Martin, S. D., McQuitty, V., and Morgan, D. N. (2019) Complexity theory and teacher education, *Oxford Research Encyclopedias*. doi: 10.1093/acrefore/9780190264093.013.479.

McDonagh, C. (2016) Ethics, rigour and validity. In B. Sullivan, M. Glenn, M. Roche, and C. McDonagh (Eds) *Introduction to Critical Reflection and Action for Teacher Researchers* (Oxon, Routledge), 93–110. doi: 10.4324/9781315693033.

Mutton, T. and 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.

Opfer, V. D. and Pedder, D. (2011) Conceptualizing teacher professional learning, *Review of Educational Research*, 81 (3), 376–407. doi: 10.3102/0034654311413609.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016) Rayyan – a web and mobile app for systematic reviews, *Systematic Reviews*, 5 (210). doi: 10.1186/s13643-016-0384-4.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., and McKenzie, J. E. (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, *BMJ*, 372 (160), n160. doi: 10.1136/bmj.n160.

Raths, J. and Lyman, F. (2003) Summative evaluation of student teacher: an enduring problem, *Journal of Teacher Education*, 54 (2), 201–216. doi: 10.1177/0022487103054003003.

Sandholtz, J. H. and 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. doi: 10.1177/0022487111421175.

Schmoker, M. (2023) *Results Now 2.0: The Untapped Opportunities for Swift, Dramatic Gains in Achievement* (Arlington, VA, ASCD).

Seidenberg, M. (2017) *Language at the Speed of Sight: How We Read, Why so Many Can't, t, and What Can Be Done About It* (New York, BasicBooks).

Snyder, H. (2019) Literature review as a research methodology: an overview and guidelines, *Journal of Business Research*, 104, 333–339. doi: 10.1016/j.jbusres.2019.07.039.

Van Der Knaap, L. M., Leeuw, F., Bogaerts, S., and Niissen, L. (2008) Combining Campbell standard and the realist evaluation approach: the best of two worlds?, *American Journal of Evaluation*, 29 (1), 48–57. doi: 10.1177/1098214007313024.

Correspondence
Sarah K. Anderson
School of Education, University of Glasgow, 11 Eldon St., Glasgow G3 6NH, UK
Email: sarah.anderson.3@glasgow.ac.uk