Transdisciplinary perspectives on validity: bridging the gap between design and implementation for technology-enhanced learning systems

Haastrecht, M.A.N. van

7

VALIDITY CRITERIA FOR TECHNOLOGY-ENHANCED LEARNING

Technological aids are ubiquitous in today's educational environments. Whereas much of the dust has settled in the debate on how to validate traditional educational solutions, in the area of technology-enhanced learning (TEL) many questions still remain. Technologies often abstract away student behaviour by condensing actions into numbers, meaning teachers have to assess student data rather than observing students directly. With the rapid adoption of artificial intelligence in education, it is timely to obtain a clear image of the landscape of validity criteria relevant to TEL. In this paper, we conduct a systematic review of research on TEL interventions, where we combine active learning for title and abstract screening with a backward snowballing phase. We extract information on the validity criteria used to evaluate TEL solutions, along with the methods employed to measure these criteria. By combining data on the research methods (qualitative versus quantitative) and knowledge source (theory versus practice) used to inform validity criteria, we ground our results epistemologically. We find that validity criteria tend to be assessed more positively when quantitative methods are used and that validation framework usage is both rare and fragmented. Yet, we also find that the prevalence of different validity criteria and the research methods used to assess them are relatively stable over time, implying that a strong foundation exists to design holistic validation frameworks with the potential to become commonplace in TEL research.

The contents of this chapter are based on: van Haastrecht, Haas, et al. (2024). Understanding Validity Criteria in Technology-Enhanced Learning: A Systematic Literature Review. Computers & Education.

7.1 INTRODUCTION

Validation in its most general sense involves evaluating evidence regarding specific claims, to assess the plausibility of these claims (Kane, 1992). Validity is a multi-faceted concept, with different validity criteria being more or less relevant in different contexts. When Cronbach and Meehl (1955) and Messick (1989) were emphasising the need for well-defined validity criteria in the second half of the twentieth century, the immense influence technology would come to have on our daily lives had yet to materialise. Concepts such as construct validity, criterion validity, and content validity seemed to cover the most vital aspects of validity in educational measurement and assessment (Kane, 1992). Then came the introduction of the varied assortment of technologies that exist today to enhance our educational environments. Students can now collaboratively improve their problem solving skills using online platforms (Stadler et al., 2020) and we can use learning analytics to understand the behaviour of students on the other side of the world in Massive Open Online Courses (MOOCs) (Douglas et al., 2020). This raised the question: can we continue to use traditional validity criteria in these decidedly non-traditional contexts?

The short answer to this question is no: we cannot simply apply old validity criteria to a new age. We need to recognise that validity argumentation must adapt when we switch from traditional classroom settings to complex, interactive environments at scale (Mislevy, 2016). External validity, commonly referred to as generalisability, is a typical criterion that we should be mindful of in technology-enhanced learning (TEL) settings. Technologies tend to abstract away the context of the learner and present educators with student data that can at best provide a summary of the actual learner context. Generalisation arguments rely on some form of comparability between one context and the next, and when students use their own devices and software, the comparability claim is challenged (Wools, Molenaar, et al., 2019). With respect to a validity criterion such as authenticity, we can use virtual reality and simulations to create more authentic educational experiences (Wools, Molenaar, et al., 2019), but in cases where technology abstracts away student behaviour, it can become difficult to assess how authentic these educational experiences really are (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023).

In a special issue examining possible links between learning analytics and assessment, the editors stated that "the field still needs a clear theoretical framework to guide the consideration of validity" (Gašević, Greiff, et al., 2022, p. 4). Recent work looking to deepen the connection between these two fields highlights that future research can improve the validity of learning construct interpretations by combining insights from different data sources (Raković et al., 2023). Likewise, several systematic reviews have stressed the need for a coherent, comprehensive framework to aid with the evaluation of TEL environments (Clunie et al., 2018; Erdt et al., 2015; Heil and Ifenthaler, 2023).

We need to clarify the current landscape of validity criteria in TEL if we are to facilitate rigorous validation in the future. Without consensus on which validity criteria could possibly be examined, we cannot expect researchers to make the right considerations.

Researchers have started to theorise what validation should look like in the technology-enhanced world of today. A common factor among novel views regarding validation is the necessity to recognise diverse perspectives (Addey et al., 2020) and diverse epistemologies (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). Recognising diverse perspectives does not exclude the possibility of reaching a common theoretical foundation. We can recognise that evidence for certain validity criteria (e.g., authenticity) may be sourced from qualitative methods applied in practice whereas evidence for other criteria (e.g., statistical validity) may result from quantitative methods based on theory, while still identifying epistemological patterns in validity criteria relations that can serve as a basis for holistic validation frameworks.

By combining insights on the validity criteria considered in TEL research, how they are defined and measured, how their prevalence has evolved over time, and how criteria relate in an epistemological sense, we can take a first step in addressing the current gaps in the literature relating to TEL validation. We aim to gain these insights in this paper by conducting a systematic review of TEL literature, to uncover how researchers have dealt with the challenging nature of TEL validation. There is, to our knowledge, no systematic review that investigates the validity arguments TEL researchers rely on to defend their conclusions. By collecting details on how validity criteria were defined and measured, we will answer the following research questions:

- **RQ**: How can we characterise the landscape of validity criteria used in TEL research?
  - **RQa**: Which validity criteria are considered in TEL research, how are they defined, and how are they measured?
  - **RQb**: How has the prevalence of different validity criteria in TEL research evolved over time?
  - **RQc**: What epistemological patterns do we observe in the connections between validity criteria in TEL research?

In what follows, we will first discuss earlier work on TEL validation and reviews of TEL literature (Section 7.2). We will then outline our systematic review methodology in Section 7.3 and present our results in Section 7.4. We discuss and interpret our results in Section 7.5, also covering some of the limitations of our methodology. Section 7.6 concludes and outlines several interesting areas for future research.

7.2 BACKGROUND

In this section, we outline the challenges posed to existing validity arguments when technology enters the picture, the solutions that have been proposed, and current open problems. Additionally, we discuss previous systematic reviews of TEL research, demonstrating how our work contributes to the existing literature.

7.2.1 *Validation of technology-enhanced learning*

Bennett and Bejar (1998) recognised over 25 years ago that the introduction of technology into our learning environments necessitated a different approach to validation. They argued that validation cannot be complete unless the underlying rationales supporting design decisions are adequately explained. Kane (1992, 2013b), whose argument-based approach to validation (Kane, 1992, 2013b) is considered to be dominant in educational assessments (Addey et al., 2020), recognised that the introduction of technology implied that different elements in the validity argument now required emphasis (Clauser et al., 2002). In the argument-based approach to validation, an inference chain is constructed pertaining to the design in question, whereby evidence is collected to inform arguments supporting the validity of each step in the inference chain. If we do not deal with novel threats to validity, such as generalisability issues caused by students using personal devices and software (Wools, Molenaar, et al., 2019), we risk weakening the links of the inference chain and undermining the trustworthiness of TEL systems (Aloisi, 2023). The challenges posed by the introduction of novel technologies have led to the conclusion that adapted validation frameworks are required to deal with our adapted world (Mislevy, 2016).

Several adapted validation frameworks have been proposed in recent years that build on the argument-based approach. Zhai et al. (2021) introduced a validity inferential network to better incorporate the impact of machine learning on today's educational assessments. Huggins-Manley et al. (2022) similarly focus on how assessments enhanced with artificial intelligence should be validated, taking a specific interest in fairness. In van Haastrecht, M. J. S. Brinkhuis, Wools, et al. (2023), a validation framework for e-assessment solutions is proposed that combines traditional insights on validity from the educational domain with information systems validity theory. However, these frameworks are rarely employed within general TEL research outside of educational measurement. This is evidenced by a recent systematic review where the authors stated that, to the best of their knowledge, no such frameworks existed (F. L. da Silva et al., 2023).

Where validation generally covers the full research cycle, including research methodology and design methods, evaluation tends to focus on the artefact produced by research and how it is used. Evaluation is more common in

TEL research than validation, but evaluation strategies are generally not comprehensive in nature. A review of TEL literature found that the majority of studies cover one or two educational aspects when evaluating the use of TEL solutions, leading the authors to question "whether educators are evaluating the use of technology in education from a holistic perspective" (Lai and Bower, 2019, p. 38). These findings were largely confirmed in a follow-up study where Lai, Bower, et al. (2022) asked educational technology experts which dimensions should be considered when evaluating TEL solutions. Although the experts could agree to a large extent that learning outcomes and technological aspects should be considered during evaluation, aspects such as design and behaviour were only considered relevant by a minority. The study concludes that theories used in TEL evaluation studies "do not comprehensively account for all dimensions of educational technology use" (Lai, Bower, et al., 2022, p. 752). The review authors describe how they validated their questionnaire, but do not discuss the relationship between the validation and evaluation of TEL research, pointing to the disconnect between current TEL studies and the body of knowledge on validity theory from educational measurement.

That is not to say that TEL researchers are unaware of approaches such as argument-based validation. In fact, several recent works in the area of learning analytics have stressed the potential of argument-based validation to yield more holistic evaluations (Gašević, Greiff, et al., 2022; van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). We have seen applications of the argument-based approach to validation in studies concerning MOOC assessment (Douglas et al., 2020), asynchronous writing tasks (T. Chen, 2022), and eye-tracking solutions for the assessment of data literacy (F. Chen et al., 2023). However, these studies use the traditional argument-based approach, rather than frameworks adapted to suit technology-enhanced environments. Combined with the general lack of comprehensive evaluations, it is evident that TEL validation is still in its infancy (Rodríguez-Triana et al., 2017).

Employing adapted validation frameworks is required to move TEL validation from infancy to maturity, but using such frameworks passively is not sufficient. We must also continuously develop these frameworks to align with the novel learning process data that becomes available due to technological advancement (Fan, van der Graaf, et al., 2022; Raković et al., 2023). Goldhammer et al. (2021) argue that a complete validity argument requires thought about process indicators right from the start of the design phase. Yet, Zumbo et al. (2023) find that there are currently no holistic validation frameworks that adequately deal with TEL process data. Furthermore, only adapting to technological advancement is insufficient. We need to actively ensure our approaches to validation appreciate the human element in the face of increasing technological influence. Future validation approaches should leave room for legitimately diverse arguments (Addey et al., 2020) that consider qualitative criteria such as fairness and trustworthiness (van Haastrecht, M. Brinkhuis,

Peichl, et al., 2023). Only then can we truly claim that TEL validation has matured.

7.2.2 *Systematic reviews of technology-enhanced learning*

Before moving forward with our review, we should ask: How have earlier systematic reviews addressed the topics of validation and evaluation? Verbert et al. (2012) review recommender systems for TEL and retrieve information on whether studies evaluated learning efficiency/effectiveness, accuracy, usefulness, and usability. They find that some studies perform no evaluation at all and that the majority of studies only consider one or two criteria during evaluation. These findings lead Verbert et al. (2012) to conclude that more comprehensive evaluation studies are needed with a more structured approach. Yet, they do not detail what such an approach may look like and which further criteria might be needed to arrive at a comprehensive evaluation. Erdt et al. (2015) similarly review recommender systems for TEL and focus explicitly on evaluation. Of the 235 studies they include, 95 performed no evaluation. The authors suggest that we need to consider evaluation from the earliest design stage and that we should use evaluation frameworks to standardise the evaluation process. However, like Verbert et al. (2012), the authors focus on evaluation, not validation. Evaluation is mostly geared at answering questions about designed artefacts and how they are used, whereas validation also critically examines the research and design methods that produced an artefact. The fixation of evaluation approaches on outcome over process naturally produces more insights regarding implementation than early design stages.

Later TEL reviews maintain this focus on outcomes. Boyle et al. (2016) review the impacts and outcomes of computer games and serious games. Rodríguez-Triana et al. (2017) review blended TEL environments, finding that usefulness and usability are the most commonly incorporated constructs in evaluations, and concluding that their findings are illustrative of a relatively young field. Clunie et al. (2018) ask whether studies investigating the efficacy of TEL resources are comprehensive, in the sense that studies go beyond measuring learner impact to also consider institutional impact. The authors find that no study considered the institutional perspective, and mention the need for "robust evaluation strategies that can provide answers to the why, how, and when questions" (Clunie et al., 2018, p. 315). Lai and Bower (2019) confirm these findings, showing that just 1.4% of studies consider the institutional environment during TEL evaluation. As with the other reviews we have discussed, neither of these reviews mention the possibility that a focus on validation rather than evaluation could be the solution.

In a 2020 tertiary review of 73 systematic reviews, Lai and Bower (2020) provide further evidence of the lack of consensus in TEL evaluation. Using the eight dimensions of evaluation from their earlier work (Lai and Bower, 2019) - including learning, behaviour, design, and the institutional environment -

they find that no systematic review covered more than five dimensions. The authors assert that there is room for a systematic review taking a broader perspective of evaluation "to more comprehensively understand the effects of using technology in education" (Lai and Bower, 2020, p. 253). In other words, at the time of the tertiary review, there was a need for the type of systematic review we are performing within this work.

Since 2020, we have seen various systematic reviews in the area of TEL, but none have tackled the broader perspective that Lai and Bower (2020) call for. Some of these reviews focus on a specific criterion such as generalisability (Abdulrahaman et al., 2020) or usability (Law and Heintz, 2021), thereby not offering a comprehensive overview. Others consider multiple criteria, but do not employ a specific framework or set of evaluation dimensions (Bond et al., 2020; F. L. da Silva et al., 2023; Heil and Ifenthaler, 2023). Further reviews take a different perspective entirely, and consider what drives the adoption of learning technologies (Q. Liu et al., 2020), evaluate how effective workshops are that prepare teachers for TEL (Ahadi et al., 2021), or analyse the survey instruments used to evaluate the integration of new technologies (Consoli et al., 2023).

We can conclude that the gap in the literature identified by Lai and Bower (2020) has not yet been addressed. We still require a systematic review that provides a comprehensive overview of the landscape of evaluation and validity criteria in TEL research. Furthermore, we have seen from the previous sections that appreciation of validation over evaluation has, implicitly if not explicitly, increased in recent years. As Clunie et al. (2018) emphasised, we need more robust strategies that address the why, how, and when questions. There is a definite, pressing need for clarity in TEL validation.

7.3 METHODOLOGY

Prior to conducting our systematic review, we formulated a protocol conforming to the PRISMA-P checklist (Moher et al., 2015) and NIRO-SR guidelines (Topor et al., 2020). The protocol prescribed the steps of our systematic review and helped to ensure that our process was in accordance with the PRISMA guideline for reporting systematic reviews (Page et al., 2021). In this section, we will describe the core elements of our review methodology and deviations from the protocol. A more detailed description of our methodology can be found in the protocol, which we have made available in an open-source project along with our data. [1] All actions have been recorded with time stamps in our open-source project, for complete transparency. To our knowledge, this is the first time in the field of TEL that a systematic review protocol was made available open access prior to publication of the review.

1 https://osf.io/g2s56/

Table 7.1: Search terms and synonyms used in our database search. Terms must be included in the abstracts or titles of studies.

	SEARCH TERMS
	"technology-enhanced learning" OR "technology enhanced learning" OR "e-learning" OR "mobile learning" OR "digital learning" OR "electronic learning" OR "distance-learning" OR "web-based learning" OR "computer-based learning" OR "virtual learning"
AND	"validity" OR "validation" OR "quality" OR "evaluation"
AND	"criteria" OR "criterion" OR "dimension" OR "type" OR "aspect"

7.3.1 *Search strategy*

For our search we used the following databases: ACM Digital Library, IEEE Xplore, PubMed, and Web of Science. We included peer-reviewed journal and conference articles written in the English language. We consciously chose not to exclude any studies based on their publication date, since we aimed to analyse the use of validity criteria over time. The search terms used are listed in Table 7.1.

The search process produced 1,566 results, of which 1,256 remained after deduplication and removal of results that were not peer-reviewed journal or conference articles. The 1,256 publications served as input for our title and abstract screening phase, where we included studies that satisfied the following criteria:

1. Study design: the study reports on a TEL intervention in a real-world environment,
2. Participants: the study concerns a population of learners or educators,
3. Technology: the study discusses a technology with a direct impact on the learning experience,
4. Validity criteria: the study evaluates the TEL intervention using at least one clearly defined validity criterion.

The reason for focusing on intervention studies is that they are able to address validity criteria covering the full spectrum from design to implementation, hopefully preventing the introduction of a bias in validity criteria purely caused by the type of study considered. Additionally, by selecting only intervention studies, we establish a focused scope for this review. In future reviews, broadening this scope could offer valuable insights. A table summarising all inclusion and exclusion criteria can be found in our protocol.

7.3.2 *Selection process*

We used the ASReview screening software (van de Schoot et al., 2021) to perform title and abstract screening. ASReview optimises the title and abstract

screening process through the use of active learning, whereby reviewers are presented with the most relevant studies first. We initialised the ASReview process with two reviewers who independently screened a set of 100 randomly sampled articles. Both reviewers agreed on the exclusion of 87 articles and the inclusion of 7 articles. For 6 articles there was initial disagreement, leading to a Cohen's kappa of $\kappa = 0.67$. All disagreements resulted from different interpretations of abstracts, rather than from a fundamentally different understanding of which studies should be included. After discussion with two further reviewers, unanimous agreement was reached and the screening process was continued.

We estimated the total number of relevant articles in our set based on the random sample of 100 studies. We included 11 of the 100 studies. Our stopping criterion for the main screening phase specified that the reviewer should stop once they had reached 95% of the estimated number of relevant papers or had encountered 20 irrelevant articles in a row. The estimated number of relevant papers was $1,256 \times 0.11 = 138.16$, implying that the criterion was to stop screening once we had reached a total of 132 ($138.16 \times 0.95 = 131.25$, rounded up) relevant papers. This approach is based on the approach used in earlier research (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021). The 132 relevant records were reached after screening 374 records in total.

After title and abstract screening, we attempted to retrieve full-text articles for all potentially relevant records. At this stage, two articles were excluded as the main text was not written in English, and one article was excluded because a similar study by the same author was included. Ten articles could not initially be retrieved. We managed to contact the authors of seven of these articles, resulting in one additional full text inclusion. In total, 12 articles were excluded at this stage, resulting in 120 remaining inclusions.

We then performed backward snowballing to identify additional studies that may have been missed through the database search. We sorted all inclusions randomly before initialising the snowballing procedure. We stopped the backward snowballing phase once we had reviewed at least 10 inclusions fully and consequently reached a point where 100 references in a row were considered irrelevant. The snowballing phase was conducted by a single reviewer, with a second reviewer screening all papers that the first reviewer marked for inclusion. We evaluated a total of 543 references, resulting in a further 34 inclusions, on which the two reviewers reached full agreement.

Figure 7.1 summarises the selection process using a PRISMA flowchart. The final step of assessing data quality is discussed next. At this stage, 107 papers remained after we had excluded 47 of our 154 inclusions based on their full text content.

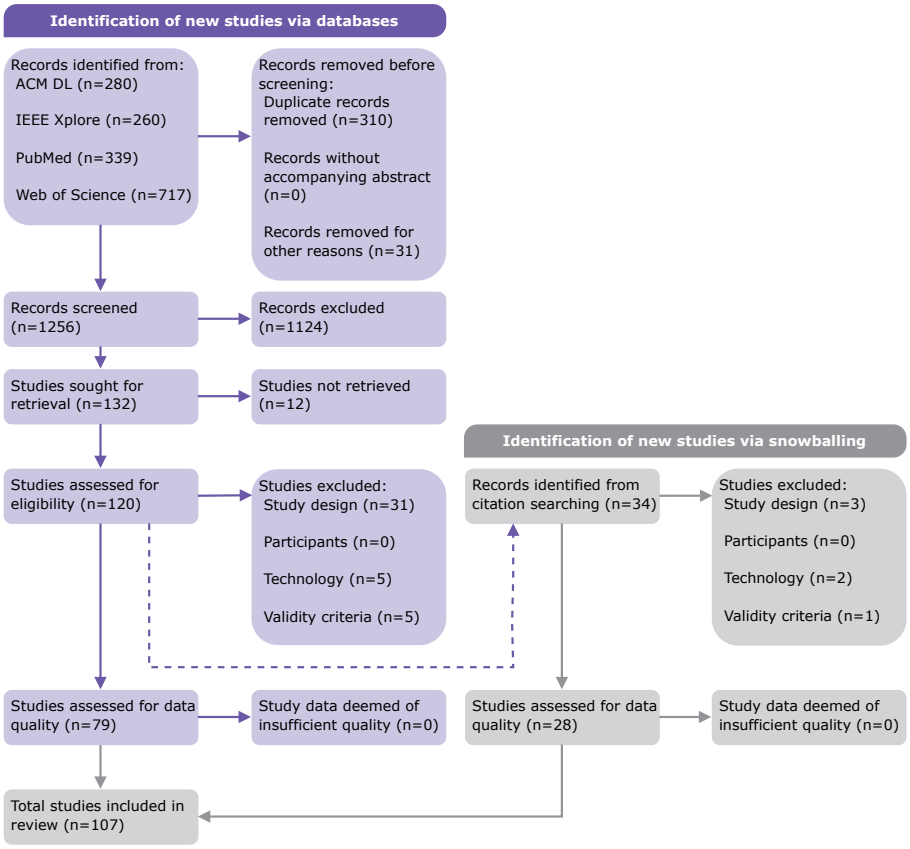

Figure 7.1: The PRISMA flowchart for our systematic review.

7.3.3 *Data extraction*

Data was extracted based on a data extraction form which can be found on our open data repository page. To ensure consistency, a pilot test of the data extraction form was conducted by two reviewers on a sample of 20 included studies. Any discrepancies were resolved through discussion. Our protocol initially specified that the pilot test would be conducted on a sample of 10 studies, but we found it necessary to have more data extraction examples to facilitate a detailed discussion, and therefore increased the sample to 20.

The data extraction form was revised before continuing with the full data extraction process, which was performed by a single reviewer. Compared to our original data extraction form as specified in our protocol, we included an item specifying the number of participants in a study and three items relating to the use of an evaluation or validation framework. We additionally noticed a focus on evaluation rather than validation in the first studies we analysed, and therefore decided to monitor the extent to which studies addressed the evaluation criteria as specified in the taxonomy of Lai and Bower (2019). Once data extraction was complete, two reviewers evaluated the completeness, accuracy, and consistency of the data. We did not encounter any issues, which is reflected in the fact that no studies were excluded due to insufficient data quality.

Table 7.2 lists the validity criteria that we consider in this systematic review. We provide possible synonyms for these criteria as observed in the literature (see e.g., Cronbach and Meehl (1955), Lincoln and Guba (1986), Mingers and Standing (2020), Straub (1989), and van Haastrecht, M. Brinkhuis, Peichl, et al. (2023)). Three changes were made compared to our protocol. Effectiveness replaced consequential validity as a main criterion, joyfulness was added as a criterion, and helpfulness was added as a synonym for usefulness. In cases where studies report validity criteria using different terminology or concepts, we attempted where possible to map these to the appropriate validity criteria listed in Table 7.2. We consciously chose not to prescriptively posit our own definitions for these validity criteria, but rather to take author's claims of assessing particular criteria at face value. If authors claim to assess trustworthiness, we included the criterion of trustworthiness for their study. To provide more insight into the implicit definitions used by researchers, we paired each included criterion with an accompanying quote from the relevant study, which can be found in the extracted data in our open access data repository.

For each validity criterion, we determined the research method used, the knowledge source for the evidence, and the outcome of how the criterion was assessed. For the research method item, we used the categories quantitative, qualitative, and mixed. If the evidence used to assess a validity criterion came from a qualitative approach such as an interview, the research method label for that criterion would be qualitative. From an epistemological point

Table 7.2: Validity criteria considered in our systematic review and their synonyms.

VALIDITY CRITERIA	SYNONYMS
Actionability	Practicability
Authenticity	Genuineness, originality, ecological validity
Confirmability	Auditability, accountability
Effectiveness	Consequential validity, impact, social validity
Consistency	-
Construct validity	Convergent validity, discriminant validity, specificity, structural validity
Content validity	Face validity, representativeness, comprehensiveness, objectivity [context: unbiased content]
Credibility	Authority
Criterion validity	Concurrent validity, predictive validity, empirical validity [context: predictive ability], accuracy
Dependability	-
Elegance	Appealingness, attractiveness, beauty, gracefulness
External validity	Generalisability, population validity, sample representativeness
Fairness	Impartiality, unbiasedness, equity
Internal validity	Causal validity
Joyfulness	Delightfulness
Meaningfulness	Significance [context: personal impact]
Parsimony	Simplicity
Relevance	Applicability, pertinence, suitability
Reliability	-
Replicability	Reproducibility, repeatability, objectivity [context: replicable research methodology]
Rigour	Thoroughness, soundness
Statistical validity	Statistical significance, empirical validity [context: correlation], statistical robustness
Transferability	Portability
Trustworthiness	Integrity
Understandability	Clarity, comprehensibility, interpretability, intuitiveness, transparancy
Usability	User-friendliness, accessibility, ease of use
Usefulness	Helpfulness, practicality, utility

of view, it is common to distinguish knowledge sourced from theoretical reasoning and knowledge sourced from practice. For the knowledge source item, we therefore used the categories theory and practice. If the evidence used to assess a validity criterion resulted from theoretical reasoning, as is often the case for statistical validity, the knowledge source label would be theory. Conversely, if the evidence resulted from feedback from students, the knowledge source label would be practice. Finally, we investigated whether the eventual outcome of validity criteria assessments was positive, negative, or mixed. If authors measured relevance and concluded based on their evidence that their solution was indeed relevant, the assessment was labelled as positive. A mixed assessment could be achieved if the evidence was inconclusive, or if certain evidence pointed to a positive assessment while other evidence pointed to a negative assessment.

7.4 RESULTS

Our main research question asks how we can characterise the landscape of validity criteria used in TEL research. To answer this question we formulated three sub-questions, which we will cover in the three subsections below.

7.4.1 *Which validity criteria are considered?*

Our first sub-question aimed to investigate which validity criteria TEL research considers and how researchers define and measure these criteria. Figure 7.2 shows how often each criterion was encountered in our inclusions. Effectiveness (82 appearances) and statistical validity (78) were the most commonly assessed criteria and Figure 7.2 shows that they tended to be positively assessed based on results from quantitative methods. In contrast, criteria such as external validity (34) and rigour (23) tended to be negatively assessed based on results from qualitative methods.

We extracted 298 criteria from our inclusions where the underlying argumentation resulted from a predominantly quantitative research method. Of these criteria, 34 (11.4%) were assessed negatively. Compare this to the 137 instances of predominantly qualitatively researched criteria, of which 77 (56.2%) were assessed negatively. Even when excluding external validity and rigour, which were often mentioned in the limitations section of research, 24 (30.0%) of the remaining 80 instances of predominantly qualitatively motivated criteria were assessed negatively. We can additionally observe that theoretically underpinned criteria were generally either argued for qualitatively and assessed negatively, or argued for quantitatively and assessed positively. This places these criteria at the extremes of the Figure 7.2 grid.

Figure 7.3 shows the research method used to assess each validity criterion instance. We excluded the two papers with more than 1,000 participants in

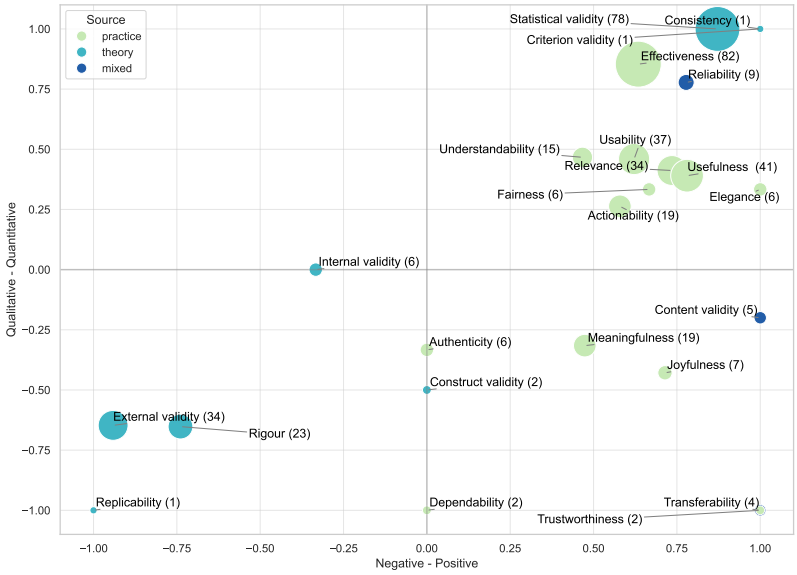

Figure 7.2: Bubble plot of validity criteria encountered in our inclusions, where each bubble is coloured depending on which knowledge source (on average) was used to assess it.

this plot to avoid readability issues due to scaling. When qualitative methods or mixed-methods were used by researchers, negative assessments were much more common than when quantitative methods were used. Since qualitative studies tend to have lower numbers of participants, one could wonder whether the connection between research method and assessment is caused by this mediating variable. To disentangle the number of participants and the research method, criteria are additionally visualised based on the number of participants in the study they were encountered in. The number of participants does not appear to be strongly correlated to the assessment.

Besides validity criteria, TEL researchers often consider constructs that are more directly tied to evaluation. With this in mind, we extracted data on the constructs used for TEL evaluation based on the list of constructs and construct themes presented in Lai and Bower (2019). Table 7.3 depicts the result of this work. Compared to Lai and Bower (2019), we observe relatively more criteria considered per paper, which can be explained by our confined search focusing only on studies that describe the criteria they use. We additionally observe less usage of established instruments and frameworks, which can partially be explained by the larger time window of our search and the fact that earlier papers used established instruments less frequently.

Figure 7.4 shows that differences with the results of Lai and Bower (2019) can be primarily attributed to a decrease in studies using established instru-

Table 7.3: Constructs used for TEL evaluation, following the exact structure of Table 6 in Lai and Bower (2019).

		Papers		Instruments	
THEMES/ASPECT (NO. OF PAPERS, %)	SUB-THEME CONSTRUCTS	NO.	%	ESTABLISHED	SELF-DEVELOPED
Learning (103, 96.3%)	Knowledge, achievement or performance	96	89.7%	25.0%	75.0%
	Cognitive load/effort (e.g., mental effort)	19	17.8%	36.8%	63.2%
	Skills development (e.g., interpersonal skills, motor skills, verbal and non verbal skills or communication skills)	39	36.4%	25.6%	74.4%
	Learning styles or learning strategies	27	25.2%	25.9%	74.1%
Affective Elements (82, 76.6%)	Perceptions, intentions or preferences	62	57.9%	24.2%	75.8%
	Engagement, motivation or enjoyment	50	46.7%	26.0%	74.0%
	Attitudes, values or beliefs	20	18.7%	30.0%	70.0%
	Emotional problems, anxiety or boredom	14	13.1%	35.7%	64.3%
	Self-efficacy	15	14.0%	20.0%	80.0%
Behavior (84, 78.5%)	Usage or participation	53	49.5%	20.8%	79.2%
	Interaction, collaboration or cooperation	52	48.6%	23.1%	76.9%
	Self-reflection, self-evaluation or self-regulation	20	18.7%	10.0%	90.0%
Design (59, 55.1%)	Course quality, course content, course structure, resources or overall design	59	55.1%	22.0%	78.0%
Technology (73, 68.2%)	Functionality	13	12.1%	38.5%	61.5%
	Perceived usefulness	45	42.1%	26.7%	73.3%
	Perceived ease of use	41	38.3%	24.4%	75.6%
	Adoption	3	2.8%	100.0%	0.0%
	Accessibility	34	31.8%	14.7%	85.3%
Teaching/Pedagogy (56, 52.3%)	Pedagogical practice, teaching strategies or teaching quality/credibility	49	45.8%	22.4%	77.6%
	Feedback	28	26.2%	32.1%	67.9%
Presence (10, 9.3%)	Social presence, co-presence or community	10	9.3%	20.0%	80.0%
	Presence in the environment	0	0.0%	-	-
Institutional Environment (11, 10.3%)	Institutional - institutional capacity, institutional intervention, institutional policy or institutional support	8	7.5%	25.0%	75.0%
	External environment/factors	3	2.8%	66.7%	33.3%

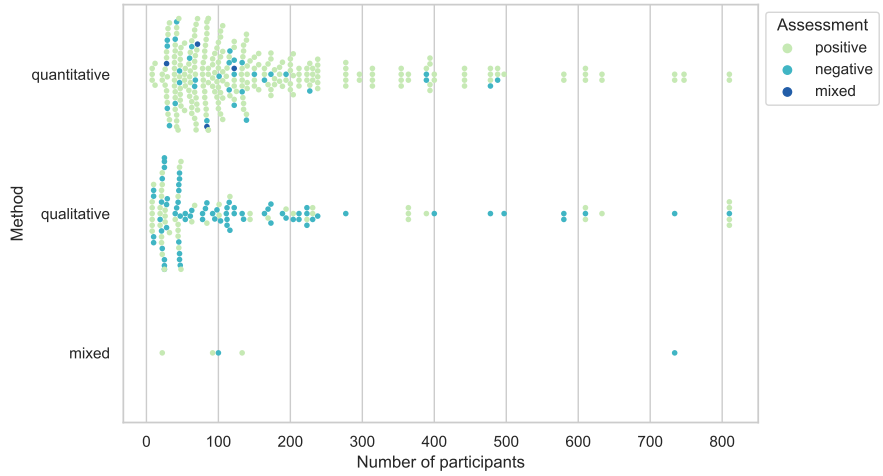

Figure 7.3: Plot of validity criteria occurrences, split by the number of participants in the corresponding study and the method used to assess the criteria. Criteria are coloured by whether the assessment was positive, negative, or mixed.

ments to evaluate learning and affective elements, and an increase in studies using self-developed instruments to evaluate technology and teaching/pedagogy. Altogether, 27 of our 107 inclusions used an established evaluation or validation framework, with the most commonly used framework employed only 4 times. This fragmentation in the use of frameworks was also found by Lai and Bower (2019), where the most common framework appeared 20 times among their 243 inclusions that applied an established instrument.

7.4.2 *How does criteria prevalence change over time?*

Our second sub-question asked how the prevalence of validity criteria has changed over time in TEL research. Figure 7.5 shows the percentage contributions of the ten most commonly encountered criteria in our inclusions over time. The bar plot stacks the top ten criteria from most frequently occurring overall at the bottom (effectiveness) to least frequently occurring at the top (understandability). The height of each individual bar within a year represents the percentage contribution of a criterion. The total height for a particular year represents the percentage contribution of the top ten criteria. We observe that the most frequently occurring criteria overall tend to be the most frequently occurring criteria per year. This is a first signal of the temporal stability of the validity criteria landscape in TEL.

Figure 7.6 is a variant of the Figure 7.2 bubble plot, but with each subplot showing criteria prevalence during the period of the subplot title. Time win-

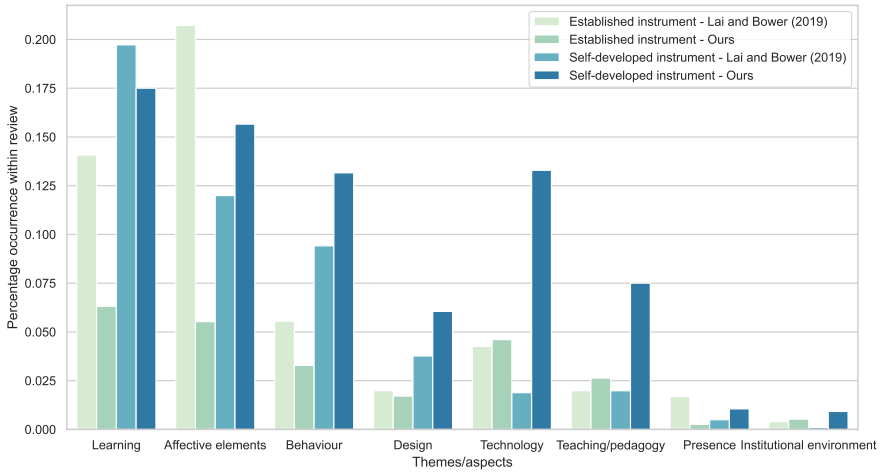

Figure 7.4: Comparison of the percentage occurrence of evaluation themes/aspects in Lai and Bower (2019) and this review.

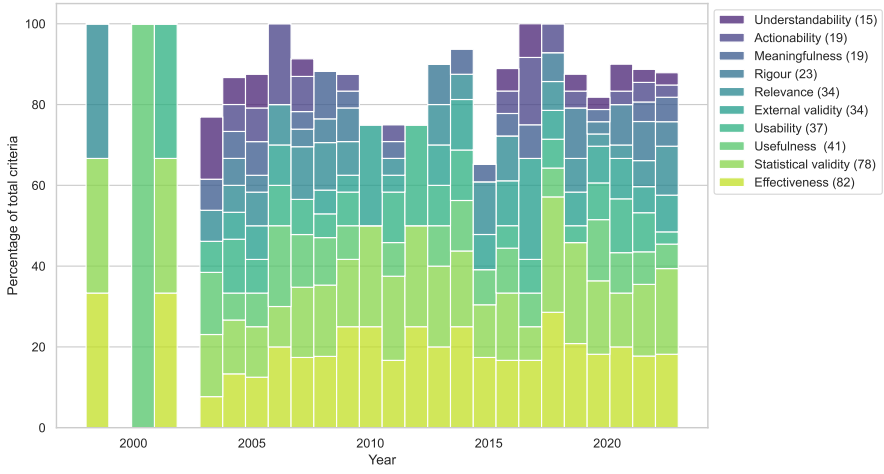

Figure 7.5: The ten most frequently encountered criteria overall, plotted per year from 1998-2023.

dows were chosen such that individual plots contain roughly equal numbers of validity criteria instances. Although we observe movement in the placement of certain criteria, the overall picture is relatively constant, with criteria that are coloured yellow and light green remaining in the top right and criteria that are coloured dark blue remaining in the bottom left. Of the 24 criteria shown in Figure 7.2, 20 are already included in the 1998-2009 plot. Of these 20, 17 are in the same quadrant overall as they were in 1998-2009, and 15 are both in the same quadrant and have the same knowledge source categorisation. The three criteria where the quadrant changes are elegance, fairness, and usefulness. In each case, the quadrant change was from bottom-right in 1998-2009 to top-right overall, meaning these positively assessed criteria were more commonly evaluated using quantitative methods in later years. Elegance and fairness had been assessed just once and twice, respectively, by 2009. The quadrant change for usefulness is more significant, as it had been assessed 15 times by 2009. However, the change from being primarily qualitatively assessed (y=-0.07) to being primarily quantitatively assessed (y=0.39) was not major. Figure 7.6 points to stable definitions and interpretations of criteria over time, thereby providing an affirmative answer to the question: Is there a common ground from which to build a comprehensive validation framework?

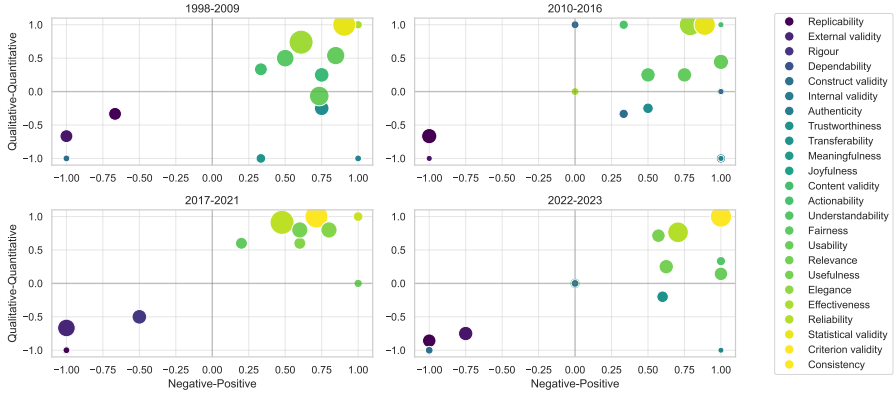

Figure 7.6: A grid of bubble plots visualising validity criteria positioning in different periods of time. Colours are determined by the position of a criterion in the overall plot of Figure 7.2.

7.4.3 *What epistemological patterns do we observe?*

Our final sub-question asked whether we observe any epistemological patterns in the connections between validity criteria. We concluded from Figure 7.2 and Figure 7.3 that there are observable relations between the method used to assess a criterion, the knowledge source used to inform this assessment,

and the eventual outcome of the assessment. However, these figures do not allow us to analyse the connections between validity criteria. Figure 7.7 represents a network visualisation of validity criteria, where the edge weights are determined by the relative co-occurrence of the target criterion in the papers where the source criterion was assessed. Node sizes are determined by how often a criterion was encountered and node colour is determined by whether a criterion was largely positively assessed (green) or largely negatively assessed (red). A reduced version of the total network is shown, as we only depict edges with co-occurrence scores of at least 90% of the maximum co-occurrence score per criterion. For example, the large edge weight for the edge going from dependability to authenticity indicates that the criterion of authenticity was encountered more often in the papers assessing dependability than we would expect based on the prevalence of authenticity as a criterion.

Figure 7.7 shows several clusters of validity criteria that are interconnected, as well as pairs of criteria such as fairness and transferability. The lack of strong connections emanating from statistical validity and effectiveness points to the ubiquity of these criteria. Even when effectiveness co-occurs with other criteria quite often, the corresponding edge weights will still be relatively small since the base probability of co-occurrence with effectiveness is high. Another explanation for the lack of strong connections could be that researchers tend to judge these criteria to be essential to their studies, regardless of the type of study. Metaphorically, effectiveness and statistical validity are acquainted to every criterion, but true friends with none.

One way Figure 7.7 can be useful is in helping to select criteria that together form an epistemologically complete set. When designing a validation framework for TEL, one might start with the inclusion of the top ten criteria shown in Figure 7.5. Figure 7.7 can then help to unearth which clusters of validity criteria would not yet be covered by this initial set, such as the pair internal validity-construct validity and the pair fairness-transferability. An extended framework that incorporates internal validity and fairness could then be considered epistemologically more comprehensive.

7.5 DISCUSSION

The results presented in the previous section provide answers to the research questions we posed, but also raise new questions that we will discuss further.

7.5.1 *A problematic hierarchy of validity criteria*

Figure 7.2, Figure 7.5, and Figure 7.6 visualise the prevalence and epistemological positioning of validity criteria. These visualisations suggest the existence of a hierarchy of validity criteria, which can be construed as problematic. At

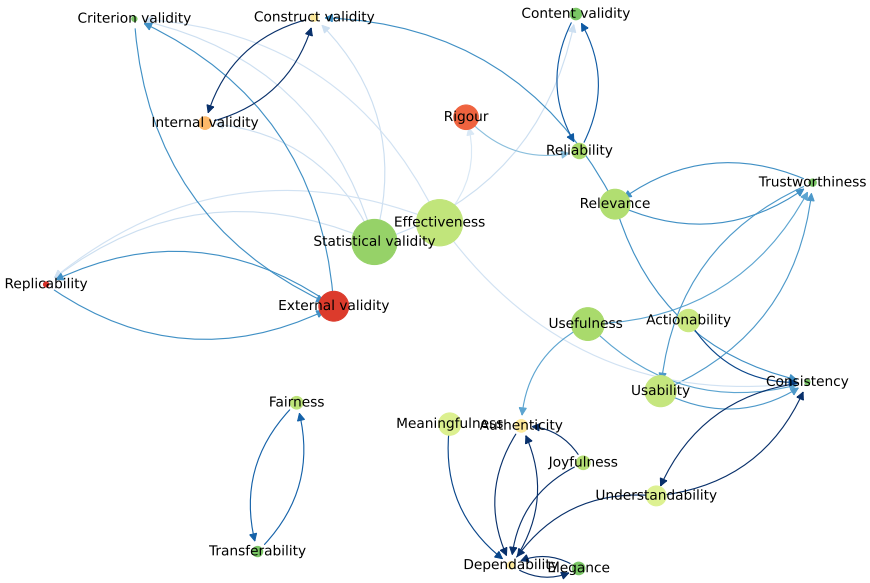

Figure 7.7: Network showing the relative co-occurrence of validity criteria.

the top of this hierarchy we find effectiveness and statistical validity. Around 75% of our inclusions assessed these criteria, and they were overwhelmingly assessed positively. However, Salehi et al. (2023) raise an important point regarding effectiveness and statistical validity that is generally not discussed in our inclusions. In a study of continuing professional development for 10,000 health workers in Ghana during the pandemic, the researchers mention in their discussion of e-learning effectiveness: "While these effect sizes are useful in painting an overall picture, with education evaluation, a 'small' effect size on a difficult-to-change variable (e.g., attitude toward recommending the vaccine) could be as valuable as a larger effect size on something easier to change (e.g., knowledge)" (Salehi et al., 2023, p. 10). Valuing particular criteria highly is not problematic in itself, but, as Salehi et al. (2023) point out, it is vital to critically contextualise validity evidence.

At the bottom of the hierarchy we find external validity, often termed generalisability, and rigour. External validity was assessed 34 times, with only one study reaching a positive conclusion. Of the 33 negative assessments, 28 times researchers mentioned that that their study focused on one educational context, and that this implies their results do not generalise to other contexts. Interestingly, the criterion of transferability, which can be seen as a counterpart to external validity (see, e.g. van Haastrecht, M. Brinkhuis, Peichl, et al. (2023)), was assessed positively 100% of the time. This points to the feasibility of

designing studies that produce generalisable results. An illustrative example is the one study that assessed external validity positively, which conducted a multi-centre randomised controlled trial (Vivekananda-Schmidt et al., 2005).

For studies that reached a negative conclusion about the rigour of their approach, common issues that were mentioned were the possibility of accidental exposure of the control group to the treatment (Tsai, 2010), the inability to mitigate certain biases due to the methodology used (Whitaker et al., 2007), and the overall lack of control over the experimental situation encountered during the COVID pandemic (Başağaoğlu Demirekin and Buyukcavus, 2022). Yet, there were positive examples too. One randomised controlled trial stated: "the major strength of this study is the robust methodology and adherence to protocol for each candidate once randomised" (Brewer et al., 2021, p. 5). A study applying a qualitative analysis of student reflections during the COVID pandemic concluded that "the strength of the study is that it provides quite a comprehensive picture of the students' experiences" (Wojniusz et al., 2022, p. 8).

Concerns have been voiced in earlier work about the troubling manner in which TEL research distinguishes between validity criteria in the upper echelons of the hierarchy, such as statistical validity, and criteria lower down, such as external validity (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). We are not calling for all studies to prioritise every validity criterion, as this is impossible. Yet, we need to ensure that as a field we do not structurally ignore certain criteria, while structurally prioritising, but not critically contextualising, other criteria.

7.5.2 *Framework foundations without structure*

Perhaps the most important finding from Section 7.4 is that there exists temporal stability in the usage and epistemological interpretation of TEL validity criteria. Figure 7.5 and Figure 7.6 convincingly showed that the same validity criteria have been prioritised by researchers throughout the last decades, and that the manner in which they have been assessed has remained remarkably constant. Naturally, one might conclude that the necessary foundations are present for a consensus validation framework.

However, we observed in Table 7.3 that usage of established frameworks is minimal. Additionally, similarly to Lai and Bower (2019), we found that there is a high degree of fragmentation in the use of frameworks. We suggested in Section 7.4.3 that our network analysis in Figure 7.7 could be a useful aid in selecting an epistemologically complete set of validity criteria for a validation framework. But a set of validity criteria is only the basis for a framework. A comprehensive framework requires a structure within which these validity criteria should be assessed and related to each other.

The argument-based approach to validation could offer the exact structure that TEL validation is currently lacking. In Section 7.2.1, we covered several

validation frameworks that have been proposed in recent years building on the argument-based approach (Huggins-Manley et al., 2022; van Haastrecht, M. J. S. Brinkhuis, Wools, et al., 2023; Zhai et al., 2021). Yet, we also highlighted that these frameworks are currently rarely employed in TEL research. There are criticisms regarding how these frameworks deal with TEL process data (Zumbo et al., 2023) and concerns whether they leave sufficient space for legitimately diverse arguments (Addey et al., 2020). Nevertheless, with the frameworks that already exist and the temporal stability present in TEL validation, there is clear promise for future holistic validation frameworks such as those based on the argument-based approach.

7.5.3 *Quantitative positivity: correlation or causation?*

We highlighted in Section 7.4.1 that there exists a correlation between the research method used to gather evidence regarding a validity criterion and the eventual assessment outcome. Not a single criterion in the quantitative half of the diagram in Figure 7.2 was on average assessed negatively. The question is whether there are any causal factors at play. Our research design was not suited to answer any causal questions regarding the relationship between research method and assessment outcome. However, we can present hypotheses that can be investigated in further research. Based on our discussion of a validity criteria hierarchy, one hypothesis is that the correlation is caused by publication bias. If predominantly quantitative criteria are considered more important than predominantly qualitative criteria, studies with negative assessments regarding quantitatively researched validity criteria would be less likely to get published than studies with negative assessments based on qualitative methods. One way to assess the hypothesis that a publication bias offers an explanation for the trend we observe in Figure 7.2, would be to survey TEL researchers. The researchers could be asked whether they consider research with negative quantitative results fit for publishing and whether they have experienced papers with negative quantitative results being rejected more often than papers of comparable quality with negative qualitative results.

Another hypothesis is that it is not the researchers, but rather the participants, that are causing the observed correlation. Quantitative approaches, such as questionnaires using Likert scales, condense constructs down to a numerical scale. In his seminal qualitative research work, Geertz (1973) delineates how qualitative methods are in search of meaning whereas quantitative methods are in search of law. One of our inclusions that applied qualitative methods was Rossiter et al. (2024), a study explaining the design and evaluation of a mobile learning resource for university students. A telling example of how qualitative methods can leave room for meaning over law comes from a student quote regarding the new resource's trustworthiness. The student explained: "I think I sort of trusted it a bit more because it felt like it was made by you for me as opposed to very general random videos that might be on

the subject area" (Rossiter et al., 2024, p. 119). A quantitative approach would not allow space for the meaning behind the student's positive assessment, and would likely abstract away this individual opinion into an aggregated number that serves as our law. A hypothesis to explain the correlation we observe in Figure 7.2 could thus be that quantitative methods leave less room for nuanced assessments and inadvertently hide negative or mixed feedback. A way to test this hypothesis would be to assess a set of constructs both quantitatively and qualitatively in a controlled environment. One could then examine whether quantitatively assessed constructs are evaluated significantly more positively.

7.5.4 *Limitations and threats to validity*

We should mention that this study has its limitations, along with potential threats to the validity of our conclusions. Firstly, although the search strategy we employed was geared at capturing all relevant studies for our systematic review, we cannot rule out the possibility that relevant papers were missed. An example of studies we may have missed are those that use wording in their title and abstract that deviate from the terminology of our search query. For example, we did not use the term 'online learning' in our original query. However, our systematic review process incorporating ASReview mitigates this risk by allowing for a broad database search with many related terms, and we additionally included a snowballing step which allowed us to identify relevant papers independently from our search query. For the case of the omitted term 'online learning,' our broad search strategy resulted in nevertheless having 22 of 107 papers including this term in the title or abstract. Additionally, only 4 of the 28 snowballing inclusions used the term 'online learning,' demonstrating that our search query did not miss disproportionately many studies for this term.

A potential threat to validity is the bias that the reviewers may have introduced into our screening process. Reviewers may have had personal biases that influenced which studies were included and how data was extracted. We believe the process we specified in our protocol and carried out for this study, where multiple reviewers were involved at each step of the systematic review, helped to minimise the risk associated with individual reviewer bias. Furthermore, by making our protocol available within an open-source project, we are transparent about our process and facilitate potential replication of our review.

Finally, although the 107 included papers and 440 extracted validity criteria constitute a comprehensive representation of the TEL literature, we have seen in Section 7.4 that certain criteria listed in Table 7.2 were either not encountered or rarely encountered. This could imply that our network analysis produced different results than if a larger set of papers would have been considered. The strictness with which reviewers followed our systematic

review protocol significantly decreases the probability that a replication study would find decidedly different results, but we would certainly welcome a large-scale systematic review that would enable deeper insights into the connections between TEL validity criteria.

7.6 CONCLUSION AND FUTURE WORK

Technological innovations have provided a diverse array of opportunities to optimise educational environments, but have also introduced new challenges in assessing the validity of novel solutions. We have seen in this chapter that the use of evaluation and validation frameworks in TEL research is rare and fragmented. However, we found that there is a clear light at the end of the validation tunnel. We demonstrated that the TEL validity criteria landscape has been remarkably stable over time. Both the types of validity criteria that are most commonly assessed and their epistemological positioning have stayed relatively constant over the past two decades. The stability in validity criteria usage and definitions offers a solid foundation from which to build future validation frameworks, where we highlighted the promise of argument-based validation to serve as the guiding structure.

There is a long road ahead before the use of holistic validation frameworks becomes commonplace in TEL research. Existing argument-based validation frameworks need to continually adapt to the changing world, with a constant need to recognise diverse perspectives and epistemologies. In our discussion section, we outlined several open questions whose answers would aid progress towards more holistic validation strategies. We observed a clear correlation between the research method used to assess validity criteria and the outcome of the assessment. Further research will need to determine whether the cause for this correlation lies with publication bias on the side of the research field, or with the inherent challenge of uncovering nuance and meaning using quantitative methods. Finally, future work will need to critically examine the problematic hierarchy of validity criteria that currently exists. We argue for a situation where validity criteria are prioritised critically based on their contextual relevance, rather than selected blindly based on their perceived importance.