Available on: https://jsshrs.ep-journals.org/ E-ISSN: 3051-3561, P-ISSN: 3051-3553

Journal of Social Science and Human Research Studies

Volume : 02 Issue : 10 October-2026

Measuring AI-Pedagogical Knowledge in Pre-Service Teachers: Development and Initial Validation of a Vignette-Based Assessment

Abdul Hafid
Department of Primary School Teacher Education, Universitas Negeri Makassar

Rukayah
Department of Primary School Teacher Education, Universitas Negeri Makassar

Muh. Faisal
Department of Primary School Teacher Education, Universitas Negeri Makassar

Nurul Mukhlisah Abdal
Department of Informatics and Computer Engineering, Universitas Negeri Makassar

Ahmad Syawaluddin
Department of Primary School Teacher Education, Universitas Negeri Makassar

Published Date: October 02, 2026 **Article DOI: 10.65150/EP-jsshrs/V2E10/2026-02**

ABSTRACT : The growing use of generative artificial intelligence (AI) in education creates a need for assessments that measure teachers' pedagogical judgement rather than relying solely on self-reported competence. This study developed and examined initial validity evidence for a vignette-based assessment of AI-pedagogical knowledge among pre-service teachers. Instrument development proceeded from construct definition and blueprinting to an initial pool of 36 researcher-developed vignettes, developer screening, a diagnostic pilot with 60 pre-service teachers, systematic item revision, and content review by three independent experts. The revised 24-item assessment was administered to an analytic sample of 484 pre-service teachers. Validity evidence was examined through expert ratings, internal structure, item and distractor functioning, score reliability and conditional precision, differential item functioning, and relations with external variables. Aiken's V coefficients ranged from .83 to 1.00 (M = .94), although the small expert panel and restricted use of rating categories warrant cautious interpretation. A one-factor model adequately represented the response data, supporting interpretation of a total score rather than separate domain scores (CFI = .990, TLI = .989, RMSEA = .034, SRMR = .065). Internal consistency was high (KR-20 = .93; categorical omega = .95), but precision was concentrated at the lower end of the ability distribution. Of the participants, 24.2% obtained the maximum score and 47.5% scored 22 or higher. No item showed non-negligible differential functioning by gender or field of study. Knowledge scores were moderately associated with AI self-efficacy (r = .36) and teaching self-efficacy (r = .33) and more weakly associated with self-reported instructional practice (r = .22). The assessment may support formative and research uses when interpreted as a single total score, but it is not yet suitable for ranking or high-stakes decisions. Further development requires more difficult items, revision of weak items and distractors, cognitive interviews, and cross-institutional evaluation.

KEYWORDS: AI-pedagogical knowledge, Pre-service teachers, Vignette-based assessment, Instrument development and validation, Generative artificial intelligence, Teacher education, Psychometric properties.

1. INTRODUCTION

In several education systems, the use of generative artificial intelligence (AI) is increasing more rapidly than the formal guidance and preparation available to teachers and students. In a nationally representative triangulation of five populations in the United States, 54% of students and 53% of English language arts, mathematics, and science teachers reported using AI for school in 2025, increases of more than 15 percentage points within one to two years (Doss et al., 2025). Institutional preparation remained uneven: as of fall 2024, 55% of teachers reported that their school or district had provided professional development related to generative AI, while by spring 2025 only 35% of district leaders reported providing students with any AI training. More than 80% of students also indicated that their teachers had never explicitly taught them how to use AI for schoolwork (Doss et al., 2025). These findings document substantial uptake alongside incomplete instructional support in the United States; they do not, by themselves, establish a uniform global pattern..

Related concerns are evident in teacher education. In a case study of pre-service EFL teachers' lesson planning, participants used generative AI for topic selection, material creation, lesson organisation, and language checking, but the study did not document its use to critique lesson plans

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

or examine their pedagogical effectiveness (Kerr & Kim, 2025). This pattern illustrates an important distinction: the ability to produce instructional material with an AI tool does not necessarily demonstrate the ability to evaluate whether that material is appropriate for a particular learning objective, learner group, or assessment purpose.

Basic operation of generative AI may require less preparation than evaluating the pedagogical quality of its outputs. A lesson plan can be fluent, well organised, and factually plausible while still misaligning cognitive demand with learning objectives. Similarly, an automatically generated feedback comment may appear supportive without providing information that a learner can act on. Identifying such limitations requires pedagogical judgement in addition to technical familiarity.

Together, these observations identify a need for an objectively scored assessment of context-specific pedagogical and ethical judgement among pre-service teachers. This study therefore develops a vignette-based assessment of AI-pedagogical knowledge and provides initial validity evidence for the intended interpretation of its scores.

2. THEORETICAL FRAMEWORK

This distinction has direct implications for how teacher AI competence is measured. Research on teachers' AI competence has produced several carefully developed instruments, but many prominent measures rely on self-report. Ning et al. (2024) developed a 39-item *AI-TPACK* scale spanning seven dimensions. Choudhury et al. (2024) subsequently examined the scale with 660 university faculty members in India and reported satisfactory reliability. Setiyawan et al. (2025) developed and validated an AI-TPACK instrument for vocational education, Chiu et al. (2025) used a Delphi procedure to construct the Teacher AI Competence Self-Efficacy scale across six dimensions, including AI pedagogy, AI assessment, and AI ethics. These instruments provide useful evidence about perceived knowledge, confidence, and readiness. Their scores, however, should not automatically be interpreted as direct evidence of demonstrated knowledge or judgement.

This distinction is particularly important when perceived competence and reported practice are collected from the same respondent, during the same administration, and with similar response formats. Under these conditions, observed associations may reflect both substantive relationships and variance attributable to the shared measurement method (Podsakoff et al., 2003). Common-method variance does not necessarily invalidate self-report findings, nor does it always inflate associations, but it can make it difficult to distinguish relationships among beliefs from relationships involving demonstrated competence. Self-report and performance-based measures therefore address different questions and are most informative when their intended interpretations are stated explicitly

To complement self-report approaches, objectively scored assessments of AI literacy have begun to emerge. Jin et al. (2025) developed the Generative AI Literacy Assessment Test (GLAT), a 20-item multiple-choice instrument validated with higher education students. GLAT scores predicted performance on AI-assisted tasks, whereas self-reported ChatGPT proficiency did not. Markus et al. (2025). similarly introduced the AI Competency Objective Scale (AICOS) for broader educational and professional populations. These developments show that the measurement landscape is no longer limited to self-report, although many available objective tests assess broad AI literacy rather than the pedagogical judgement required for classroom decisions.

The distinction has also been examined directly among educators.Zhang et al. (2026). developed self-reported and objective measures across shared AI literacy dimensions and found only a weak relationship between the resulting factor scores. Their results indicate that perceived and demonstrated AI literacy can diverge within a teacher sample. However, a discrepancy should not automatically be interpreted as miscalibration because its meaning depends on whether the two measures cover sufficiently aligned content.

Potential consequences of inaccurate AI-related judgement can be illustrated by the use of AI-text detectors in academic-integrity decisions. Liang et al. (2023) found substantial false-positive rates when widely used detectors were applied to essays written by non-native English speakers. A teacher who is confident in such a detector without understanding its limitations may make an unjustified decision affecting a student. This example does not establish how frequently such decisions occur, but it demonstrates why perceived capability and demonstrated knowledge should not be treated as interchangeable.

A useful precedent for separating knowledge from self-efficacy comes from teacher education research. Depaepe & König (2018) examined general pedagogical knowledge, self-efficacy, and reported instructional practice in 342 pre-service teachers. Knowledge was assessed using an objective test, whereas self-efficacy and instructional practice were measured by self-report. They found no statistically significant association between general pedagogical knowledge and self-efficacy. Self-efficacy was associated with all reported instructional-practice scales, while general pedagogical knowledge was associated only with practices concerning student support and the provision of structure.

These findings support treating knowledge and self-efficacy as conceptually and empirically distinguishable, but they do not establish that the association between them is exactly zero or that the stronger self-report associations were caused by common-method variance. Instead, they suggest that the magnitude and pattern of the relationships should be estimated directly within each domain. In the AI context, this distinction is especially relevant because frequent access to AI tools may increase familiarity and confidence without necessarily producing equivalent gains in pedagogical judgement.

Vignette-based and situational judgement approaches offer one way to assess the application of domain knowledge under controlled conditions. Rather than asking respondents how capable they believe themselves to be, these approaches present a defined situation and require respondents to evaluate or select a response. Situational judgement test performance has been interpreted as reflecting knowledge about the relative effectiveness of possible actions within a domain (Lievens & Motowidlo, 2016). Such assessments do not directly observe classroom behaviour, but they can sample judgement in realistic contexts more closely than general self-efficacy ratings. Applications focused specifically on AI-pedagogical and ethical judgement remain limited, providing a basis for further instrument development without assuming that no related objective measures exist.

To define the construct assessed by the proposed instrument, the study draws on the UNESCO AI Competency Framework for Teachers. The framework organises 15 competencies into five dimensions: a human-centred mindset, ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional learning. Each dimension progresses through three levels labelled *Acquire*, *Deepen*, and *Create* (Miao &

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Cukurova, 2024). The framework was developed partly in response to limited formal provision; as of 2022, only seven countries were reported to have developed national AI competency frameworks or training programmes for teachers (Miao & Cukurova, 2024).

The present study focuses on AI pedagogy and ethics of AI because these dimensions are most directly related to instructional decision-making and the protection of learners. This choice does not imply that the remaining dimensions cannot be assessed objectively. Instead, it defines a deliberately narrower construct. In this study, *AI-pedagogical knowledge* refers to the knowledge required to judge whether an AI-related instructional decision is pedagogically appropriate and ethically defensible within a specified educational context. The assessment draws on competencies described at the *Acquire* and *Deepen* levels; it does not classify respondents at those proficiency levels unless such interpretations are supported by a separate standard-setting procedure.

On this conceptual basis, and consistent with contemporary validity theory, the study examines evidence based on test content, internal structure, item and distractor functioning, score reliability and precision, differential item functioning, and relations to other variables rather than treating validity as a single coefficient or a completed property of the instrument (American Educational Research Association et al., 2014). Three research questions guide the study:

RQ1. What evidence based on test content supports the intended interpretation of scores from the vignette-based assessment of AI-pedagogical knowledge?

RQ2. What evidence from internal structure, item and distractor functioning, score reliability and precision, and differential item functioning across gender and field-of-study groups supports the intended interpretation of the assessment scores?

RQ3. What are the magnitude and direction of the associations between AI-pedagogical knowledge scores and domain-aligned AI self-efficacy, teaching self-efficacy, and self-reported instructional practice among pre-service teachers?

The study makes two intended contributions. Methodologically, it provides initial validity evidence for a context-specific assessment of the pedagogical and ethical knowledge required to evaluate AI-related classroom decisions. Substantively, it estimates the relationships between objectively scored AI-pedagogical knowledge, self-efficacy, and reported instructional practice. These analyses clarify construct relationships without treating vignette responses as direct observations of teaching quality or establishing causal effects of knowledge or self-efficacy on practice.

3. METHODS

3.1 Research design

This study used a multi-phase instrument-development design to develop a vignette-based assessment of AI-pedagogical knowledge and obtain initial evidence supporting its score interpretation. The completed sequence comprised construct definition and blueprint development, preparation of a 36-item pool, developer review and reduction to 24 items, a diagnostic pilot with 60 pre-service teachers, systematic revision of all response options, independent review by three experts, and a main field administration that yielded an analytic sample of 484 pre-service teachers (Figure 1).

The diagnostic pilot was formative and was not treated as validation of the revised assessment. The main study examined evidence based on test content, internal structure, item and distractor functioning, score reliability and conditional precision, differential item functioning, and relations with external variables.

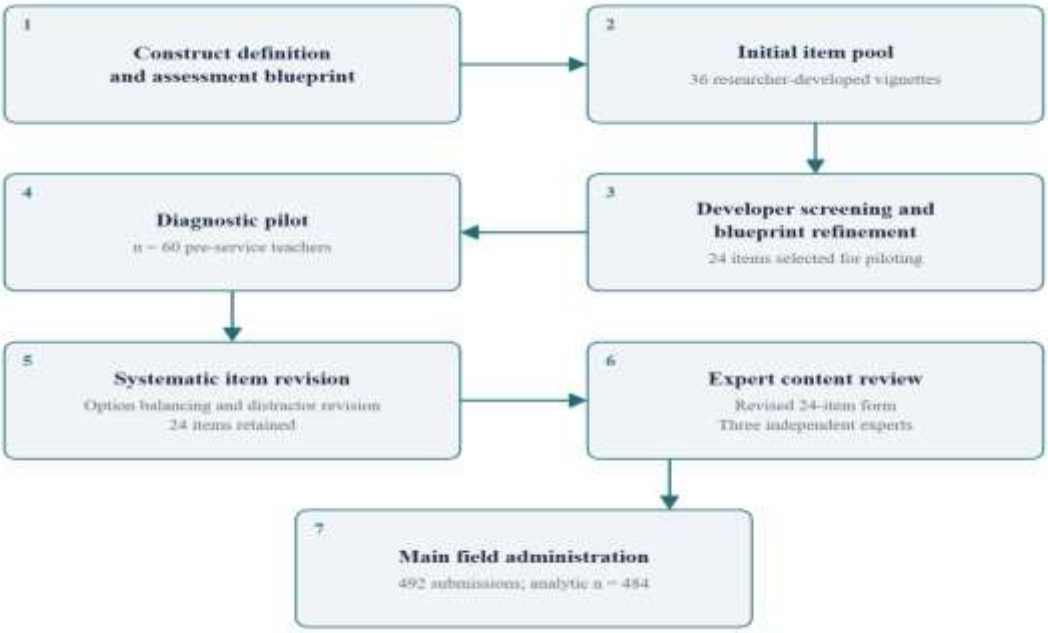


Figure 1. Completed sequence of instrument-development stages

3.2 Construct definition and intended score interpretation

AI-pedagogical knowledge was defined as the knowledge required to evaluate and justify instructional decisions when artificial intelligence is involved in teaching, learning, or assessment. It encompassed four content domains: appraisal of the pedagogical quality of AI-generated output;

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

decisions about when and how AI should be used in instruction; protection of assessment validity and academic integrity; and anticipation of ethical, bias, privacy, and equity risks affecting students.

The construct excluded technical knowledge of AI architectures and algorithms, operational proficiency with particular products, general attitudes towards AI, and self-reported AI self-efficacy. These boundaries guided item writing because dependence on product familiarity or favourable attitudes would introduce variance outside the intended score domain.

The blueprint was informed by the UNESCO AI Competency Framework for Teachers (Miao & Cukurova, 2024), particularly the AI pedagogy and ethics of AI dimensions at the Acquire and Deepen levels. The remaining dimensions and the Create level were outside the deliberately defined scope of the assessment. The UNESCO levels served as design references and were not used to classify respondents into proficiency categories. Interpretation of a total score was made conditional on evidence for a sufficiently coherent general dimension; otherwise, the four domain scores would have been prioritised.

3.3 Instrument blueprint

The 24-item main-study form represented four content domains, with six items allocated to each (Table 1). Every item was linked to one primary indicator recorded in its specification and rationale. Items were presented in a repeating B1-B2-B3-B4 sequence rather than in domain blocks to reduce the possibility that item position would reveal the intended content domain.

Table 1. Blueprint of the 24-item AI-pedagogical knowledge assessment.

Code	Content domain	UNESCO competency block	Level	Items
B1 (EO)	Appraisal of AI output	AI pedagogy - validating and selecting tools	Acquire	6
B2 (ID)	Instructional decision-making	AI pedagogy - integration into strategy	Deepen	6
B3 (AS)	Assessment and integrity	AI pedagogy - learning assessment	Deepen	6
B4 (EB)	Ethics, bias, and student data	Ethics of AI - principled and safe use	Acquire / Deepen	6

3.4 Item development

3.4.1 Initial item pool and reduction

The initial pool of 36 vignette items was drafted by the researcher from the construct definition and assessment blueprint rather than adapted from an existing instrument. Developer review reduced the pool to 24 items before empirical administration. Items were removed when they duplicated content, represented their assigned indicator weakly, depended excessively on technical knowledge or familiarity with a particular AI product, or could not support defensible distractor rationales. The reduction preserved six items in each content domain and was based on construct coverage and item specifications rather than item statistics.

3.4.2 Item format and scoring

Each item comprised a classroom-based vignette, a single stem, and four response options. The assessment used a best-answer format: several options could appear initially plausible, but one was intended to be more defensible after the instructional purpose, available evidence, consequences for students, assessment validity, and ethical considerations were weighed. Responses were scored dichotomously, with 1 assigned to the keyed option and 0 to the other options. Each item was accompanied by a written rationale explaining why the key was preferable and why each distractor represented incomplete or pedagogically problematic reasoning. The keys remained provisional until independent expert review was completed.

3.5 Diagnostic pilot and systematic item revision

The initial 24-item form was administered to 60 pre-service teachers from 14 class groups and six subject specialisations within one provider institution. Item difficulty, corrected point-biserial discrimination, distractor-selection frequencies, total-score distributions, and KR-20 were examined to identify form-level and item-level problems requiring revision.

The pilot identified a pronounced relationship between option length and key status. Option length was therefore treated as a potential construct-irrelevant cue, although the pilot data could not establish whether respondents consciously used a longest-option strategy. All 24 items subsequently underwent option-level revision. Options were rewritten using parallel grammatical structures and comparable visual length; unique wording, absolutes, grammatical mismatches, and unnecessary repetition from the stem were removed. Never-selected or implausible distractors were replaced with alternatives representing recognisable pedagogical errors (Haladyna & Rodriguez, 2013).Key positions were balanced across A, B, C, and D, and references to particular product names or model versions were avoided.

A maximum difference of 15 characters between the longest and shortest options was used as a revision heuristic rather than as a deletion rule. Semantic clarity and substantive accuracy took priority when strict length matching would have obscured meaning. The 15-character rule and the option-length diagnostics were introduced after the pilot and are therefore reported as post hoc safeguards against construct-irrelevant cues.

3.6 Expert content review

Three experts with complementary backgrounds in AI in education or educational technology, pedagogy and pre-service teacher education, and educational measurement independently reviewed the revised 24-item form. Using a five-point scale, they rated each item for alignment with its indicator, clarity, authenticity of the classroom context, accuracy and uniqueness of the keyed response, and plausibility of the distractors. They could also identify alternative interpretations and provide written revision comments.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Aiken's V and 95% confidence intervals were calculated for each item-criterion combination (Aiken, 1985; Penfield & Giacobbi, Jr., 2004). A coefficient below .80 flagged an item for substantive review but did not require automatic deletion. Decisions integrated the quantitative ratings, the uncertainty associated with a three-expert panel, written comments, the defensibility of the key, and consequences for blueprint coverage. The panel results were therefore treated as initial evidence based on test content rather than as conclusive validation.

3.7 Main field administration and participants

The revised survey was administered online in Indonesian to pre-service teachers enrolled at one professional teacher education provider. The sections were presented in the order background information (Part A), AI self-efficacy (Part C), AI-pedagogical knowledge (Part B), general teaching self-efficacy (Part D), and self-reported instructional practice (Part E). Part C preceded Part B to reduce the possibility that experience with the knowledge test would influence self-efficacy judgements. Participants were instructed not to use external sources or AI tools while completing Part B.

The administration produced 492 submissions. Eight records associated with four duplicated student identification numbers were removed because the valid submission within each duplicated pair could not be identified, leaving 484 participants for analysis. The analytic sample was 55.2% female, with a mean age of 24.31 years (SD = 2.18), and represented nine subject specialisations. Physical education was the largest group (46.5%). All participants reported recent practicum, microteaching, or equivalent teaching experience, so Part E was applicable throughout the analytic sample.

3.8 Measures used for relations to other variables

Parts C-E were included to examine relations between objectively scored AI-pedagogical knowledge and theoretically relevant self-reported constructs. These measures were not treated as criterion measures that could establish the validity of Part B by themselves.

3.8.1 AI self-efficacy Part C

Part C was developed specifically for this study following Bandura (2006) guidance for constructing domain-specific self-efficacy scales; it was not adapted from a pre-existing Bandura scale. The 16 items paralleled the four content domains in Part B, with four items assessing confidence in evaluating AI outputs, making instructional decisions, protecting assessment and academic integrity, and addressing ethics, bias, and student data. Responses ranged from 1 (not at all confident) to 6 (completely confident). No items were reverse scored. Because the measure was newly constructed, its factor structure and reliability were evaluated in the present sample.

3.8.2 Teaching self-efficacy Part D

Part D used the 12-item short form of the Teachers' Sense of Efficacy Scale (Tschannen-Moran & Hoy, 2001). The items represent efficacy for student engagement, instructional strategies, and classroom management. Each item was rated from 1 (not capable at all) to 9 (completely capable), with verbal anchors at odd-numbered response categories. The items were administered in Indonesian, and reliability and structure were re-estimated in the present sample rather than inferred from the original study.

3.8.3 Self-reported instructional practice Part E

Part E comprised 16 items covering student support, providing structure, cognitive activation, and classroom management, with four items per domain. Student support, cognitive activation, and classroom management followed the Three Basic Dimensions framework (Praetorius et al., 2018). Providing structure was retained as a separate domain following the self-reported instructional-practice operationalisation used by Depaepe & König, (2018). Participants rated the frequency of each practice from 1 (never) to 5 (almost always) with reference to their most recent practicum, microteaching, internship, or teaching experience during the preceding six months. Scores represented reported practice and were not interpreted as direct observations of teaching quality.

3.9 Data analysis

Data screening. Records were checked for duplicate identifiers, missing responses, coding errors, and invariant response patterns. A uniform response pattern on a self-report scale was treated as a quality flag rather than an automatic exclusion criterion because it could represent a genuine judgement. Sensitivity analyses compared the principal estimates with and without flagged cases.

Evidence based on test content. Aiken's V coefficients, confidence intervals, expert comments, judgements about the key and distractors, and resulting revisions summarised the content review. Direct response-process data were not available and were not inferred from the option-length or distractor analyses.

Item design and classical functioning. For the pilot and main-study forms, option lengths were summarised by key status, the number of items in which the key was the longest option, the score obtainable by always choosing the longest option, and the correlation between total score and the number of longest options selected. Main-study item analyses included proportion correct, corrected point-biserial discrimination, distractor-selection frequencies, distractor point-biserial correlations, and score distributions. KR-20 was used as a descriptive index of internal consistency. Thresholds prompted substantive review rather than automatic item deletion.

Internal structure. Eigenvalues from the tetrachoric correlation matrix and parallel analysis were examined before categorical confirmatory factor analysis. Using the robust weighted least-squares estimator (WLSMV), a one-factor model was compared with the four-correlated-factor model specified by the blueprint. Model fit was evaluated jointly using the chi-square statistic, CFI, TLI, RMSEA with its 90% confidence interval, and SRMR. Standardised loadings and residual correlations were also inspected. A total score was interpreted only if the fitted structure supported a sufficiently coherent general dimension.

Item-response modelling and score precision. The retained one-factor categorical model was expressed on the two-parameter logistic metric to summarise item discrimination, item difficulty, and test information. Rasch estimates were also obtained as secondary diagnostics. The equal-discrimination assumption was evaluated with Andersen's likelihood-ratio test, and Rasch infit, outfit, person separation reliability, and targeting were examined (Andersen, 1973; Mair & Hatzinger, 2007).Score precision was summarised with KR-20, categorical omega, test information, conditional standard errors, conditional reliability, and the distribution of raw scores. Ceiling effects were examined because high overall reliability can coexist with weak precision at the upper end of the score continuum.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Differential item functioning and fairness. DIF was examined by gender and field of study. Because physical education was the only specialisation large enough for a stable focal-group comparison, it was compared with all other specialisations combined. Logistic regression used the rest score as the matching variable and tested uniform and non-uniform DIF jointly (Swaminathan & Rogers, 1990). Holm correction controlled multiplicity, and Nagelkerke change in R-squared was interpreted using the criteria proposed by Jodoin & Gierl, (2001). Andersen's test split by gender was reported as a supplementary global diagnostic. DIF across institutions could not be examined because all respondents came from one institution.

Relations to other variables. Observed Pearson correlations with 95% confidence intervals were calculated between Part B scores and Parts C-E. Spearman correlations and Rasch person estimates were used as sensitivity checks for the association between Parts B and C. The one-factor and four-correlated-factor models for Part C were compared using robust maximum likelihood, and an item-level joint measurement model for Parts B-E was estimated with WLSMV to obtain latent correlations. Correlations between matched and unmatched Part B and Part C domains were compared descriptively. Steiger's test compared dependent correlations that shared one variable (Steiger, 1980). All associations were interpreted as cross-sectional and non-causal.

3.10 Ethical considerations

The diagnostic pilot and main field administration were conducted with institutional ethical approval and informed consent. Participation was voluntary, responses were treated as confidential, and individual scores were not shared with instructors or programme administrators or used in academic evaluation. Identifying information was stored separately from response data and was accessible only to the research team.

4. RESULTS

4.1 Sample and data screening

Of the 492 submitted responses, eight records sharing four student identification numbers were removed. The duplicated submissions differed on 24 to 43 items, so the valid submission could not be identified. The analytic sample therefore comprised 484 pre-service teachers (55.2% female; *M*age = 24.31, *SD* = 2.18). Physical education was the largest specialisation (46.5%), followed by guidance and counselling (14.3%), social studies (8.7%), civic education (8.3%), primary education (6.4%), science (5.4%), special education (4.8%), vocational education (3.3%) and arts (2.3%). Most participants had accumulated between 1 and 40 hours of teaching (80.4%). A fifth (20.9%) had never attended AI training and a further 53.7% had attended fewer than eight hours, yet 47.5% reported using AI daily or almost daily.

There were no missing responses in Parts B to E. A total of 118 respondents (24.4%) gave identical ratings across all items of at least one self-report scale. These cases were retained because uniform ratings can reflect genuine judgements. A sensitivity analysis excluding them left every substantive estimate practically unchanged: KR-20 for Part B was .928 with and .924 without these cases, and the correlation between Part B and AI self-efficacy was .36 and .34, respectively.

4.2 RQ1: Evidence based on test content

Test content

The 24 vignettes were distributed evenly across the four sub-dimensions derived from the UNESCO AI Competency Framework for Teachers (Miao & Cukurova, 2024): evaluating AI outputs (EO), instructional decision-making (ID), assessment and academic integrity (AS), and ethics, bias and student data (EB), with six items in each.

Three experts rated every item independently: one in educational evaluation and psychometrics, one in AI in education, and one in educational technology for teacher education. Each item was rated on five criteria on a five-point scale: alignment with the indicator, clarity of the item, authenticity of the classroom context, correctness and uniqueness of the key, and plausibility of the distractors. Content validity was quantified with Aiken's *V* (Aiken, 1985), and items were retained when every criterion reached $V \geq .80$.

All 120 criterion-level coefficients met the criterion (Table 1). Aiken's *V* ranged from .83 to 1.00 with a mean of .94, and the five criteria were rated almost identically (means .93 to .94). At item level, mean *V* ranged from .90 (B04, B09 and B16) to .97, and no item fell below the threshold on any criterion. The three experts also rated the representativeness of each sub-dimension's six items for its blueprint at the maximum of 5. On this evidence no item was rejected or revised, and the experts recorded no written comments or revision suggestions.

Two features of these ratings limit how much they can support. First, the experts used only the two highest scale points, so the resulting coefficients have a restricted range. Agreement indices were correspondingly uninformative: exact agreement between pairs of experts ranged from .56 to .66, which is close to chance when only two categories are used, and the intraclass correlation was effectively zero because the items showed almost no rated variance. Second, with three raters, the 95% confidence interval for $V = .83$ extends from .55 to .95, so the values for individual items are estimated imprecisely. The panel evidence is therefore best read as agreement that the item set as a whole is appropriate, rather than as a basis for discriminating among items.

Table 1. Content validity of the 24 items: Aiken's V by criterion and sub-dimension (three experts, five-point scale)

Criterion / sub-dimension	Mean *V*	Minimum *V*	Maximum *V*	Items with *V* < .80
Alignment with the indicator	.93	.83	1.00	0
Clarity of the item	.94	.83	1.00	0
Authenticity of the classroom context	.94	.83	1.00	0
Correctness and uniqueness of the key	.93	.83	1.00	0
Plausibility of the distractors	.94	.83	1.00	0
EO: evaluating AI outputs (6 items)	.93	.83	1.00	0
ID: instructional decision-making (6 items)	.95	.83	1.00	0

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Criterion / sub-dimension	Mean *V*	Minimum *V*	Maximum *V*	Items with *V* < .80
AS: assessment and integrity (6 items)	.93	.83	1.00	0
EB: ethics, bias and student data (6 items)	.93	.83	1.00	0
All ratings (24 items × 5 criteria)	.94	.83	1.00	0

Note. Aiken's V = Σ(rating − 1) / [n(c − 1)] with n = 3 raters and c = 5 scale points. Item-level V values are given in Table 3. The ratings used only the scale points 4 and 5, so V = .83 corresponds to two ratings of 5 and one of 4.

4.3 RQ2: Evidence from internal structure, item and distractor functioning, score precision and differential item functioning

Item design and distractor functioning

Two aspects of item design and empirical functioning were examined: the extent to which option features unrelated to the construct could account for scores, and the functioning of the distractors (Table 2).

In the pilot version (1.0, n = 60), the keyed answer was the longest option in 23 of 24 items, and the number of longest options a respondent selected correlated almost perfectly with the total score (r = .998). In the revised version used here, keys and distractors were of near-identical length (85.9 and 84.7 characters on average), and options within an item differed by 5.2 characters on average. The key was the longest option in seven items, close to the six expected by chance. A respondent who always chose the longest option would score 7 out of 24, and the correlation between total score and the number of longest options selected fell to .22. One item (B06) still showed a 30-character gap between its key and its shortest distractor.

All 72 distractors were selected by at least one respondent, and every distractor correlated negatively with the total score (point-biserial r = −.53 to −.02). The distractors therefore attracted lower-scoring respondents, as intended. However, 25 distractors (34.7%) were chosen by fewer than 5% of respondents, and only five items had all three distractors above that threshold. Item B01 was the one clear anomaly. Its distractor stating that the objectives omitted teaching methods and media was chosen by 20.2% of respondents, and their mean total score (18.5) was almost identical to the sample mean (18.9; point-biserial r = −.03). Knowledgeable respondents thus appear to have found this distractor defensible.

Table 2. Indicators of construct-irrelevant option cues and distractor functioning in the pilot and main administrations

Indicator	Pilot (v1.0, *n* = 60)	Main study (v2.0, *n* = 484)
Items in which the key is the longest option	23 of 24	7 of 24
Mean length of key / distractors (characters)	91 / 76	85.9 / 84.7
Score obtained by always choosing the longest option	23	7
Correlation of total score with number of longest options chosen	.998	.22
Items with at least one distractor never chosen	3	0
Distractors chosen by < 5% of respondents	—	25 of 72
Distractors with non-negative point-biserial correlation	—	0 of 72

Note. Pilot values are taken from the pilot analysis reported in the Method section. Dashes indicate indicators not computed for the pilot.

Internal structure

The first eigenvalue of the tetrachoric correlation matrix was 14.71 and the second was 1.49, a ratio of 9.9. Parallel analysis suggested retaining four factors, but confirmatory factor analysis with the WLSMV estimator did not yield a defensible four-dimensional solution. The correlated four-factor model mirroring the content domains was inadmissible, with latent correlations between .98 and 1.04 and a non-positive-definite latent covariance matrix. By contrast, the one-factor model fitted the data well: $\chi^2(252) = 390.33$, CFI = .990, TLI = .989, RMSEA = .034, 90% CI [.027, .040], SRMR = .065 (Table 4). The one-factor model was therefore retained for score interpretation. Standardised loadings ranged from .32 (B01) to .96 (B18), and 22 of 24 items loaded at .55 or above. The largest absolute residual correlation was .24 (B01 with B20).

Item functioning

Classical item statistics are reported in Table 3. Item difficulty (p) ranged from .56 to .86 (M = .79), and corrected item–total correlations ranged from .23 to .78, with none below .20. Item parameters converted from the one-factor solution to the two-parameter logistic (2PL) metric showed discriminations from 0.57 to 5.83 and difficulties from −1.63 to −0.44. All 24 items were therefore located below the mean of the latent distribution. The very high discriminations of B18 and B22 (5.83 and 4.23) should be read with caution, as such values are often inflated when responses cluster near the ceiling.

Andersen's likelihood-ratio test rejected the equal-discrimination assumption of the Rasch model (median split: $\chi^2(20) = 56.65$, p < .001; three items were excluded from this test because of perfect responding within a subgroup). Rasch estimates were therefore treated as secondary diagnostics. Infit mean-squares ranged from 0.61 to 1.66 and outfit mean-squares from 0.51 to 2.51. Three items underfitted, with infit or outfit above 1.30: B01 (1.66 and 2.51), B08 (1.45 and 1.54) and B11 (1.34 and 1.61); B02 sat on the criterion (outfit 1.31). Six items overfitted, with outfit below 0.70 (B14, B17, B18, B19, B22 and B24), which is consistent with their high 2PL discriminations. The mean person location was 2.07 logits above the mean item difficulty.

Score precision

Internal consistency was high: KR-20 = .93, categorical omega = .95, and Rasch person separation reliability = .81 (excluding extreme scores). These summary coefficients, however, conceal highly uneven precision along the latent continuum. Test information peaked at θ = −0.92 and

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

exceeded 5, which corresponds to a reliability of about .80, only between θ = −2.40 and θ = 0.26 (Figure 1b). Conditional reliability fell to .55 at θ = 1.0 and to .21 at θ = 2.0.

The score distribution showed the same pattern (Figure 1a). Of the 484 respondents, 117 (24.2%) obtained the maximum score and 230 (47.5%) scored 22 or above. The distribution also had a secondary cluster: 76 respondents (15.7%) scored 10 or below, close to the chance level of six. When the analysis was restricted to respondents scoring 20 or above (n = 307), KR-20 dropped to .40. The high overall coefficient therefore largely reflects the separation between the low-scoring cluster and the rest of the sample.

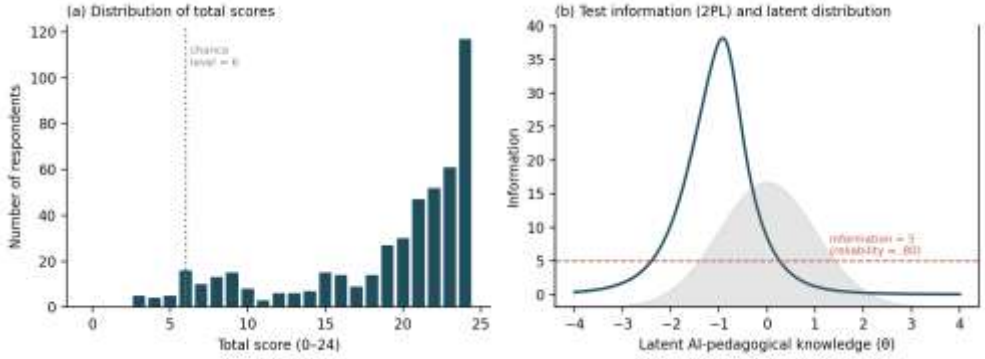


Figure 1. (a) Distribution of Part B total scores (N = 484). (b) Test information function based on 2PL parameters, with the standard normal latent distribution shaded.

Differential item functioning

All participants were enrolled at a single institution, so differential item functioning (DIF) across institutions could not be examined. DIF was instead examined across gender and across specialisation, comparing physical education with all other specialisations, because physical education was the largest single group. Logistic regression (Swaminathan & Rogers, 1990) was used, with the rest score as the matching variable and uniform and non-uniform DIF tested jointly. Effect sizes were evaluated with the Nagelkerke ΔR^2 criteria of Jodoin and Jodoin & Gierl (2001), under which values below .035 are negligible.

For gender, the largest ΔR^2 was .030 (B20), and no item remained significant after Holm correction. For specialisation, the largest ΔR^2 was .015 (B07), again with no significant items. The Andersen test split by gender was significant ($\chi^2(23) = 40.55$, p = .013). Because no single item showed DIF of non-negligible size, this result more plausibly reflects the variable discrimination already identified than item-specific bias.

Table 3. Item statistics for Part B (N = 484)

Item	Dim.	V	p	$r$it	λ	a	b	δ	Infit	Outfit	ΔR^2 G	ΔR^2 S
B01	EO	0.93	0.56	0.23	0.32	0.57	-0.44	1.86	1.66	2.51	0.011	0.012
B02	ID	0.95	0.85	0.44	0.64	1.41	-1.63	-0.62	1.17	1.31	0.016	0.009
B03	AS	0.93	0.83	0.61	0.80	2.30	-1.17	-0.34	0.89	0.82	0.002	0.003
B04	EB	0.90	0.69	0.59	0.77	2.06	-0.66	0.86	1.03	1.01	0.018	0.007
B05	EO	0.92	0.83	0.56	0.76	1.98	-1.25	-0.37	0.99	0.78	0.009	0.012
B06	ID	0.93	0.83	0.44	0.63	1.37	-1.55	-0.43	1.20	1.19	0.020	0.003
B07	AS	0.93	0.73	0.64	0.82	2.41	-0.76	0.55	0.93	0.90	0.011	0.015
B08	EB	0.95	0.72	0.39	0.54	1.09	-1.07	0.67	1.45	1.54	0.003	0.000
B09	EO	0.90	0.83	0.54	0.74	1.88	-1.30	-0.41	1.03	1.00	0.005	0.005
B10	ID	0.97	0.66	0.67	0.85	2.77	-0.49	1.11	0.87	0.81	0.000	0.003
B11	AS	0.93	0.64	0.40	0.55	1.13	-0.64	1.28	1.34	1.61	0.004	0.002
B12	EB	0.95	0.82	0.67	0.87	2.97	-1.06	-0.28	0.80	0.70	0.002	0.005
B13	EO	0.95	0.76	0.68	0.86	2.87	-0.82	0.30	0.83	0.88	0.004	0.011
B14	ID	0.95	0.86	0.60	0.82	2.48	-1.33	-0.77	0.85	0.56	0.005	0.002
B15	AS	0.92	0.78	0.56	0.74	1.85	-1.06	0.09	1.06	0.88	0.000	0.004
B16	EB	0.90	0.84	0.55	0.77	2.06	-1.28	-0.48	1.00	0.90	0.001	0.005
B17	EO	0.93	0.85	0.68	0.90	3.56	-1.13	-0.57	0.74	0.61	0.016	0.005
B18	ID	0.97	0.79	0.78	0.96	5.83	-0.84	0.03	0.61	0.51	0.001	0.002
B19	AS	0.93	0.84	0.64	0.85	2.73	-1.17	-0.48	0.82	0.56	0.003	0.012
B20	EB	0.95	0.82	0.58	0.80	2.31	-1.15	-0.30	0.96	0.92	0.030	0.005
B21	EO	0.97	0.82	0.58	0.78	2.13	-1.17	-0.28	0.97	0.77	0.005	0.000

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Item	Dim.	*V*	*p*	*r*it	λ	*a*	*b*	δ	Infit	Outfit	ΔR^2 G	ΔR^2 S
B22	ID	0.95	0.79	0.74	0.93	4.23	-0.87	0.01	0.70	0.65	0.003	0.002
B23	AS	0.93	0.86	0.60	0.84	2.69	-1.30	-0.77	0.84	0.75	0.004	0.003
B24	EB	0.95	0.85	0.66	0.89	3.40	-1.19	-0.67	0.75	0.52	0.001	0.003

Note. Dim. = sub-dimension (EO = evaluating AI outputs; ID = instructional decision-making; AS = assessment and integrity; EB = ethics, bias and student data). *V* = mean Aiken's *V* across the five expert criteria; *p* = proportion correct; *r*it = corrected item–total correlation; λ = standardised loading, one-factor WLSMV model; *a*, *b* = 2PL discrimination and difficulty converted from the factor solution ($a = 1.702\lambda/\sqrt{(1-\lambda^2)}$); δ = Rasch item difficulty (logits); Infit/Outfit = Rasch mean-square fit; ΔR^2 = Nagelkerke effect size for logistic-regression DIF by gender (G) and specialisation (S).

Table 4. Fit of the measurement models

Model	χ^2 (*df*)	CFI	TLI	RMSEA [90% CI]	SRMR
Part B, one factor (WLSMV)	390.33 (252)	.990	.989	.034 [.027, .040]	.065
Part B, four correlated factors (WLSMV)[a]	380.37 (246)	.990	.989	.034 [.027, .040]	.064
Part C, one factor (MLR)	463.68 (104)	.863	.842	.113	.070
Part C, four correlated factors (MLR)	286.26 (98)	.931	.915	.083	.058
Joint model: B, C, D, E (WLSMV)	2879.10 (2204)	.932	.930	.025 [.023, .028]	.056

Note. [a] Inadmissible solution: latent correlations between .98 and 1.04. Scaled χ^2 and robust fit indices are reported; the Part C models were estimated with MLR because the response scale has six points.

4.4 RQ3: Relations with AI self-efficacy, teaching self-efficacy and self-reported instructional practice

AI-pedagogical knowledge was positively associated with all three self-report measures (Table 5). Its association with domain-aligned AI self-efficacy was moderate: the observed correlation was $r = .36$, 95% CI [.28, .43], with Spearman's $\rho = .27$ and $r = .31$ [.22, .38] when Rasch person estimates replaced total scores. In the joint measurement model, the latent correlation was .42, 95% CI [.33, .50]. The two constructs thus shared about 13% of their variance at the observed level and 17% at the latent level.

The knowledge test related to general teaching self-efficacy to a similar degree (latent $r = .37$) and more weakly to self-reported instructional practice (.25). The three self-report measures, by contrast, were more strongly intercorrelated (latent $r = .52$ to .60). The observed correlation between knowledge and AI self-efficacy was significantly smaller than that between AI self-efficacy and general teaching self-efficacy (Steiger's test, $t = -4.84$, $p < .001$).

Domain alignment did not produce stronger associations. Spearman correlations between the four Part B sub-scores and the four Part C subscales ranged from .20 to .31. The mean of the four matched pairs (.24) was the same as the mean of the twelve unmatched pairs (.24). In Part C itself, the four-factor model improved on the one-factor model, although absolute fit was mixed (CFI = .931, TLI = .915, RMSEA = .083, SRMR = .058) and the subscales remained highly intercorrelated (.72 to .88).

Table 5. Correlations among knowledge, self-efficacy and practice

	1	2	3	4
1. AI-pedagogical knowledge (Part B)	—	.42 [.33, .50]	.37 [.29, .45]	.25 [.16, .35]
2. AI self-efficacy (Part C)	.36 [.28, .43]	—	.60 [.54, .66]	.52 [.46, .59]
3. Teaching self-efficacy (Part D)	.33 [.25, .41]	.57 [.50, .62]	—	.58 [.53, .64]
4. Instructional practice (Part E)	.22 [.14, .31]	.48 [.41, .55]	.55 [.49, .61]	—

Note. Observed Pearson correlations with 95% CIs below the diagonal; latent correlations from the joint WLSMV model with 95% CIs above the diagonal. All $p < .001$. Scale means: Part C = 4.33 (1–6), Part D = 6.78 (1–9), Part E = 4.23 (1–5).

5. DISCUSSION

This study examined validity evidence for a vignette-based assessment of AI-pedagogical knowledge and used the resulting scores to estimate their associations with AI self-efficacy, teaching self-efficacy and self-reported instructional practice. Four findings stand out. First, revising the options substantially reduced the length cue that had dominated the pilot version. Second, responses in this sample were adequately represented by a general factor rather than empirically separable content domains. Third, score precision was concentrated in the lower half of the ability distribution. Fourth, knowledge was moderately associated with AI self-efficacy and teaching self-efficacy and more weakly associated with self-reported instructional practice, supporting the distinction between objectively scored knowledge and self-reported competence.

The contrast between the two versions shows that systematic revision substantially reduced the potential influence of option length. In the pilot, selecting the longest option would have yielded almost the same score as answering the items correctly. In the main study, the same strategy would have produced 7 of 24 points, below the sample's lower quartile. Together with the distractor results, this suggests that the revised options functioned more appropriately, although these analyses do not reveal how respondents interpreted or reasoned through the items (Haladyna & Rodriguez, 2013).

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Expert judgement and empirical functioning did not fully converge. Every item cleared the Aiken's *V* criterion comfortably, including items that later underfitted the measurement model, and the panel proposed no revisions. Because the experts used only the top two scale points, the ratings could not separate stronger from weaker items. This is a well-known constraint of rating-based content validation and an argument for treating panel review as a screen for serious content problems rather than as evidence of item quality. Reviewing items against pilot data, and asking experts to rank or compare items rather than only to rate them, would make the panel stage more informative.

The distractor analysis is more mixed. Every distractor attracted lower-scoring respondents, but a third were rarely chosen, which reduces many items to an effective choice between two or three options. B01 needs particular attention. It was both the hardest item and the most poorly fitting, and one of its distractors appealed to respondents across the ability range. One plausible explanation is that the widely taught ABCD format for learning objectives includes a condition component, which respondents may equate with methods and media. The key and this distractor should therefore be reviewed by the expert panel before the item is used again.

The four UNESCO-derived content domains were not empirically distinguishable in this sample: their latent correlations approached or exceeded 1. Responses were adequately represented by a general factor, supporting use of the total score and advising against interpretation of domain sub-scores at this stage. This result does not invalidate the framework. The framework is normative, describing what teachers should be able to do (Miao & Cukurova, 2024), rather than an empirically established factor structure. One possible interpretation is that AI-pedagogical knowledge at the pre-service stage is not yet differentiated into distinct areas. Instead, performance may reflect a general capacity to scrutinise AI outputs, protect students' thinking and treat AI-generated evidence cautiously across the four content domains.

The main limitation of the current version is its targeting. The KR-20 of .93 indicates high overall internal consistency, but the conditional analysis shows that precision is concentrated below the latent mean. Nearly a quarter of respondents answered every item correctly, and the test provided little differentiation across roughly the upper 40% of the latent distribution. The current form may therefore provide useful formative information about respondents at the lower end of the score distribution, but it is not suitable for ranking or certifying more knowledgeable candidates. Extending the scale upwards will require more difficult items at the deeper level of the framework, with more closely competing response options.

The unequal discriminations and rejection of the Rasch model favour a 2PL scoring model for future calibration. No item showed non-negligible DIF by gender or by the field-of-study comparison used here. This provides no indication of large item-specific bias in these comparisons, but it does not by itself establish full score comparability across groups. Moreover, because the study involved only one institution, comparability across institutions could not be examined and remains an open question.

The moderate correlation between objectively scored knowledge and domain-aligned self-efficacy (r = .36; latent .42) supports treating the two as distinct constructs. Self-efficacy shared considerably more variance with other self-reports than with the knowledge test, which suggests that common method variance contributes substantially to associations among self-report measures of teacher competence. This pattern is consistent with Depaepe & König (2018), who found little relationship between pre-service teachers' general pedagogical knowledge and their self-efficacy. The present study extends that finding to the AI domain. The practical implication is that self-report measures of AI competence, which dominate the current literature, should not be taken as proxies for what pre-service teachers actually know about teaching with AI.

Aligning the content of the two instruments did not strengthen the association: matched and unmatched sub-dimensions correlated equally. Together with the unidimensional structure of the test, this suggests that the link between knowledge and self-efficacy operates at a general level. It does not reflect domain-specific calibration in which, for example, knowledge about assessment informs confidence about assessment in particular.

The weaker association with self-reported instructional practice (latent r = .25) indicates that vignette performance and reported teaching behaviour overlap only modestly. Because instructional practice was self-reported and measured concurrently, the association should not be interpreted as evidence that greater AI-pedagogical knowledge causes better teaching practice. It instead provides limited relational evidence that the knowledge scores are connected to, but not interchangeable with, reported practice.

6. CONCLUSION

This study developed and provided initial validity evidence for a 24-item vignette-based assessment of AI-pedagogical knowledge among pre-service teachers. Content review by three experts supported the relevance and appropriateness of the item set, with Aiken's V values ranging from .83 to 1.00. These estimates should nevertheless be interpreted cautiously because the panel was small and the experts used only the two highest rating categories. The revision process substantially reduced the option-length cue identified during the diagnostic pilot, and all distractors functioned in the intended direction, although several were rarely selected and B01 contained a distractor that remained defensible across the ability range.

The internal-structure results supported interpretation of a single total score rather than separate scores for the four content domains. Overall internal consistency was high, with KR-20 of .93 and categorical omega of .95. However, these coefficients concealed uneven conditional precision. Nearly one quarter of participants obtained the maximum score, and almost half scored 22 or higher, indicating that the assessment differentiated respondents more effectively at lower levels of AI-pedagogical knowledge than at the upper end. No item showed non-negligible differential functioning by gender or the field-of-study comparison examined, but the single-institution sample does not establish score comparability across institutions.

The current form may therefore be useful for formative assessment and research when results are reported as a total score. It should not yet be used for ranking, screening, proficiency classification, or high-stakes certification. Further development should introduce more difficult vignettes, strengthen weak distractors, and reconsider B01, B08, and B11. Cognitive interviews are also needed to obtain direct response-process evidence, followed by evaluation of the revised form across multiple institutions. These steps would strengthen the interpretation, precision, and generalisability of scores while preserving the assessment's focus on pedagogical and ethical judgement in AI-related classroom situations.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

7. LIMITATIONS

This study has two main limitations. All participants were drawn from a single institution, and the sample distribution across fields of study reflected the programmes operating during data collection, with physical education constituting the largest group; therefore, generalisability to other institutional settings remains to be examined. In addition, cognitive interviews were not conducted, so the study provides no direct evidence about how respondents interpreted the vignettes or reasoned through the response options. The item-design and distractor analyses should therefore not be interpreted as direct response-process evidence.

REFERENCES

1) Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012
2) American Educational Research Association, A. E. R. A., American Psychological Association, A. P. A., & National Council on Measurement in Education, N. C. on M. in E. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
3) Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch Model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180
4) Bandura, A. (2006). Guide to the construction of self-efficacy scales. In F. Pajares & T. Urdan (Ed.), *Self-efficacy beliefs of adolescents* (Vol. 5). Information Age Publishing.
5) Chiu, T. K. F., Ahmad, Z., & Çoban, M. (2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, *30*(5), 6667–6685. https://doi.org/10.1007/s10639-024-13094-z
6) Choudhury, S., Deb, J. P., Pradhan, P., & Mishra, A. (2024). Validation of the Teachers AI-TPACK Scale for the Indian Educational Setting. *International Journal of Experimental Research and Review*, *43*, 119–133. https://doi.org/10.52756/ijerr.2024.v43spl.009
7) Depaepe, F., & König, J. (2018). General pedagogical knowledge, self-efficacy and instructional practice: Disentangling their relationship in pre-service teacher education. *Teaching and Teacher Education*, *69*, 177–190. https://doi.org/10.1016/j.tate.2017.10.003
8) Doss, C. J., Bozick, R., Schwartz, H. L., Chu, L., Rainey, L. R., Woo, A., Reich, J., & Dukes, J. (2025). *AI Use in Schools Is Quickly Increasing but Guidance Lags Behind: Findings from the RAND Survey Panels*. RAND Corporation. https://doi.org/10.7249/RRA4180-1
9) Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. In *Developing and Validating Test Items*. https://doi.org/10.4324/9780203850381
10) Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence*, *9*, 100436. https://doi.org/10.1016/j.caeai.2025.100436
11) Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2
12) Kerr, R. C., & Kim, H. (2025). From Prompts to Plans: A Case Study of Pre-Service EFL Teachers' Use of Generative AI for Lesson Planning. *English Teaching*, *80*(1), 95–118. https://doi.org/10.15858/engtea.80.1.202503.95
13) Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779
14) Lievens, F., & Motowidlo, S. J. (2016). Situational Judgment Tests: From Measures of Situational Judgment to Measures of General Domain Knowledge. *Industrial and Organizational Psychology*, *9*(1), 3–22. https://doi.org/10.1017/iop.2015.71
15) Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9). https://doi.org/10.18637/jss.v020.i09
16) Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI competency objective scale (AICOS). *Computers and Education: Artificial Intelligence*, *9*, 100485. https://doi.org/10.1016/j.caeai.2025.100485
17) Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084
18) Ning, Y., Zhang, C., Xu, B., Zhou, Y., & Wijaya, T. T. (2024). Teachers' AI-TPACK: Exploring the Relationship between Knowledge Elements. *Sustainability*, *16*(3), 978. https://doi.org/10.3390/su16030978
19) Penfield, R. D., & Giacobbi, Jr., P. R. (2004). Applying a Score Confidence Interval to Aiken's Item Content-Relevance Index. *Measurement in Physical Education and Exercise Science*, *8*(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3
20) Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879
21) Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM – Mathematics Education*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4
22) Setiyawan, A., Soeharto, S., Wijaya, T. T., Korenova, L., & Lavicza, Z. (2025). Measuring Teachers' competencies for AI integration: Development and validation of the AI-TPACK in vocational education. *Computers and Education Open*, *9*, 100319. https://doi.org/10.1016/j.caeo.2025.100319
23) Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245
24) Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

25) Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1
26) Zhang, S., Xiao, R., Botelho, A. F., Liao, G., Chiu, T. K. F., Stamper, J., & Koedinger, K. R. (2026). How to Assess AI Literacy: Misalignment Between Self-Reported and Objective-Based Measures. *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, 405–414. https://doi.org/10.1145/3785022.3785088

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Abdul Hafid