JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Early Access Articles, 1–20. https://doi.org/10.18608/jla.2026.9133

Learning-Aware Reliability Estimation for Tutor Skill Assessment Using Large Language Models

Conrad Borchers[1], Danielle R. Thomas[2], Jionghao Lin[3], Kenneth R. Koedinger[4]

Abstract

Assessment is foundational to learning analytics, especially in evaluating instructional interventions and guiding improvement in online learning environments. With the growing use of large language models (LLMs) to score open-ended responses, questions arise about the reliability of these model-generated scores, particularly in short pre-post formats where learners are expected to improve. This study introduces a novel method for estimating test reliability that adjusts for learning gains using a Rasch-based split-half approach. We validated this approach through simulations under realistic conditions of missing data and score changes, demonstrating tangible improvements in reliability estimation compared to baseline methods. Applying this method to a dataset of 985 tutors completing 12 online lessons, we find that GPT-4–based scoring achieves satisfactory reliability, with open-ended responses (0.733) outperforming multiple-choice items (0.652). Both item types jointly yielded the highest reliability (0.774). Hence, as few as 14 open-ended items (across an average of 3 to 4 completed lessons) were sufficient to surpass common reliability thresholds of 0.7 or higher. Principal component analysis revealed a skill structure with a strong primary dimension shared across almost all lessons and interpretable subdimensions—socio-emotional, cognitive, and fairness-related tutoring skills—supporting a bifactor-like model. These findings demonstrate that GPT-4 and similar LLMs can be effectively used for formative assessment of complex instructional skills in online and personalized learning contexts, provided their reliability is empirically verified. This study contributes an open-source, learning-aware framework for scalable and reliable AI-supported assessment in learning analytics contexts.

Notes for Practice

- Results suggest using at least 14 open-ended assessment items across lessons to achieve reliable estimates ($\geq$ 0.7) of tutor skill in short-form online training.
- Open-ended responses, when scored with GPT-4, yield more reliable assessments of tutoring skills than multiple-choice questions.
- Tutoring skills exhibit a primarily unidimensional structure, with interpretable subdimensions such as cognitive, socio-emotional, and pedagogical domains marginally improving model fit.
- Large language model (LLM)-based scoring is reliable for assessing tutor skill, but pre-post formats where learning gains occur and personalized learning with missing student-item combinations require novel reliability estimation methods.
- The proposed split-half method based on adjusted Rasch models can help assess reliability even in contexts with few items, missing data, or learning gains.

Keywords

Reliability, large language models, assessment, short answer grading, tutor training, online learning.

Submitted: 13/06/2025 — **Accepted:** 23/02/2026 — **Published:** 18/05/2026

[1] *Email: cborchers@cmu.edu Address: Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. ORCID iD: https://orcid.org/0000-0003-3437-8979*

[2] *Email: drthomas@cmu.edu Address: Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. ORCID iD: https://orcid.org/0000-0001-8196-3252*

[3] *Email: jionghao@cmu.edu, jionghao@hku.hk Address: Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA; The University of Hong Kong, Pok Fu Lam Road, Hong Kong SAR, China. ORCID iD: https://orcid.org/0000-0003-3320-3907*

[4] *Email: koedinger@cmu.edu Address: Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. ORCID iD: https://orcid.org/0000-0002-5850-4768*

1. Introduction

Assessment of learning and performance is a cornerstone of learning analytics, providing essential insights into student understanding and informing instructional decisions. In the context of tutoring systems and online training platforms, assessments typically include both multiple-choice and open-ended items, aiming to capture a broad range of skills from factual recall to higher-order reasoning (Gurung et al., 2024). With the growing integration of artificial intelligence (AI) in education, large language models (LLMs) have become increasingly used to automate the grading of open-ended responses. These models have demonstrated promising consistency when benchmarked against human grading, often achieving high agreement levels using classification metrics such as accuracy and F1-score (D. R. Thomas et al., 2025b; Lin et al., 2024). However, consistency alone does not guarantee a trustworthy assessment system. A critical yet often overlooked property is *reliability*—the consistency and stability of scores across repeated or highly similar measurements. For example, a model might agree with human scores on average yet produce inconsistent results across the same or highly similar items or across pre- and post-tests. These inconsistencies can arise from at least two distinct sources. One source concerns the psychometric properties of the assessment design, including short test length, missing responses, and student learning between assessments. Another source is variability in the scoring process itself, such as stochastic variation in LLM outputs. For instance, LLMs are inconsistent when prompted multiple times with the same input due to their stochastic nature (L. Wang et al., 2024; Jang & Lukasiewicz, 2023; Y. Liu et al., 2023) or might be overly responsive to phrasing and wording of highly similar items (Y. Liu et al., 2023). These inconsistencies can have significant implications for educational interventions, where assessments serve as the basis for instructional feedback, evaluation of learning gains, and program efficacy in learning analytics.

This study makes three primary contributions to address this gap. First, we introduce a novel method for estimating test reliability in short pre-post assessments that explicitly adjusts for learning gains. Conventional reliability estimates, such as Cronbach's α or split-half reliability, implicitly assume static trait levels (DeVellis, 2006). These assumptions are often violated in pre-post studies, where genuine learning is expected to occur between assessments. Another assumption typically violated in personalized learning is that students traverse only a small subset of items, leading to substantial missing data (L. Zhang et al., 2024). By accounting for these common violations, our method provides a more accurate and interpretable estimate of score reliability under educational interventions with pre-post assessment—a common scenario in learning analytics research. Past approaches used in learning analytics either focus on modelling growth (e.g., additive factors model (AFM), performance factors analysis (PFA)) or define reliability through classical internal consistency measures (e.g., Cronbach's α) for a single assessment occasion. In contrast, our contribution is a Rasch-based split-half *reliability estimator* that produces a test-level reliability coefficient for short pre-post assessments by measuring the consistency of person ability estimates while explicitly accounting for learning. Classical internal-consistency estimates (e.g., Cronbach's α and standard split-half) assume a stable trait. In short pre-post assessments, real learning between measurements violates this assumption so that classical reliability can misattribute true change to measurement error.

Second, we empirically evaluate the reliability of LLM-based scoring across multiple short training assessments in the context of scenario-based tutoring. These assessments, designed to train tutors on key socio-emotional and pedagogical skills, include both multiple-choice and open-ended questions. While many studies have shown that LLMs can approximate human grading in these contexts (Borchers et al., 2025; Grévisse, 2024), their reliability, specifically how consistently they produce scores across assessment instances, has not been systematically examined. Given the stochastic and often hard to interpret nature of these models, scoring the same response twice may yield divergent outputs (Y. Liu et al., 2023). This variability undermines confidence in automated assessment systems unless carefully measured and controlled.

Third, as a secondary contribution, we examine the dimensionality of tutoring skills using factor analytic approaches. There is little empirical work on whether tutoring skills—such as giving praise, supporting a growth mindset, or guiding emotion regulation—reflect a unified construct or represent multiple distinct competencies. For example, training on praise and mindset may appear separately in a professional development curriculum. Still, it remains unclear whether tutors who succeed in one area will necessarily succeed in another. Understanding whether these skills generalize or are domain specific has direct implications for assessment design: a unidimensional structure would allow for streamlined assessments. In contrast, a multidimensional structure would necessitate a more comprehensive and granular evaluation. Our findings provide evidence for the degree to which tutor performance can be explained by general versus specific skill dimensions. Our analysis relates to reliability assessment, as it can serve as a blueprint for other learning analytics researchers seeking to use LLMs for assessment in domains where the factor-analytic structures of skills are unknown. To this end, we open-source all our analysis code and LLM prompts, and all study data will be made available upon request[1].

In sum, this study addresses the following research questions:

- **RQ1:** How can test reliability be estimated for short pre-post assessments while adjusting for learning gains?

[1] https://github.com/conradborchers/tutor-assessment-reliability

- **RQ2:** How reliable is LLM-based scoring of tutor skills with GPT-4 across multiple short assessments?
- **RQ3:** What is the underlying dimensional structure of tutoring skills?

2. Related Work

2.1 Reliability in Learning Assessment and Analytics

Reliability is core to learning analytics, ensuring that assessments consistently and accurately measure meaningful constructs. *Reliability* refers to the consistency of assessment scores—whether similar results are obtained across repeated test items, measurements, or raters (Nunnally & Bernstein, 1994). *Validity*, on the other hand, concerns the extent to which an assessment measures what it is intended to measure (Nunnally & Bernstein, 1994). Validity is beyond the scope of the present study, but reliability is generally considered a prerequisite to validity (Bannigan & Watson, 2009). While both properties are a key objective of assessment in learning analytics (Scheffel et al., 2017; Divjak et al., 2023), the reliability of tutor assessments—particularly in online training contexts—remains underexplored.

Reliability has long been emphasized in automated scoring and natural language processing (NLP)-based assessment. For example, O. L. Liu and colleagues (2014) reviewed both the promise and the constraints of automatically scoring constructed-response science items, and they highlighted the fact that consistency becomes harder to maintain when rubrics require multi-part reasoning and richer conceptual explanations. In language assessment, reliability considerations extend to spoken responses. Xi and colleagues (2008) documented the development and evaluation of SpeechRater for scoring spontaneous speech in a practice assessment setting, and they explicitly linked system development and evaluation to the speaking construct, the intended context of use, and empirical performance relative to human scores. More recently, hybrid human-plus-automated pipelines have been proposed to improve dependability by automatically detecting responses that are difficult to score (e.g., due to high noise, empty responses, or system failures) and routing those cases to human raters, while leaving the majority of responses to be scored automatically (Yoon & Zechner, 2017). This quality-control perspective is reflected in later SpeechRater reporting that details filtering models for nonscorable responses and other updates intended to stabilize automated scoring in practice (Chen et al., 2018). Although this body of work has established reliability practices for many constructed-response tasks, the reliability of tutor assessments, particularly in online training contexts, remains underexplored.

Past research in learning analytics has typically evaluated LLM scoring by comparing it to human-graded responses, using metrics such as accuracy, precision, F1, or Cohen's κ to assess consistency (D. R. Thomas et al., 2025b; Dai et al., 2024; Cao et al., 2025). This is different from reliability in that these metrics measure agreement with a reference (i.e., human) standard rather than the internal stability of the scores themselves across repeated or parallel assessments. While comparisons of LLM and human scoring are important for establishing alignment with human judgment, they often overlook reliability—whether the model (e.g., an LLM) produces *consistent scores* under similar conditions. Without this consideration, scores may vary unpredictably, limiting their usefulness for tracking progress or making instructional decisions over time and at scale.

Without a reliable assessment of open-ended and increasingly automatically scored open-ended responses (Dai et al., 2024), feedback to tutors may be inconsistent, undermining the effectiveness of training (D. R. Thomas et al., 2025b). Traditional methods, such as inter-rater and test-retest reliability (Gwet, 2001), can estimate assessment reliability but are not suited for contexts where tutors learn between assessments (such as practice modules in online courses), as they assume that students do not improve (DeVellis, 2006). This study addresses these gaps by using factor analysis and novel reliability estimation methods on pre-post learning-gain data to evaluate tutor assessments scored by LLMs.

A final contribution of this work lies in applying item response theory (IRT) to the estimation of reliability in short-form assessments where learning occurs between measurement occasions. While IRT-based models such as the AFM (Cen et al., 2006) and PFA (Pavlik et al., 2009) are widely used in learning analytics to model learner growth over time, they are primarily used to estimate skill acquisition parameters rather than to derive test-level reliability estimates. Further, they require a skill model that maps items to skills, which is not always available in short-form assessments. In classical IRT, reliability is typically characterized through test information functions or marginal reliability computed for a single assessment occasion, and these quantities are not defined for pre-post designs in which the latent ability is expected to change between measurements (De Ayala, 2013). The present study differs in that it uses a Rasch-based split-half procedure to estimate reliability from the consistency of person ability estimates across balanced item subsets, while explicitly modelling time effects to account for learning. This allows reliability to be estimated in short assessments with missing data and genuine learning gains, where standard α-based and single-occasion IRT reliability summaries are not directly applicable.

2.2 Short-Form Assessments Including LLMs

Short-form assessments are widely used in learning analytics for their efficiency and scalability, particularly in settings where repeated measurement is needed to track change over time. A key application area of such assessments is to study instructional design, whose aim is to improve student learning outcomes (Mangaroska & Giannakos, 2018). These instruments, often

composed of a few multiple-choice or open-ended items, are especially common in online training programs that require lightweight diagnostics across many sessions (D. R. Thomas et al., 2025b). While multiple-choice items make grading easier, they often fail to capture the complexity of instructional practice. Open-ended prompts, by contrast, might probe deeper pedagogical understanding but require more nuanced evaluation (Gurung et al., 2024).

To address the evaluation challenges posed by open-ended items, recent work has explored using LLMs to score open-ended responses. Studies have shown that LLMs can approximate human ratings in domains such as essay evaluation and instructional feedback classification (Dai et al., 2024; Cao et al., 2025; Seßler et al., 2025), offering a scalable alternative for assessing complex tutor behaviours. For instance, LLMs have been used to classify types of praise in tutoring interactions with high alignment to expert judgments (Lin et al., 2025).

Despite these advances, key questions remain about the reliability of LLM-based scoring when used in short-form assessments administered over time. Past research has focused on whether GPT-generated *questions* are reliable when administered to learners (Bhandari et al., 2024), not whether GPT-generated *scores* are reliable. Where LLMs are used for scoring, a common use case in learning analytics, reliability is typically not evaluated (Yan et al., 2024; Misiejuk et al., 2025). First, LLMs and their hyperparameters (e.g., temperature), stochastic inconsistency, and occasional hallucination are all known to impact open-response grading accuracy (Borchers et al., 2025). Second, in many learning analytics contexts, data consist of a limited number of items, which could amplify the impact of individual scoring inconsistencies. Repeated measures also require that scores remain stable across similar conditions, which is not a tenable assumption, as students often learn between pre- and post-assessments. Taken together, there are many reasons to believe that the reliability of LLM assessments might not be satisfactory in learning analytics contexts, but methods and practices to assess reliability are lacking in the field.

This study directly investigates these issues by analyzing how reliably LLMs score tutor responses across repeated short assessments focused on pedagogical and socio-emotional competencies around an intervention in which they learn. Our findings provide evidence of LLMs' suitability for formative assessment in educational training, particularly when consistency over time is essential. We further provide a reproducible method to assess reliability across pre- and post-measurements.

2.3 Tutoring Skills

While numerous tutor training programs target socio-emotional and pedagogical competencies (Chine et al., 2022; National Student Support Accelerator, 2023), little empirical work has examined whether these skills represent a unidimensional or multidimensional construct. For instance, training modules on praise and growth mindset are often treated as discrete topics, yet it remains unclear whether success in one predicts success in the other (Chhabra et al., 2022). This question is critical for assessment design: if tutoring skills are largely unidimensional, shorter assessments may suffice; if they are multidimensional, broader evaluation batteries are required.

Several established tutoring models provide structured competency frameworks for tutor development. The National Student Support Accelerator (NSSA) developed the Tutoring Quality Improvement System, which provides research-based quality standards for high-impact tutoring programs aligned with their Framework for High-Impact Tutoring (National Student Support Accelerator, 2025). Similarly, the International Tutor Training Program Certification (ITTPC) of the College Reading and Learning Association (CRLA) organizes competencies into three tiered levels, focusing on topics including session management, active listening and responding, and tutoring ethics (O'Neil & Schotka, 2020). Saga Education's Saga Coach training program emphasizes three core dimensions of effective tutoring: Relationships, Ratio, and Rigor, delivered through 19 self-paced training modules (Saga Coach, 2025; Saga Education, 2021). Despite the variety of tutoring competency frameworks published by practitioners, the scientific evaluation of dimensionality across tutoring competencies has been underexplored, a contribution of the present study.

Previous research has shown that scenario-based training can improve tutor performance on specific skills (D. Thomas et al., 2023), but it does not address whether those improvements generalize across competencies. Moreover, the choice of assessment format—multiple-choice versus open-ended—may differentially capture underlying skill structures (Butler, 2018). This study contributes to the literature by applying factor analysis to LLM-graded responses across diverse tutor-training assessments, providing evidence on the dimensionality of tutoring skills in online training contexts.

3. Data and Study Context

3.1 Participants and Sample

A total of 985 college-student tutors at a private university in the northeastern United States, accounting for 3,467 lesson completions, completed at least one of 12 tutor training lessons (see Section 3.2) between the 2022/23 and 2024/2025 academic years. On average, participants completed $M = 3.52$ lessons (SD $= 2.41$). As not all students completed all lessons, we systematically investigate the impact of missing data on our reliability estimation procedure in Section 4.1.1.

Delivered through an online tutoring platform, the 12 lessons align with key tutoring competencies, as perceived by professional educators, spanning the cognitive and affective dimensions of tutoring (Chhabra et al., 2022). Tutor responses

to open-ended questions and multiple-choice questions (MCQs) were recorded in DataShop, a widely used repository for educational log data in learning analytics research (Koedinger et al., 2010). Participant privacy was maintained, and consent was sought in compliance with institutional review board protocols. Tutors generally exhibited improved performance after completing the short scenario-based lessons, with average score increases of approximately 4.4% between lesson completions. However, the standard deviation in gain scores substantially varied by lesson (SD = 8.2%). We account for this variability in learning gains when validating our reliability assessment method (see Section 4.1.1).

3.2 Lesson Instructional Design and Objectives

In this study, each of the 12 lessons implements a modified *predict-observe-explain* instructional model, illustrated in Figure 1, which follows a structured, cyclical process to guide learning through situational experience (Gibbs, 1988). Initially, tutor trainees encounter a tutoring scenario (e.g., a student making a math error) and are asked to *predict* the best response through two prompts: (1) an open-response question (open) and (2) an MCQ. Following this prediction phase, tutor trainees *explain* their reasoning by answering (3) another open-ended question (open) and (4) a corresponding MCQ. Next, they observe evidence-based best practices and receive delayed corrective feedback before explaining the reasoning behind the recommended responses. The cycle concludes with a post-test, where tutor trainees respond to a new but similar tutoring scenario and repeat the modified predict-explain process for items (5)–(8). The post-test consists of four items (two MCQs and two open-response questions) that serve as the primary measure of tutor performance.

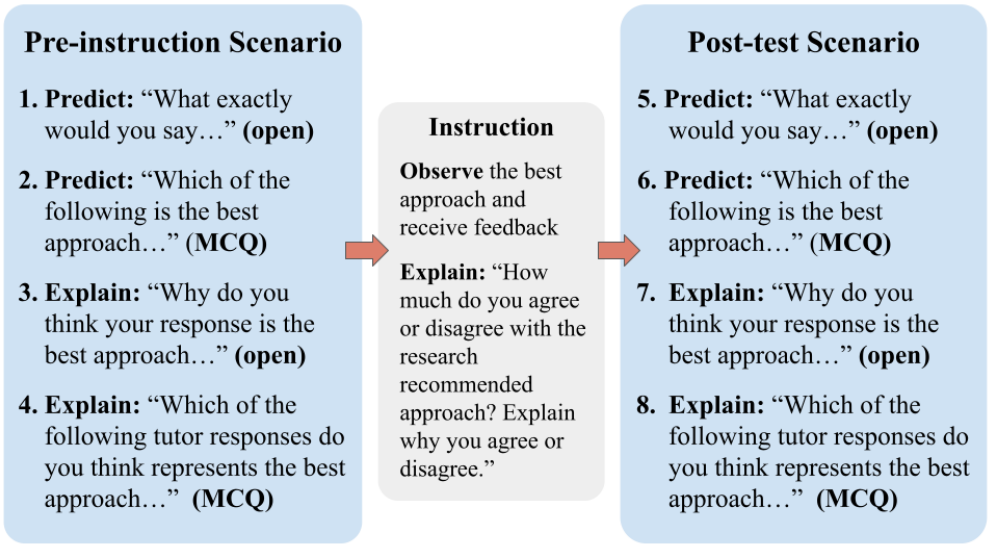


Figure 1. Lesson instructional design illustrating a modified predict-observe-explain instructional cycle adapted from D. Thomas and colleagues (2023). Tutors are presented with analogous scenarios during pre-instruction and post-test, which are used interchangeably to counterbalance difficulty.

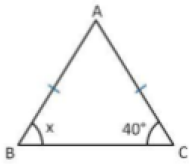
The pre-instruction and post-test scenarios are analogous, requiring tutors to apply the same learning objective. This pre-instruction serves as our pre-test measurement and is referred to as pre-*instruction* because learners receive feedback on their attempts *after all attempts have been submitted and the pre-test measurement completed*. To ensure counterbalancing of scenario difficulty, the scenarios are randomly swapped between the pre-instruction and post-test. Each predicted question contains an open response and a corresponding MCQ. Figure 2 illustrates an example of a scenario for the lesson *Determining What Students Know*. Tutors are tasked with providing textual responses predicting the best response (Q1 in Figures 1 and 2) followed by selecting the best multiple-choice response (Q2 in Figures 1 and 2).

This work involves 12 tutor lessons, each designed to support the development of key skills across socio-emotional, motivational, and content-related aspects of tutoring.

1. **Addressing Microaggressions:** Tutors learn to recognize microaggressions in classroom interactions, understand their impact on student identity and belonging, and apply strategies to appropriately and sensitively address them.
2. **Avoiding Unconscious Assumptions:** Tutors reflect on how assumptions about students' abilities, interests, or behaviour can arise unconsciously and practise strategies for more equitable and individualized engagement.
3. **Building Cultural Competence:** This lesson focuses on helping tutors understand and value students' cultural backgrounds and understand how to integrate culturally responsive practices into instruction.

Pre-Instruction Scenario (from *Determining What Students Know*)

You are working with a student named Cindy on her math homework. She is having trouble solving a geometry problem dealing with triangles. She shows you the following diagram displaying a triangle and states that she has to determine the value of angle x (shown right). Cindy says, "I don't know what to do."



1. What exactly would you say to Cindy to begin helping her solve the math problem? [**Predict-Open**]

Type your response here.

2. Which of the following tutor's responses below do you think is the most effective tutor response in helping Cindy? [**Predict-MCQ**]

A. "Cindy, this is an isosceles triangle which has two congruent sides and angles. Given this hint, can you solve the problem now?"
B. "To begin, what type of triangle is this? This information will be very helpful in solving the problem. "
C. "Let's talk about how to begin, Cindy. What do you know about the triangle?"
D. "So, you start. What do you think is the answer? How many degrees is angle x?"

3. Why do you think the tutor's response you selected in (2) will best support Cindy in solving the math problem? [**Explain-Open**]

Type your response here.

4. Which of the following statements aligns with the rationale you chose and explained in (2) and (3)? [**Explain-MCQ**]

A. By asking the student how they want to begin and asking them what they know, tutors can determine what the student already knows, or assess prior knowledge.
B. Asking a student a question about the problem, such as the type of triangle, will give the student useful information. Tutors should always help students find key information to solve the problem.
C. By giving a student the information they need to solve the math problem, they will get the correct answer quickly. When students solve problems quickly they gain confidence.
D. By having the student attempt to solve the problem in the beginning, tutors can see where the student is wrong and then correct them.

Figure 2. Sample scenario from the lesson *Determining What Students Know*. Tutors first *predict* the best approach providing textual responses (Q1), followed by selecting the best multiple-choice response (Q2), assessing the same learning objective. Then tutors *explain* the best approach in an open response (Q3), followed by selecting the best multiple-choice response (Q4). The correct multiple-choice selections are shown in green.

4. **Determining What Students Know:** Tutors develop the ability to assess prior knowledge effectively, including identifying misconceptions and asking guiding questions to surface students' understanding.
5. **Exploring Implicit Bias:** This lesson supports tutors in identifying their own potential biases, understanding how implicit bias can affect student interactions, and learning methods to minimize its influence in tutoring.
6. **Giving Effective Praise:** Tutors learn to identify features of effective praise, explain how praise can increase motivation, and apply strategies by responding to students through well-crafted, motivational praise.
7. **Helping Students Manage Inequity:** Tutors are trained to recognize situations where a student may be experiencing inequity related to learning and to apply strategies to help students manage and cope with such inequities effectively.
8. **Narrowing Opportunity Gaps:** Tutors examine systemic and classroom-level factors that contribute to opportunity gaps and practise applying strategies to support more equitable learning outcomes.
9. **Reacting to Errors:** Tutors are trained to respond productively when students make mistakes, using approaches that maintain motivation and turn errors into learning opportunities.
10. **Responding to Negative Self-Talk:** This lesson teaches tutors to recognize when a student is engaging in negative self-talk and to respond using evidence-based strategies that support student confidence and resilience.
11. **Supporting a Growth Mindset:** Tutors learn to identify fixed-mindset behaviours in students and respond in ways that foster a growth mindset, strengthening students' belief in their ability to learn and improve.
12. **Using Motivational Strategies:** Tutors explore intrinsic and extrinsic motivation and practise applying strategies to engage students through motivational responses aligned with each type.

3.3 Data Preprocessing

We preprocessed the data through a series of steps to ensure consistency and completeness. First, we included only tutor trainees who completed entire lessons, defined as responding to both pre-tests and post-tests, resulting in eight responses per lesson. We then aggregated the data into a structured format, with each row representing a unique combination of tutor, trainee, lesson, and item. Each item was labelled according to its measurement point (pre-test or post-test), and binary outcomes were generated. MCQs were automatically scored using templated solutions.

For open-ended responses, we analyzed open-response grades from open-source datasets identified in prior research (D. R. Thomas et al., 2025a, 2025a). This research used GPT-4turbo classifiers, selected for this study based on their representation across all 12 lessons of past work. Specifically, past research used scoring rubrics to first code human-labelled ground truth and then validate each classifier against it. The ground-truth data featured reported inter-rater reliability (IRR) for all lessons ranging from 0.64 to 0.88 for *predict* and 0.65 to 0.91 for *explain* open responses for the grading of human ground-truth data, which are common and acceptable inter-rater agreements for ill-defined domains such as situational judgment testing in scenario-based training, such as those featured in the present study (D. Thomas et al., 2023; D. R. Thomas et al., 2025b, 2025a; Lin et al., 2025). We refer to these publications for in-depth documentation of the data-coding process. Based on these human codes, a few-shot prompting approach was implemented for grading using learner-generated responses in conjunction with human-scored examples to help generate accurate *predict* (Q1 and Q5) and *explain* (Q3 and Q7) open-response types. Prompt development was an iterative process, refined through multiple feedback cycles. Strategies included context framing (e.g., "You are a tutor evaluator…"), incorporating human-coded response examples, and using chain-of-thought prompting (Wei et al., 2022) to encourage reasoning. To ensure consistency, the model temperature was set to 0, and the responses were truncated at 300 tokens to maintain conciseness. F1 scores range from 0.71 to 0.85 across lessons. In a study of 383 tutor trainees from three lessons (*Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*), researchers reported that the few-shot learning approach yielded an average F1 score of 0.84 and an AUC score of 0.85 (Lin et al., 2025). Comparable accuracies for the other lessons and full LLM prompt templates are reported in D. R. Thomas and colleagues (2025b, 2025a).

4. Analytical Methods

4.1 Proposed Method for Reliability Assessment (RQ1)

To address **RQ1**—how to estimate test reliability in short pre-post assessments while accounting for learning gains—we propose a modified split-half reliability procedure rooted in classical test theory. Our approach is based on the Rasch model and retains its core assumptions: a dominant underlying skill dimension, monotonic item response functions, and approximate local independence of item responses conditional on person ability and item difficulty. Unlike a classical Rasch model, however, we

do not assume a time-invariant latent trait: we explicitly relax the stability assumption by including a time effect to account for systematic learning gains between pre- and post-assessment, while preserving the remaining Rasch measurement assumptions for reliability estimation.

Both Cronbach's α and split-half reliability are standard measures of internal consistency that assess how well a set of items captures a common latent construct. Cronbach's α summarizes internal consistency by averaging inter-item correlations. At the same time, split-half reliability involves dividing the test into two halves and computing the correlation between total scores on each half (and then upward-adjusting the result for the fact that half the items are expected to lead to lower reliability estimates). We validate that our method closely approximates the widely used Cronbach's α through simulation (Section 4.1.1). This validation is based on the idea that Cronbach's α is approximately equivalent to the average of all possible split-half reliabilities (Warrens, 2015). This equivalence was demonstrated using the Flanagan–Rulon adjustment for split-half reliability (Warrens, 2015). In this study, we use the common Spearman–Brown adjustment for split-half reliability, which in itself leads to highly similar results as the Flanagan–Rulon adjustment, assuming that the response variance in each split is not substantially different (Walker, 2005). This assumption is theoretically justified and methodologically appropriate, as the goal of this study is to validate our method against the widely used Cronbach's α reliability estimate–our goal is *not* to show exact equivalence but to approximate convergence based on simulation. The reasons we use adjusted split-half reliability are its strength and its flexibility, particularly suited to data with missing responses or repeated measures, through which tutor trainees learn.

Our procedure computes the correlation between person scores derived from two test halves and then applies the *Spearman–Brown correction* to adjust for test length. Because reliability increases with the number of items, the raw split-half correlation typically underestimates full-test reliability. The Spearman–Brown formula compensates for this by estimating the reliability the full-length test would achieve and was originally described over 100 years ago (Spearman, 1910; Brown, 1910) and is still widely used today (Warrens, 2015).

Conventional reliability estimates such as Cronbach's α assume that the underlying trait—for instance, tutor skill—remains constant across administrations. However, in educational interventions, this assumption is often violated, as learners are expected to improve from pre-test to post-test. Ignoring these learning gains might lead to an underestimate of reliability by misattributing true growth to random measurement error. To adjust for this, we extend the split-half method by modelling a person's abilities using Rasch models and comparing estimates across balanced item subsets. The term *balanced* denotes averaging across many random partitions, which is expected to even out noise introduced by splitting the data in particular ways. Specifically, we fit mixed-effects logistic regression models of the form

$$\text{logit}(P(\text{response}_{ijk})) = \beta_0 + u_{item_j} + u_{student_i} + u_{time_k} + \varepsilon_{ijk}, \tag{1}$$

where $u_{student_i}$ captures individual differences in tutor skill, u_{item} models item difficulty, and u_{time} models time-specific effects of learning between pretest and posttest. The outcome P_{ijk} is the probability of a correct response for tutor i on item j at time k; β_0 represents the fixed intercept, and ε_{ij} the residual error. This model is equivalent to a Rasch model, often used in estimating student ability in one-time assessments without learning gains, except for the time effect u_{time} (Bond & Fox, 2013).

We use Algorithm 1 to estimate reliability using a split-half approach analogous to Cronbach's α, which is approximately equivalent to the average of all possible split-half correlations (as described earlier in this section). For each balanced item split, we fit Rasch models to each half, extract student ability estimates (θ_{s1} and θ_{s2}) (corresponding to $u_{student_i}$ in Equation 1), compute their Pearson correlation, and apply the Spearman–Brown correction: $r_{\text{SB}} = \frac{2r}{1+r}$, where r is the obtained correlation. This process is repeated across all valid item splits or, if that number of splits is very large, a sufficiently large random subset (e.g., 1,000) to yield an average reliability estimate. This average is the final reliability estimate of our method. Specifically, balanced item splits were generated by repeatedly sampling random split-halves of the item set without replacement. For each split, the Rasch model was fit to each half independently, and reliability was computed from the correlation of student ability estimates. Reliability was estimated by averaging Spearman–Brown-corrected correlations across a large number of unique random splits (1,000 in our analyses). Missing responses were not imputed in our main Rasch-based split-half procedure. Instead, for each split-half, we fit the Rasch (mixed-effects logistic) model using only the observed student–item responses in that half, yielding student ability estimates from the available data. We systematically investigate the robustness of our method to different levels of missingness in the data in Section 5.1.

Novelty of the approach. This approach offers a principled adaptation of classical reliability estimation methods for short, learning-sensitive assessments with missing data. Explicitly accounting for potential gains over time and leveraging item-level modelling yields improved estimates of score reliability in personalized instructional settings, such as tutor training. While split-half reliability and Rasch models are well established individually, existing reliability estimators assume static traits, and Rasch models with time effects are typically used for estimating growth rather than for deriving test-level reliability. Our contribution lies in combining these ideas to estimate reliability under learning: we obtain person ability estimates from balanced item splits using a Rasch model that explicitly includes a time effect and then assess the consistency of these estimates

Algorithm 1: Compute split-half reliability with Spearman–Brown correction for adjusted student scores.

Input: Assessment data d_{irt} with binary responses
Output: Average corrected reliability

1. Identify all unique assessment items in d_{irt}.
2. Generate balanced item splits (ensuring coverage of time and item types).
3. **For each** split do:
   - Fit Rasch model to $half1_{items}$ to obtain θ_{s1}.
   - Fit Rasch model to $half2_{items}$ to obtain θ_{s2}.
   - Compute $r = \text{cor}(\theta_{s1}, \theta_{s2})$.
   - Apply Spearman–Brown correction: $r_{\text{SB}} = \frac{2r}{1+r}$.
4. Aggregate corrected correlations (r_{SB}) across splits.
5. Report mean as an estimate of overall reliability.

using a Spearman–Brown-corrected split-half correlation. This integration allows reliability to be estimated even when pre-post gains and missing student–item combinations would invalidate standard α-based approaches.

4.1.1 Reliability Estimation Validation

To validate our reliability estimation procedure under realistic conditions, we conducted a simulation study modelled on the structure of our instructional dataset, which includes 12 lessons with four binary-scored items each, completed by 985 students. Our goal was to simulate item-level response data that converges to a Cronbach's α of over 0.8, a commonly accepted threshold for good internal consistency (Gliem & Gliem, 2003). We then tested, after adding learning gain and missing data to this simulated data set, whether our method could appropriately recover a reliability estimate close to Cronbach's α on the original data set.

We aimed to generate data from a distribution that would correspond to a Cronbach's α of about 0.8. However, such a distribution can only be indirectly defined, as the standard deviation of item difficulties influences reliability (Walker, 2005). Therefore, we tuned the parameters of a Rasch model (see Section 4.1) and observed its Cronbach's α, noted it, and then generated data with that model to see if our reliability estimation procedure can approximate that Cronbach's α under assumptions of learning gain and missingness. Specifically, we began by generating person abilities (θ) from a standard normal distribution and item difficulties (b) from a normal distribution. We first tuned the standard deviation (SD) of item difficulties in a separate simulation step. Based on a simple grid search of item difficulty SDs between 0.05 and 2 in steps of 0.05, we identified 1.8 as producing a Cronbach's α of 0.875, which we deemed satisfactory for our simulation and closest to 0.8 while still being above 0.8.

Once the item difficulty spread yielding an empirical α near 0.8 was identified, we simulated datasets using the parameters producing that alpha while incorporating two *additional* design complexities: (1) pre-post structure, where a log-odds gain was applied to post-test ability distributions to reflect expected learning gains (0 to 1, in steps of 0.2), and (2) varying levels of missing data (0% to 50%, steps of 10%). We then used our reliability estimation procedure (Section 4.1) on each simulated data set and compared whether its reliability estimate was truthful or biased relative to the expected reliability of 0.875 (identified in the step above). We did so by running 100 simulations each and averaging the reliability of our procedure. We note that our search space is consistent with our sample, with an average learning gain of 4.4% in score increases, corresponding to a Cohen's d of 0.12 and a log-odds effect size, based on conversations reported in Chinn (2000), of about 0.22. We similarly assumed a moderate student-level SD in learning gains of 0.3 across all simulations. To see how our method compares to alternative methods of estimating reliability in our context, we also computed Cronbach's α as a baseline for (a) different learning gains by treating pre- and post-test observations as distinct items (without gain adjustment) and (b) different levels of missingness by imputing item means (average scores on items) to missing values, which is often done in research (Béland et al., 2016). This contrast shows whether our method (in addition to being consistent with the 0.875 estimate) also yields less divergence and bias from 0.875 than existing methods.

4.2 RQ2: Estimating the Reliability of LLM-Based Scoring across Lessons

To evaluate the reliability of LLM-based scoring across multiple short assessments of tutor skills (RQ2), we applied the adjusted split-half procedure described in Section 4.1. This method accounts for short test lengths, missing responses, and repeated measures in between learning interventions by computing the correlation between person ability estimates from independently calibrated test halves and adjusting the result using the Spearman–Brown formula.

We estimated internal consistency at two levels: (1) across all items pooled from multiple lessons and (2) separately for each lesson. Where sample sizes permitted, we also disaggregated reliability estimates by item type (e.g., multiple-choice vs. open-ended). This multi-level estimation allowed us to examine whether aggregating item responses across lessons produces more stable and reliable measures of tutor skill than individual short assessments alone.

4.3 RQ3: Assessing Dimensionality of Tutor Skills

Because combining lessons assumes that they measure a coherent underlying construct, we investigated the dimensionality of lesson-level scores to study the dimensionality of tutor skills (RQ3). First, we examined the inter-lesson correlation matrix to identify descriptive relationships among related assessments. Second, we compared the fit of a unidimensional Rasch model—assumed to be a single latent tutor skill—to a model with independent dimensions for each lesson. These two model fits included the random intercept for measurement time, as described in Section 4.1. Model fit comparisons based on the Akaike information criterion (*AIC*) and the Bayesian information criterion (*BIC*) informed the tenability of aggregating scores into a composite metric.

To identify the optimal number of latent dimensions, we triangulated several dimensionality reduction techniques commonly used in learning analytics (Baker, 2025). We first conducted Horn's parallel analysis with 1,000 iterations to establish a threshold for eigenvalues that exceed those expected by chance (Dinno, 2009). Next, we compared scree plots from three principal component analyses (PCAs), each using a different approach to handling missing data: (1) mean imputation at the lesson level, (2) pairwise-complete correlation matrices without imputation, and (3) cross-validated PCA using leave-one-out reconstruction error (Josse & Husson, 2012), which is often considered a gold standard in the field (Baker, 2025). We then incorporated the resulting factor structures into a series of hierarchical models (following the same procedure as described in the previous paragraph), comparing their fit using *AIC* and *BIC* against a baseline Rasch model (assuming a unidimensional tutor skill; see above). This multi-method strategy provided converging evidence for a low-dimensional representation of tutor trainee performance that remained stable across preprocessing and modelling choices.

We treated individual lessons as the unit of analysis. We finally interpreted the resulting factor-analytic structure using factor loadings (Yong & Pearce, 2013) and the common threshold of 0.4 for factor loadings (Baker, 2025), indicating whether a lesson was included in a factor. We then informally interpreted the resulting groupings of tutor lessons (see Section 3.2 for a full overview).

5. Results

5.1 RQ1: Validation of the Proposed Method for Reliability Assessment

First, we performed a convergent validation of our reliability estimation procedure to see if it would align with a Cronbach's α of 0.875 when sampling from distributions that, on average, yield such an α (see Section 4.1.1). As shown in Figure 3, our proposed method yielded satisfactorily consistent reliability estimates across varying levels of learning gains and missing data, aligning approximately with the gold-standard Cronbach's α metric for reliability, which we used as the input parameter of the simulation (see Section 4.1.1). While the pattern of reliability estimates was sometimes above and below the Cronbach's α parameter (due to random chance during simulation), our procedure showed no substantial bias in estimating reliability and is accurate to a margin of about ± 0.015 for missing data and ± 0.01 for learning-gain differences. These findings support the convergent validity of our approach.

To see how our method compares to alternative methods of estimating reliability in our context, we also computed Cronbach's α as a baseline for (a) different learning gains by treating pre- and post-test observations as distinct items (without gain adjustment) and (b) different levels of missingness by imputing item means (average scores on items) to missing values, which is often done in research (Béland et al., 2016) (see Section 4.1.1). As the blue trend line in Figure 3 shows, the unadjusted baseline tends to systematically underestimate reliability for scenarios with learning gain (especially for larger gain, up to about 0.2) and for higher missingness of 40% or more (which is the case in our data set, for example), where its estimates diverge by 0.03 or more from the expected value. While these differences are promising, we also note some overestimation of reliability for low-to-moderate missingness (10–30%), which we return to in the discussion.

5.2 RQ2: Overall Reliability of LLM-Based Scoring

To assess the overall reliability of LLM-based scoring across lessons, we applied our proposed split-half procedure, which adjusts for short assessments and potential learning gains (see Section 4.1). This approach estimates the internal consistency of

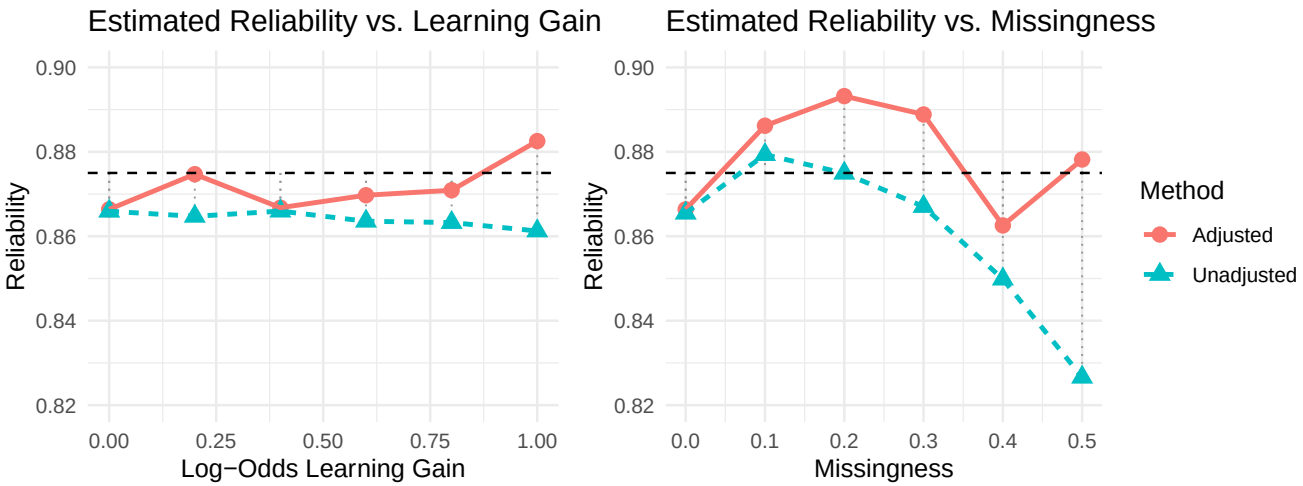


Figure 3. Estimated reliability of the proposed method (red) as a function of learning gain (left) and missing data proportion (right). Across both simulations, the method yields reliability estimates that closely align with the simulation input Cronbach's α parameter (dotted line) and, overall, more truthful estimates than unadjusted baselines (blue).

scores derived from both multiple-choice and open-ended responses, leveraging Rasch-based ability estimates and applying the Spearman–Brown correction across a large set of item splits.

Table 1. Split-half reliability by item type, using Rasch-based person scores and the Spearman-Brown correction.

Source	Reliability
All Items (Pooled across Lessons)	0.774
Multiple-Choice Items Only	0.652
Open-Ended Items Only	0.733

As shown in Table 1, reliability estimates were highest and reached a satisfactory level (Post, 2016) when pooling both multiple-choice and open-ended items across all lessons ($r = 0.774$), indicating strong internal consistency. Disaggregated analyses reveal that open-ended responses ($r = 0.733$) were scored more reliably than multiple-choice items ($r = 0.652$).

We also estimated reliability at the individual lesson level. Due to limited data per lesson—each comprising only four items—reliability scores were notably lower on average. Across the 12 lessons, the mean split-half reliability was $M = 0.60$ (SD $= 0.04$). Given the brevity of each lesson's assessment and the small number of items, these results are consistent with the prediction that shorter assessments with fewer items will lead to lower reliability (Post, 2016; Nunnally & Bernstein, 1994). Reliability estimates by item type (multiple choice vs. open-ended) *within* single lessons were not reported for multiple-choice items due to insufficient item counts for stable estimation (one per measurement time).

These findings underscore a key trade-off in short assessment design: while individual lessons yield limited reliability, aggregating assessments across lessons produces substantially more stable and reliable scores. This increase in reliability is only justifiable if the lessons have a common underlying construct—a question we address through dimensionality analysis in the following section (Section 5.3).

5.3 RQ3: Factor Structure of Tutoring Skills

To investigate the underlying structure of tutoring skills, we examined whether scores across the 12 lessons reflected a unidimensional or multidimensional construct (**RQ3**). We began by comparing a unidimensional Rasch model with one that allowed a separate latent skill per lesson. The unidimensional model provided a superior fit, with a lower Akaike information criterion ($AIC = 20,931.51$) compared to the multidimensional model ($AIC = 21,379.80$), suggesting that the assumption of a single latent tutoring skill is more parsimonious and better supported by the data. The same was true when comparing both models based on the Bayesian information criterion ($BIC = 20,963.79$ compared to $21,412.08$, respectively).

Nonetheless, we explored whether any underlying factors, grouping lessons, could better account for the observed patterns across lessons. Table 2 presents a correlation matrix of lesson-level scores. Several strong inter-lesson correlations emerged. In particular, *Supporting a Growth Mindset* and *Responding to Negative Self-Talk* were highly correlated ($r = 0.57$), as were *Using Motivational Strategies* and *Helping Students Manage Inequity* ($r = 0.48$). These findings suggest that some lessons

Table 2. Lower triangle of the correlation matrix between lessons. Asterisks denote statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Lesson (Abbreviation)	ST	EIB	AUA	Mot.	Inq.	BCC	Err.	SK	GM	AM	NOG	Pr.
Self-Talk (ST)	.											
Exploring Implicit Bias (EIB)	0.38*	.										
Avoiding Unconscious Assumptions (AUA)	0.33*	0.41***	.									
Motivation (Mot.)	0.37***	0.40*	0.59***	.								
Inequity (Inq.)	0.48***	0.50***	0.22*	0.48***	.							
Building Cultural Competence (BCC)	0.17	0.23*	0.27**	0.05	0.20	.						
Errors (Err.)	0.26***	−0.08	−0.00	0.42***	0.19**	0.25	.					
Student Knowledge (SK)	0.28***	−0.04	0.22	0.41***	0.13	0.28*	0.39***	.				
Growth Mindset (GM)	0.57***	0.59***	0.14	0.57***	0.38***	−0.21	0.41***	0.34***	.			
Addressing Microaggressions (AM)	0.05	0.21*	0.17	0.25	0.12	−0.02	−0.09	0.05	0.33	.		
Narrowing Opportunity Gaps (NOG)	0.09	0.19	0.37***	0.48*	0.16	0.06	0.11	0.21	0.18	0.15	.	
Praise (Pr.)	0.36***	−0.05	0.18	0.49***	0.29***	0.15	0.38***	0.35***	0.46***	0.26	0.35*	.

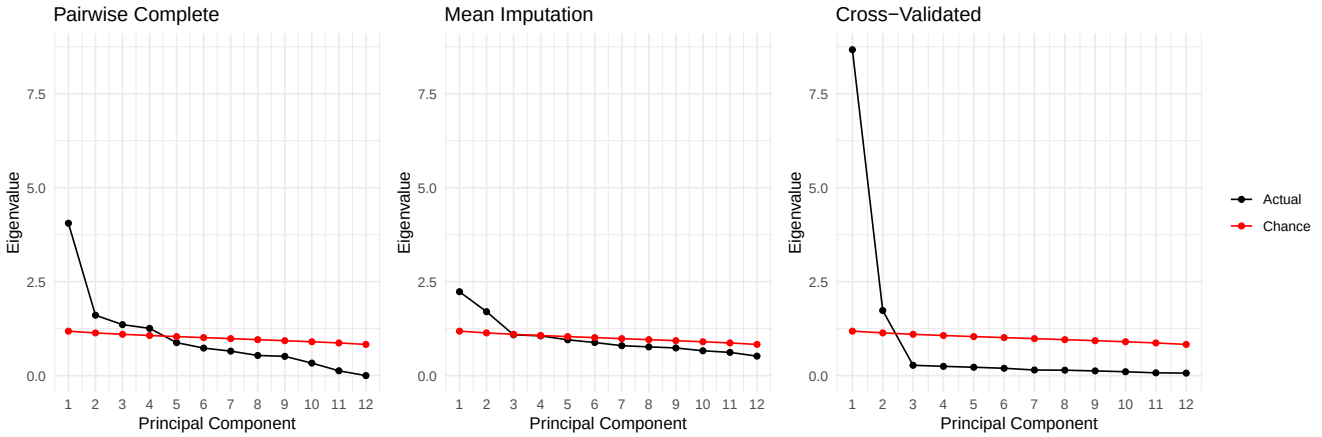


Figure 4. Scree plots from three PCA methods: pairwise-complete (left), mean imputation (centre), and cross-validated (right). Black lines show actual eigenvalues; red lines indicate chance-level eigenvalues from Horn's parallel analysis.

cluster around common constructs—particularly those involving socio-emotional and motivational competencies—while others, such as content-focused lessons, display more independence.

To quantify the underlying structure and assess its robustness, we conducted multiple PCAs to gather converging evidence (see Section 4.3). Figure 4 displays scree plots from PCA under three preprocessing strategies: pairwise-complete correlation matrices, mean imputation, and cross-validated reconstruction. In each case, we overlaid eigenvalues from Horn's parallel analysis (red) as an at-chance benchmark. Across all methods, the first two components consistently exceed chance, indicating a stable low-dimensional structure. However, the pairwise-complete method suggests that four factors may be above chance, while mean imputation also suggests an elbow point at three dimensions. Hence, we compared model fit across the three proposed solutions: two-, three-, and four-factor models. To evaluate how these factor structures support modelling, we grouped lessons based on their strongest PCA loadings and estimated hierarchical logistic models using two-, three-, and four-factor solutions. Table 3 shows model comparisons using *AIC* and *BIC*.

Table 3. Model comparison based on *AIC* and *BIC* for different dimensionality assumptions. Lower values indicate a better fit.

Model	*AIC*	*BIC*
Rasch	20931.51	20963.79
PCA-Grouped (2)	20933.50	20973.85
PCA-Grouped (3)	20931.13	20971.48
PCA-Grouped (4)	**20918.85**	**20959.20**
Lessons as Factors	21379.80	21412.08

Based on Table 3, the four-factor model provided the best overall fit, achieving the lowest *AIC* and *BIC* values among all models tested. While the three-factor model slightly improved upon the Rasch baseline in terms of *AIC*, it did not offer a consistent advantage across both criteria. The two-factor model performed worse than the Rasch model on *BIC*, suggesting limited benefit from that level of dimensionality. Finally, the model, which treated each lesson as its own dimension, showed a

substantially worse fit. Together, these results support a four-factor solution.

Table 4. Item–factor associations from the four-factor solution using a loading cutoff of 0.4. Items may load on multiple factors if they meet the threshold.

Item	Factors (loading $\geq$ 0.4)
Giving Effective Praise	F1, F2, F3
Reacting to Errors	F1, F2
Using Motivational Strategies	F1, F4
Addressing Microaggressions	F1
Helping Students Manage Inequity	F1
Supporting a Growth Mindset	F1, F2, F3
Responding to Negative Self-Talk	F1, F2, F3
Determining What Students Know	F1, F2
Avoiding Unconscious Assumptions	F1, F3
Building Cultural Competence	F2
Exploring Implicit Bias	F1
Narrowing Opportunity Gaps	F1, F3

The resulting factor loadings, summarized in Table 4, reveal that **Factor 1 is shared across nearly all lessons**, suggesting a dominant general proficiency dimension underlying tutor performance. Most lessons load on Factor 1 alongside one or more secondary factors, indicating partial unidimensionality. *Giving Effective Praise*, *Reacting to Errors*, *Supporting a Growth Mindset*, *Responding to Negative Self-Talk*, *Determining What Students Know*, and *Building Cultural Competence* load on Factor 2, pointing to a motivational tutoring knowledge dimension (i.e., where tutors need to apply interpersonal competencies to help motivate a student). Factor 3 captures fairness-related themes, such as *Avoiding Unconscious Assumptions* and *Narrowing Opportunity Gaps*, while Factor 4, somewhat surprisingly, is defined primarily by *Using Motivational Strategies*.

Together, these findings support a structure anchored by a strong general factor and enriched by interpretable subdimensions. This partial unidimensionality suggests that while a single proficiency score may be appropriate for broad assessment, finer-grained diagnostics are possible when data quantity allows. This finding aligns with the fact that the *AIC* and *BIC* fit margins in Table 4 between the unidimensional Rasch model, and the two-, three-, and four-factor solutions are small.

6. Discussion

Assessment (i.e., grading learner behaviour to generate feedback and other analytics) is a core method in learning analytics. In recent years, assessment has been increasingly done by LLMs. Their use of assessment has been commonly justified and validated by comparing their scoring to human-graded responses, using metrics such as accuracy, precision, or Cohen's κ to assess consistency (D. R. Thomas et al., 2025b; Dai et al., 2024; Cao et al., 2025).

Reliability, generally considered a prerequisite to validity (Bannigan & Watson, 2009), refers to the degree to which scores are consistent across repeated or equivalent assessments and is typically not assessed in this line of work (Yan et al., 2024; Misiejuk et al., 2025). However, awareness of this issue is increasing in the context of content authoring with LLMs (Bhandari et al., 2024). LLM's stochastic nature, hyperparameter settings, and hallucination may all contribute to a lack of reliability in assessment (Borchers et al., 2025). We thus see a crucial gap in the field: the need for novel methods to measure reliability in learning analytics contexts where LLMs are used for assessment, especially when learners often improve and learn between measurements.

This study advances our understanding of the reliability and structure of tutor skill assessments, especially those scored by LLMs in short-form online training contexts. We contribute (1) a novel method for estimating test reliability that accounts for learning gains, (2) empirical evidence supporting the internal consistency of LLM-based scoring, and (3) new insights into the factor structure of tutoring competencies. These findings inform both the design of tutor training programs and the implementation of AI-supported assessment in learning analytics.

6.1 RQ1: Learning-Aware Reliability Estimation

Reliable measurement is crucial for detecting learning gains in short pre-post formats. Classical estimates like Cronbach's α assume static traits across repeated assessments, which are often violated in intervention settings. We proposed a Rasch-based, split-half reliability method that adjusts for test length and learning effects. Simulations demonstrated its robustness under various missingness and gain conditions and convergent validity in relationship to the widely used Cronbach's α (Warrens, 2015). *After* introducing missingness and learning gain to the data and re-estimating Cronbach's α as a baseline, we also

demonstrated that our method systematically improves reliability estimation. Specifically, applying Cronbach's α without pre-post adjustment resulted in systematic underestimation of reliability relative to our method, especially for larger gain values. Hence, our proposed method is particularly relevant for contexts where students are expected to learn. Notably, researchers and practitioners often wish to use pre- *and* post-test items to estimate reliability and assess learners in other contexts. In these cases, we recommend adopting our methods in future learning analytics applications involving growth-sensitive assessments, using our open-source code (see URL in Section 1). While the reliability estimation accuracy increases attributable to our method might seem small to some (e.g., up to 0.02 for learning gain and up to 0.06 for missing data), these statistical differences may translate into practical differences of high or low magnitude depending on context (e.g., scale, high-stakes vs. low-stakes assessment). We also note that the increases observed were computed for assumptions based on our data set (e.g., sample size). Therefore, future work is encouraged to replicate our method in other contexts (to assess whether larger improvements can be made) and to study the impact of improved reliability estimation on downstream applications (e.g., the impact of less reliable assessments on feedback accuracy and dropout prediction models). As the present study is an initial step toward broader adoption of our method, these evaluations are beyond its scope. Finally, we note that our method showed a slight overestimation of reliability for low-to-moderate rates of missing data (10–30%), which might also be due to the nature of our data set (as the unadjusted reliability estimates showed a parallel trend line). Hence, while our method shows promise in dealing well with high rates of missing data (40% or more)—common in personalized learning; see L. Zhang and colleagues (2024)—more studies of our method in more learning analytics contexts are desirable.

Our proposed method has broader implications for learning analytics beyond the assessment of short pre-post evaluations. In particular, we see strong potential for adapting this learning-aware reliability estimation to knowledge-tracing models, which are widely used to monitor learner knowledge across sequences of learning activities (Abdelrahman et al., 2023). While our current approach focuses on estimating the reliability of assessments with a fixed pre-post structure, many knowledge-tracing frameworks—such as the individualized Additive Factors Model (iAFM) (R. Liu & Koedinger, 2017) and individualized Bayesian Knowledge Tracing (iBKT) (Pardos & Heffernan, 2010)—similarly estimate student-specific latent traits, often representing initial or evolving proficiency or mastery. Future work may apply our method to these contexts.

In particular, these knowledge-tracing and cognitive-growth models are increasingly used to generate actionable insights in teacher dashboards and learning analytics tools (Borchers et al., 2024), yet the reliability of their outputs is rarely examined. We argue that the core idea behind our method—splitting assessment data to evaluate the consistency of derived student ability estimates—can be extended to these longitudinal models. Rather than splitting by items as in our Rasch-based setup, one could construct step-level or temporal splits of student learning trajectories (e.g., early vs. later attempts) and assess the correlation of student parameter estimates across these partitions, again applying a Spearman–Brown correction to account for data reduction.

Such an extension would provide a principled way to quantify the reliability of growth and knowledge-tracing model outputs, offering important reliability checks for their use in real-world decision-making. It also allows researchers and practitioners to identify when model-based analytics are sufficiently reliable to inform feedback. We see this as a promising direction for future work, connecting psychometric reliability theory with state-of-the-art learner modelling techniques in educational data mining and learning analytics.

6.2 RQ2: Reliability of LLM-Based Scoring

To assess the reliability of LLM-based scoring across tutor training lessons, we applied our proposed split-half procedure (Section 4.1). LLM-based scoring demonstrated moderate to high internal consistency, with a reliability of 0.733. This result exceeds commonly accepted minimal standards of 0.7 for a reliable psychometric measurement (Post, 2016). It suggests that reliable tutor assessment can be achieved with as few as 28 items (four pre-test items and four post-test items across an average of 3.5 lessons per student), as used in our design. Open-ended responses scored by the LLM were more reliable than multiple-choice items, supporting the increasing use of LLMs for open-response assessment (Dai et al., 2024). Our finding also supports the argument that open-response assessments can capture knowledge more holistically (D. R. Thomas et al., 2025b), as they allow for diverse expressions of understanding. When scored reliably, such variation across items can enhance the measurement of underlying knowledge constructs. At the lesson level, reliability averaged around $r = 0.60$, consistent with the expectation that fewer items lead to lower reliabilities (Warrens, 2015). This reliability increased with the number of lessons completed: participants completed an average of 3.5 lessons, yielding an overall reliability of 0.774 when combining multiple-choice and open-ended responses. Combining both, therefore, led to the most reliable assessment in our context.

Looking ahead, several LLM configuration factors are known to influence scoring reliability, including temperature, model architecture, model size, prompting strategy, and use of self-consistency. Each of these can impact both scoring validity and alignment with human ratings (Borchers et al., 2025). In future work, these parameters should be systematically varied to examine their effects on scoring reliability. An open question remains: how many items are needed to achieve reliable scoring in different conditions? While our current design achieves acceptable reliability with an average of 28 items, further work is needed to explore the generalizability of this threshold across models and domains. We consider this question to be outside the

scope of the present study but important for future work.

6.3 RQ3: Dimensionality of Tutoring Skills

To investigate the underlying structure of tutoring skills, we examined whether assessment scores across the 12 lessons reflected a unidimensional or multidimensional construct. As prior work in tutor training has emphasized a variety of socio-emotional and pedagogical competencies (Chhabra et al., 2022; National Student Support Accelerator, 2023), it remains an open question whether these competencies are separable constructs or reflect a general latent proficiency that can be assessed holistically.

We first compared the fit of a unidimensional Rasch model—assumed to represent a single latent trait of general tutoring skill—to a multidimensional model with a separate dimension for each lesson. The unidimensional model yielded a superior fit, suggesting that, despite topical diversity, tutor performance across lessons was quite well explained by a general underlying ability. Our findings support the notion of *partial unidimensionality*, where skills across content- and socio-emotional–focused lessons share sufficient commonality to justify aggregation for high-level assessment (Reise, 2012). This partial unidimensionality was especially evident in all lessons (except one), which loaded onto a single factor in a four-factor solution, exhibiting a marginally better fit than a single-factor solution.

Despite this partial evidence for unidimensionality, scree plots and Horn's parallel analysis consistently identified between two and four factors with eigenvalues exceeding chance-level values. The best-fitting model—a four-factor solution—revealed interpretable groupings aligned with known domains in the tutor-training literature, including cognitive tutoring, motivational support, and fairness-based instruction (D. Thomas et al., 2023; Butler, 2018).

Taken together, this structure of tangible subdimensions with a single, strong factor resembles bifactor models in psychometrics, where a general factor accounts for shared variance. At the same time, orthogonal subfactors capture specific skill dimensions (Reise, 2012). Although our analysis did not explicitly fit a bifactor model, the observed pattern suggests that such a structure may provide a fruitful direction for future confirmatory work.

6.4 Implications for Learning Analytics

Our results have practical implications for both assessment design and instructional planning. From a tutor assessment standpoint, the partial unidimensionality supports the use of aggregated scoring for formative evaluation—an approach that can reduce testing time while maintaining adequate reliability, as demonstrated in Section 5.2. At the same time, the existence of interpretable subdimensions suggests opportunities for more targeted feedback (Lin et al., 2025). For instance, identifying whether a tutor struggles more with motivational support than with error correction may guide personalized professional development.

More broadly, our findings align with theoretical accounts of tutoring as a multifaceted practice involving both pedagogical knowledge and socio-emotional intelligence (R. E. Wang et al., 2024; Robinson, 2022). In learning analytics, modelling such multidimensional constructs supports a richer understanding of learner trajectories and can enhance the interpretability of dashboards and analytics tools used by instructors and program evaluators (Baker, 2025; Borchers et al., 2024). As such, dimensionality analysis is not merely a psychometric exercise but a foundational step toward designing effective, scalable, and interpretable AI-supported feedback and assessment in learning analytics. In addition to our domain-specific findings regarding the dimensionality of tutor skills, our method and procedures for assessing reliability in learning settings can serve as a blueprint for other learning analytics researchers and practitioners assessing competencies during online learning, including massive open online courses, learning management systems, and other forms of educational technology (Fischer et al., 2020).

Additionally, our contribution can be used as a measurement layer for learning analytics research that relies on LLM-derived scores. Measurement in learning analytics is inherently noisy, and conclusions about learning and intervention effectiveness are bounded by the dependability of the measures that feed downstream models and decisions (Bergner, 2017; Winne, 2020). In practice, this means reliability estimates should be reported alongside LLM-scored outcomes before those outcomes are used as predictors, mediators, or evaluation endpoints. This is particularly relevant for applications that translate analytics into action, where the field has called for tighter links between analytics, decision-making, and impact (Wise et al., 2021). For example, reliability-aware LLM scoring can strengthen early-warning and dropout-risk pipelines that incorporate open-ended reflections or tutor responses, since low reliability can inflate noise and produce unstable risk signals. For system designers building LLM-scored assessments and dashboards, our results suggest treating reliability as a design constraint rather than only a retrospective statistic. Recent work distinguishes agreement with human grades from score consistency and shows that rubric and prompt design can materially affect consistency, which supports making scoring protocols explicit and versioned (D.-W. Zhang et al., 2024). In operational deployments, reliability estimates can directly drive workflow policies (e.g., selective human review for low-stability cases) to improve dependability while retaining scalability.

6.5 Limitations

Several limitations warrant mention. First, our sample consisted of college-aged tutors completing online modules, which limited the generalizability of our findings to other educator populations. Second, the item pool disproportionately emphasized

socio-emotional content, possibly inflating internal consistency among those skills. Third, our analysis did not explore individual item factor loadings or identify items with weak psychometric properties—this remains a task for future research. Fourth, our analysis did not systematically explore different language models or compare different versions of the employed GPT-4turbo model. Across the multi-year data collection for our sample, we believe multiple GPT-4turbo versions used for grading are represented in our data. While changing model versions in closed-source models like GPT-4 can introduce minor changes to grading behaviour, we believe this change is inevitable for large-scale, long-term projects like ours. We therefore leave the question of model stability, which could be investigated using our open-source code, for future work. Finally, while LLM-based scoring was benchmarked against human annotations, further work is needed to assess fairness across linguistic and cultural subgroups. Moreover, integrating real-world tutoring outcomes (e.g., student learning gains) would strengthen construct validity, as it remains an open question whether a tutor trainee who completes our modules with good grades will actually become a good and effective tutor in real-world tutoring practice. A further limitation is that our simulations assume missing-at-random mechanisms and stable response behaviours; in real practice, missingness may depend on ability, engagement, or item difficulty, which could interact with the modest 10–30% overestimation pattern we observed. Future work should therefore examine nonrandom missingness processes and varied item pools to assess whether this effect generalizes and to identify potential corrections.

7. Conclusion

Reliable assessment is crucial to learning analytics, particularly in online and AI-supported training environments where feedback enhances learner outcomes. This study introduced a novel method for estimating reliability in short assessments that explicitly accounts for learning gains and applied it to evaluate the internal consistency and structure of tutoring skill assessments scored by LLMs. Our findings offer concrete contributions to both learning analytics researchers and practitioners who use online assessments, particularly those scored by AI and LLMs. Specifically, we contribute an open-source framework for evaluating score reliability in pre-post learning contexts, where students improve over time, and in personalized learning contexts characterized by missing student-item observations. This study also establishes rigorous foundations for using LLMs in the formative assessment of professional tutoring skills. The three central study implications are as follows.

First, we demonstrate that LLM-based scoring—particularly using GPT-4—achieves satisfactory reliability, with a split-half reliability of $r = 0.733$ for open-ended responses. This exceeds the reliability of MCQs ($r = 0.652$), supporting prior claims that open-response formats provide meaningful and potentially richer evidence of complex tutoring competencies. Notably, combining both item types across short assessments yielded a robust overall reliability of 0.774, surpassing commonly accepted minimal psychometric thresholds and supporting the rapid increase in the use of LLMs for formative, open-ended online assessments at scale.

Second, our proposed reliability estimation method adjusts for learning gains via a Rasch-based split-half procedure and exhibits convergent behaviour with classical reliability estimates such as Cronbach's α when their assumptions approximately hold. In simulation, we define a ground-truth reliability by tuning complete, single-time Rasch data to yield a target α, and then introduce learning gains and missingness. Under these conditions, commonly used practices (such as treating pre- and post-test items as distinct items without gain adjustment or imputing item means to missing responses) systematically underestimate reliability. In contrast, our method consistently recovers values closer to the simulated reliability across realistic ranges of learning gain, missing data, and short test length. This makes the approach particularly suitable for intervention studies and growth-sensitive assessments, and more generally for evaluating the reliability of model-based scores, including potential extensions to knowledge tracing.

Third, we find evidence for a *bifactor-like structure* in tutor skills. While a unidimensional model of general tutoring skill fits the data well and supports reliable aggregation across lessons, we also identified interpretable subdimensions—motivational strategies, cognitive tutoring, and fair instruction—suggesting that targeted feedback and adaptive instructional planning are possible when sufficient data are available. Most lessons load onto a dominant general factor, affirming the viability of streamlined assessments, while the existence of domain-specific factors highlights opportunities for more personalized feedback.

Finally, our results offer clear implications for assessment design: with as few as 14 open-ended items (i.e., two open-ended pre-test and two post-test responses across an average of 3.5 short lessons), LLM-based scoring can support the reliable assessment of tutoring skills at scale. This affirms the viability of using GPT-class models in real-world instructional settings, provided that their reliability is empirically verified using open-source learning-aware methods, such as those proposed in this study. While multiple-choice items in our sample led to comparatively low reliability, a larger pool of such items may yield more competitive reliability estimates while being more efficient than open-response assessment, which is subject to future work alongside considerations of item quality.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings, and conclusions expressed in this material are those of the authors.

References

Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, *55*(11), 1–37. https://doi.org/10.1145/3569576

Baker, R. S. (2025). *Big data and education* (9th ed.). University of Pennsylvania.

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Béland, S., Pichette, F., & Jolani, S. (2016). Impact on Cronbach's α of simple treatment methods for missing data. *The Quantitative Methods for Psychology*, *12*(1), 57–73. https://doi.org/10.20982/tqmp.12.1.p057

Bergner, Y. (2017). Measurement and its uses in learning analytics. In *Handbook of learning analytics* (1st ed.). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.003

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence*, *7*, 100284. https://doi.org/10.1016/j.caeai.2024.100284

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press. https://doi.org/10.4324/9781315814698

Borchers, C., Thomas, D. R., Lin, J., Abboud, R., & Koedinger, K. R. (2025). Augmenting human-annotated training data with large language model generation and distillation in open-response assessment. *arXiv preprint arXiv:2501.09126*. https://doi.org/10.48550/arXiv.2501.09126

Borchers, C., Wang, Y., Karumbaiah, S., Ashiq, M., Shaffer, D. W., & Aleven, V. (2024). Revealing networks: Understanding effective teacher practices in AI-supported classrooms using transmodal ordered network analysis. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 371–381). ACM. https://doi.org/10.1145/3636555.3636892

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Cao, J., Zhao, C. Q., Chen, X., Wang, S., Schunn, C., Koedinger, K. R., & Lin, J. (2025). From first draft to final insight: A multi-agent approach for feedback generation. *arXiv preprint arXiv:2505.04869*. https://doi.org/10.48550/arXiv.2505.04869

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *Intelligent tutoring systems. ITS 2006. Lecture notes in computer science* (pp. 164–175, Vol. 4054). Springer. https://doi.org/10.1007/11774303_17

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C. L., Leong, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRater^{SM} v. 5.0 engine. *ETS Research Report Series*, *2018*(1), 1–31. https://doi.org/10.1002/ets2.12198

Chhabra, P., Chine, D., Adeniran, A., Gupta, S., & Koedinger, K. (2022). An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In *Proceedings of the 2022 Society for Information Technology & Teacher Education International Conference*, 11 April 2022, San Diego, California, USA (pp. 1812–1817). Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/primary/p/220956/

Chine, D. R., Chhabra, P., Adeniran, A., Gupta, S., & Koedinger, K. R. (2022). Development of scenario-based mentor lessons: An iterative design process for training at scale. In *Proceedings of the Ninth ACM Conference on Learning at Scale* (L@S 2022), 1–3 June 2022, New York, New York, USA (pp. 469–471). ACM. https://doi.org/10.1145/3491140.3528262

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22⟨3127::AID-SIM784⟩3.0.CO;2-M

Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, 100299. https://doi.org/10.1016/j.caeai.2024.100299

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Dinno, A. (2009). Implementing Horn's parallel analysis for principal component analysis and factor analysis. *The Stata Journal*, *9*(2), 291–298. https://doi.org/10.1177/1536867X0900900207

Divjak, B., Svetec, B., Horvat, D., & Kadoić, N. (2023). Assessment validity and learning analytics as prerequisites for ensuring student-centred learning design. *British Journal of Educational Technology*, *54*(1), 313–334. https://doi.org/10.1111/bjet.13290

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Further Education Unit.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Proceedings of the 2003 Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, 8–10 October 2003, Columbus, Ohio (pp. 82–88). ScholarWorks Indianapolis. https://hdl.handle.net/1805/344

Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-06026-5

Gurung, A., Vanacore, K., Mcreynolds, A. A., Ostrow, K. S., Worden, E., Sales, A. C., & Heffernan, N. T. (2024). Multiple choice vs. fill-in problems: The trade-off between scalability and learning. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 507–517). ACM. https://doi.org/10.1145/3636555.3636908

Gwet, K. (2001). *Handbook of inter-rater reliability*. STATAXIS Publishing Company.

Jang, M., & Lukasiewicz, T. (2023). Consistency analysis of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2023), 6–10 December 2023, Singapore (pp. 15970–15985). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.991

Josse, J., & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, *56*(6), 1869–1879. https://doi.org/10.1016/j.csda.2011.11.012

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 43–56). CRC Press. https://doi.org/10.1201/b10274-10

Lin, J., Chen, E., Han, Z., Gurung, A., Thomas, D. R., Tan, W., Nguyen, N. D., & Koedinger, K. R. (2024). How can I improve? Using GPT to highlight the desired and undesired parts of open-ended responses. *arXiv preprint arXiv:2405.00291*. https://doi.org/10.48550/arXiv.2405.00291

Lin, J., Han, Z., Thomas, D. R., Gurung, A., Gupta, S., Aleven, V., & Koedinger, K. R. (2025). How can I get it right? Using GPT to rephrase incorrect trainee responses. *International Journal of Artificial Intelligence in Education*, *35*, 482–508. https://doi.org/10.1007/s40593-024-00408-y

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

Liu, R., & Koedinger, K. R. (2017). Towards reliable and valid measurement of individualized student parameters. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM 2017), 25–28 June 2017, Wuhan, China (pp. 135–142). International Educational Data Mining Society. https://educationaldatamining.org/EDM2017/proc_files/proceedings.pdf

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*. https://doi.org/10.48550/arXiv.2308.05374

Mangaroska, K., & Giannakos, M. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, *12*(4), 516–534. https://doi.org/10.1109/TLT.2018.2868673

Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics*, *12*(1), 12–31. https://doi.org/10.18608/jla.2025.8591

National Student Support Accelerator. (2023). Toolkit for tutoring programs. https://nssa.stanford.edu/tutoring

National Student Support Accelerator. (2025). Tutoring quality standards. https://nssa.stanford.edu/tqis/quality-standards

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

O'Neil, S., & Schotka, R. (2020). *CRLA ITTPC standards, outcomes, and assessments* (2nd ed.). College Reading and Learning Association. https://cdn.ymaws.com/crla.net/resource/resmgr/soas/crla_ittpc_standards_outcome.pdf

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *User modeling, adaptation, and personalization. UMAP 2010. Lecture notes in computer science* (pp. 255–266, Vol. 6075). Springer. https://doi.org/10.1007/978-3-642-13470-8_24

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis: A new alternative to knowledge tracing. *Frontiers in Artificial Intelligence and Applications*, *200*(1), 531–538. https://doi.org/10.3233/978-1-60750-028-5-531

Post, M. W. (2016). What to do with "moderate" reliability and validity coefficients? *Archives of Physical Medicine and Rehabilitation*, *97*(7), 1051–1052. https://doi.org/10.1016/j.apmr.2016.04.001

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Robinson, C. D. (2022). A framework for motivating teacher-student relationships. *Educational Psychology Review*, *34*(4), 2061–2094. https://doi.org/10.1007/s10648-022-09706-0

Saga Coach. (2025). Saga Coach Program. https://saga.org/products/saga-coach/

Saga Education. (2021, May). National tutoring nonprofit launches free, online training to help scale tutoring programs [press release]. https://saga.org/national-tutoring-nonprofit-launches-free-online-training-to-help-scale-tutoring-programs/

Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017). The proof of the pudding: Examining validity and reliability of the evaluation framework for learning analytics. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data driven approaches in digital education. EC-TEL 2017. Lecture notes in computer science* (pp. 194–208, Vol. 10474). Springer. https://doi.org/10.1007/978-3-319-66610-5_15

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 462–472). ACM. https://doi.org/10.1145/3706468.3706527

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., & Koedinger, K. (2023). When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (LAK 2023), 13–17 March 2023, Arlington, Texas, USA (pp. 250–261). ACM. https://doi.org/10.1145/3576050.3576089

Thomas, D. R., Borchers, C., Kakarla, S., Lin, J., Bhushan, S., Guo, B., Gatz, E., & Koedinger, K. R. (2025a). Do tutors learn from equity training and can generative AI assess it? In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 505–515). ACM. https://doi.org/10.1145/3706468.3706531

Thomas, D. R., Borchers, C., Kakarla, S., Lin, J., Bhushan, S., Guo, B., Gatz, E., & Koedinger, K. R. (2025b). Does multiple choice have a future in the age of generative AI? A posttest-only RCT. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 494–504). ACM. https://doi.org/10.1145/3706468.3706530

Walker, D. A. (2005). A comparison of the Spearman-Brown and Flanagan-Rulon formulas for split half reliability under various variance parameter conditions. *Journal of Modern Applied Statistical Methods*, *5*(2). https://doi.org/10.22237/jmasm/1162354620

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *NPJ Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01029-4

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. *Research Square preprint*. https://doi.org/10.21203/rs.3.rs-5363154/v1

Warrens, M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. In R. Millsap, D. Bolt, L. van der Ark, & W. Wang (Eds.), *Quantitative psychology research. Springer proceedings in mathematics & statistics* (pp. 293–300, Vol. 89). Springer. https://doi.org/10.1007/978-3-319-07503-7_18

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.52202/068431-1800

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457. https://doi.org/10.1016/j.chb.2020.106457

Wise, A. F., Knight, S., & Ochoa, X. (2021). What makes learning analytics research matter. *Journal of Learning Analytics*, *8*(3), 1–9. https://doi.org/10.18608/jla.2021.7647

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRater^{SM} v1.0. *ETS Research Report Series*, *2008*(2), i–102. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 101–111). ACM. https://doi.org/10.1145/3636555.3636856

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

Yoon, S.-Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, *93*, 43–52. https://doi.org/10.1016/j.specom.2017.08.001

Zhang, D.-W., Boey, M., Tan, Y. Y., & Jia, A. H. S. (2024). Evaluating large language models for criterion-based grading from agreement to consistency. *NPJ Science of Learning*, *9*(1), 79. https://doi.org/10.1038/s41539-024-00291-1

Zhang, L., Lin, J., Borchers, C., Cao, M., & Hu, X. (2024). 3DG: A framework for using generative AI for handling sparse learner performance data from intelligent tutoring systems. *arXiv preprint arXiv:2402.01746*. https://doi.org/10.48550/arXiv.2402.01746