

# A Validity Argument Framework for Automated Item Generation Systems

**Euigyum Kim**  
Sogang University  
Seoul, Republic of Korea  
egkim@sogang.ac.kr

**Hyo Jeong Shin**  
Sogang University  
Seoul, Republic of Korea  
hshinedu@sogang.ac.kr

**Alina A. von Davier**  
Duolingo  
Pittsburgh, PA, USA  
avondavier@duolingo.com

## Abstract

We introduce an evidence-centered, argument-based framework for validating AI-generated assessment items. The framework organizes three claims—content comparability, construct validity, and psychometric functioning—with assumptions and evidence. Applying the framework to critical thinking assessments, we compared AI-generated and human-authored items and identified distinct mechanisms underlying the validity of AI-generated items.

## 1 Introduction

Amid the increasing adoption of generative artificial intelligence (AI) in education, the human-in-the-loop (HITL) strategy has been emphasized to address the inherent limitations of AI (Chen et al., 2023). While early HITL research has focused on human roles in monitoring and refining AI-generated output (Memarian and Doleck, 2024), recent discussion has expanded to a human-centered AI perspective (Shneiderman, 2022) that centers human judgment and agency throughout the process of designing, developing, and implementing generative AI systems (von Davier and Burstein, 2024). Accordingly, proactive human monitoring, verification, and responsible use are considered essential throughout the development and use of AI-generated educational content to ensure pedagogical appropriateness, validity, and ethical safety (Kowal et al., 2025; Miao and Holmes, 2023).

Among various applications of AI in education, automated item generation (AIG) is a prominent area of interest in educational measurement research (Graf et al., 2026; von Davier, 2018). In prior studies on AIG, a rigorous HITL process, particularly through expert engagement, is regarded as essential to support the reliability and validity of AI-generated items (Attali et al., 2022; Kim et al., 2025b; Kowal et al., 2025). Moreover, recent studies perform psychometric analysis to in-

vestigate whether AI-generated items can demonstrate psychometric properties comparable to those of human-authored items (HAIs) (Bhandari et al., 2024; Shin et al., 2025). Collectively, many studies continue to use expert judgment and psychometric evaluation, either separately or in combination, as primary sources of validity evidence.

However, given the argument-based view of validity by Kane (1992, 2001, 2013), the validity of AI-generated items cannot be justified solely on the basis of human judgments or psychometric indicators. Rather, we argue that the validity of AI-generated items should be established through an integrated framework that demonstrates their structural, content-related, and functional comparability to rigorously validated HAIs within a coherent chain of claims and supporting evidence.

Therefore, using a critical thinking (CT) assessment as a case study, we propose an evidence-centered, argument-based framework for evaluating the validity of AIG systems through comparison with HAIs. By integrating three claims, their corresponding assumptions, and supporting evidence, the framework examines whether AI-generated items can provide sufficient validity evidence prior to operational administration. More specifically, the three claims address content comparability, construct validity, and psychometric functioning, and are operationalized through computational linguistic analysis, qualitative expert review, and quantitative psychometric analyses.

## 2 Theoretical Background

### 2.1 Validity Argument

Validity is not an inherent property of a test, but a dimension of interpretation constructed through argumentation (Cook et al., 2015; Kane, 1992; Messick, 1989). Validation, in turn, is the process of rigorously substantiating a specific decision by articulating claims and assumptions, and supporting

them through an empirical evidence-based verification process (Kane, 2013).

Specifically, Cook et al. (2015) comprehensively reviewed Kane's works and emphasized that validation requires an evaluation of coherence and completeness within an interpretive framework. They argued that the four steps of scoring, generalization, extrapolation, and implications can be substantiated through the inference and interpretation of various forms of evidence (Cook et al., 2015).

As a notable example of applying a validity argument, Jin et al. (2024) proposed a systematic validation framework for science learning progression (LP), evaluating whether LPs can coherently align curriculum, instruction, and assessment. Jin et al. (2024) organized three claims about the LP itself, the student's proficiency measured by the LP, and the instructional use of the LP and the associated test. Each claim was linked to explicit assumptions and evaluated using theoretical rationales and empirical evidence. During the validation process, Jin et al. (2024) revealed limitations in discriminating high-performing students, highlighting that validation is not a one-time confirmation but an iterative process of refinement.

Likewise, the validation process extends beyond simple evidence-based verification to include integrative interpretations within broader analytical processes (Kane, 2001, 2013), and such an approach also provides opportunities to strengthen validity by identifying critical limitations and sources of misalignment (Jin et al., 2024). This perspective has important implications for AIG research, suggesting that the validity of AI-generated items must be established through an argument-based framework, wherein intended interpretations and uses are articulated as explicit claims and systematically supported by empirical evidence.

## 2.2 Validity Research in AI-Driven Evaluation

Van Moere and Downey (2016) noted that existing validity arguments should be extended to address AI-driven assessments while accounting for the distinctive properties of AI. With the rapid advancement of generative AI, these concerns have emerged as a major agenda in AI-based educational evaluation systems (Hao et al., 2024).

Notably, Burstein et al. (2022) extended traditional validity frameworks into digital contexts by proposing the digitally-informed chain of inferences (DCI) model. The DCI builds on the six-step inferential chain including digital-specific layers,

thus facilitating traditional extrapolation methods to address novel questions that are particularly relevant to AI-driven assessment (Burstein et al., 2022).

Similarly, Care and Maddox (2021) broadened the scope of validity arguments by emphasizing the importance of test-taker experience. They noted that rigid standardization, while often framed as a guardrail of fairness, may unintentionally introduce construct-irrelevant variance and disproportionately disadvantage diverse learners, especially those with disabilities (Care and Maddox, 2021).

Although applications of validity arguments to AIG itself remain limited, prior studies collectively suggest the importance of establishing validity while considering the distinctive characteristics of generative AI. Rather than being established in isolation, the validity of AI-generated items needs to be examined in comparison with HAIs, particularly as recent approaches increasingly incorporate HITL and human-AI collaboration to leverage the complementary strengths of humans and AI (Kim et al., 2025b; Kowal et al., 2025). Such comparison, in turn, should be conducted through a systematic argumentation process that articulates validity claims and the evidence supporting them. Therefore, building on prior research, this study aims to broaden current discussions on the validity of AI-based evaluations by comparing AI-generated items to HAIs.

## 3 Elements of the Validity Framework

Drawing on the validity argument approach (Kane, 1992, 2013), this study conceptualizes a framework for evaluating the validity of AI-generated items through a sequential, cumulative, and integrated examination of their comparability with HAIs, as illustrated in Figure 1. The framework consists of three main claims, each supported by corresponding assumptions and evidence.

**Claim 1: AI-generated items should demonstrate content comparability.** The first claim concerns whether AI-generated items exhibit content characteristics sufficiently comparable to those of HAIs. Leveraging rigorously validated HAIs as a benchmark, we assume that AI-generated items do not exhibit significant differences in linguistic characteristics (**Assumption 1**) and semantic properties (**Assumption 2**) relative to HAIs. The assumptions can be evaluated through the quantitative analysis of computational linguistic indices (**Evidence 1**), semantic similarity measures

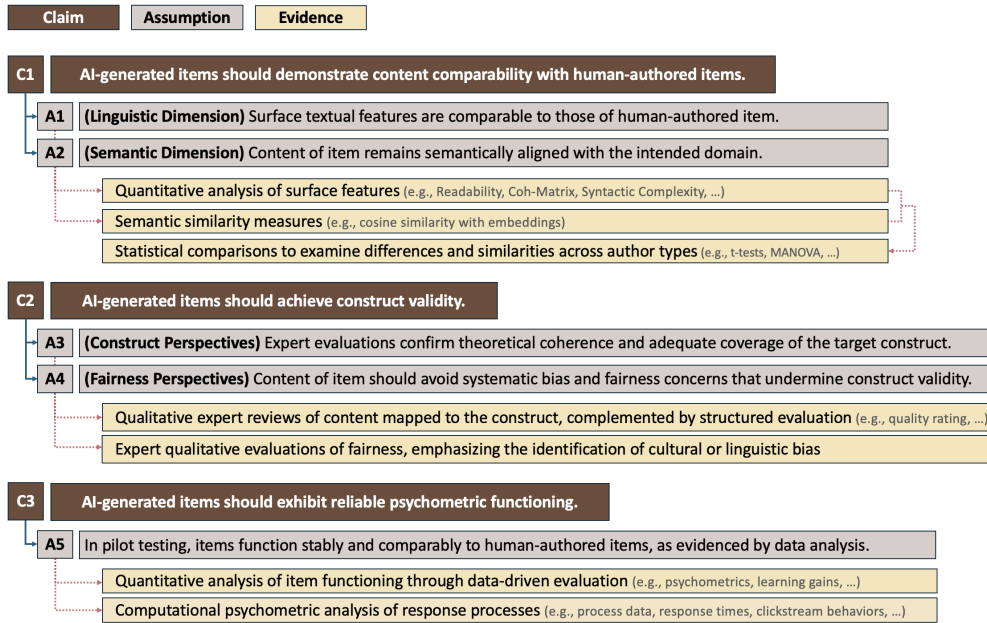


Figure 1: Proposed evidence-centered, argument-based validity framework for AIG systems.

(**Evidence 2**), and statistical comparisons between AI-generated items and HAIs (**Evidence 3**).

**Claim 2: AI-generated items should achieve construct validity.** The second claim concerns whether AI-generated items adequately represent the target construct. We assume that subject matter experts can verify that the items adequately represent the intended measurement construct (**Assumption 3**) and do not introduce structural biases that threaten validity (**Assumption 4**). These assumptions can be examined through expert reviews on the construct validity (**Evidence 4**), together with structured qualitative evaluations of potential sources of bias (**Evidence 5**).

**Claim 3: AI-generated items should exhibit reliable psychometric functioning.** For this claim, we assume that AI-generated items exhibit reliable psychometric properties and support interpretations comparable to those of HAIs through pilot testing (**Assumption 5**). This can be examined through psychometric modeling, specifically within an item response theory (IRT) framework (**Evidence 6**), or by directly analyzing the cognitive test-taking processes of examinees (**Evidence 7**).

Overall, the validity of AI-generated items can be articulated through an evidence-centered argument that evaluates their characteristics and functioning relative to rigorously developed HAIs. With this framework, we seek to not only illuminate the distinct attributes of AI-generated items but also

examine whether they can function as effectively as validated HAIs.

## 4 Research Methods

### 4.1 Dataset

To illustrate the practical application of the proposed framework, we analyzed CT assessment data from Shin et al. (2025), where CT items were developed by both human experts and AI. The data comprised 96 items administered to over 700 sixth-grade students, demonstrating high psychometric reliability. For this study, we selected a subset of 48 multiple-selection items (24 human-authored, 24 AI-generated), in which examinees judge each of four statements and responses are scored polytomously. Items were deliberately selected to have four human-authored and four AI-generated items in each of the six PACIER sub-skills, ensuring balanced comparison across development sources. An example CT assessment item is shown in Figure 2.

Figure 2: An illustrative example of a CT item.

Table 1: Description of computational linguistic and semantic indices with their computation methods.

| Feature                           | Description                                                                                                                                                                                                                      |
|-----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Root Type–Token Ratio (RTTR)      | Lexical diversity measured as the number of distinct word types divided by the square root of the total number of tokens (Guiraud, 1960), particularly useful for short texts. Higher values indicate greater lexical diversity. |
| Average Sentence Length (ASL)     | Average number of words per sentence (Kincaid et al., 1975). Higher values indicate longer sentences.                                                                                                                            |
| Maximum Dependency Depth (MDP)    | Maximum depth of syntactic dependency trees, reflecting syntactic complexity (Montani et al., 2023). Higher values indicate more complex structures.                                                                             |
| Lexical Density                   | Proportion of content words in a text (Biber, 1988). Higher values indicate denser informational content.                                                                                                                        |
| Dale–Chall Index                  | Readability measure based on the proportion of unfamiliar words and average sentence length (Chall and Dale, 1995). Higher scores correspond to greater reading difficulty and to higher grade-level bands.                      |
| Flesch–Kincaid Grade Level (FKGL) | Readability index expressed as required grade level for comprehension (Kincaid et al., 1975). Higher values indicate greater reading difficulty.                                                                                 |
| Sentiment Score                   | Overall sentiment polarity score ranging from $-1$ (negative) to $+1$ (positive), computed using the VADER sentiment analyzer (Hutto and Gilbert, 2014).                                                                         |
| Semantic Similarity               | Cosine similarity between sentence embeddings calculated using all-MiniLM-L6-v2 model, based on the Sentence-BERT framework, with higher values indicating greater semantic alignment (Reimers and Gurevych, 2019).              |

Each item consisted of a passage, a question, and four statements. Students were required to read the passage and question and then apply their CT skills to determine which statement represented the most appropriate response. As the questions and options are consistent across all items, only the passages and statements were included in the analyses.

## 4.2 Linguistic and Semantic Feature Analysis

To systematically collect empirical evidence for **Claim 1**, we computed the linguistic features of all items. Building on prior research that employed linguistic and semantic properties as criteria for evaluating the quality of AI-generated items (Kim et al., 2025b), we selected indicators that capture surface-level linguistic structure and readability, semantic aspects of content, and affective characteristics. The descriptions of selected features are summarized in Table 1.

Furthermore, statistical analyses were conducted to compare these features with those of HAIs. Because linguistic features often exhibit non-normal distributions and the number of items per group was small, we employed the non-parametric two-sample Kolmogorov–Smirnov (K–S) test which provides a robust evaluation of group differences, making it particularly suitable for linguistic features that often exhibit non-normal distributions (Massey, 1951).

## 4.3 Human Expert Review

To establish empirical grounding for **Claim 2**, we conducted a rigorous qualitative evaluation involving two experts in CT. Consistent with prior research emphasizing the importance of expert involvement in AIG (von Davier and Burstein, 2024), we requested experts to independently review the AI-generated items using a structured rubric comprising three criteria adapted from Kim et al. (2025a): (i) alignment with the intended PACIER sub-skills, (ii) linguistic and content-related issues that could threaten construct validity, and (iii) potential cultural or linguistic biases that might undermine test fairness.

## 4.4 Item Parameter Estimation

In order to evaluate **Claim 3**, we analyzed pilot-test data with IRT models focusing on estimating item discrimination parameters. A concurrent calibration (von Davier and von Davier, 2007) was conducted for all 96 items using the Generalized Partial Credit Model (Muraki, 1992) placing AI-generated items and HAIs on a common scale and thereby enabling direct comparison of item parameters across the two development sources. Although our analysis focused on a subset of items (48 among the 96 items), scale comparability was ensured through concurrent calibration of the full item set.

## 4.5 Mixed-Effects Linear Modeling

To integrate the validity evidence gathered from the content (**Claim 1**) and psychometric (**Claim 3**) analyses, we employed generalized mixed-effects linear models (Rabe-Hesketh et al., 2004) to examine how linguistic features of item content contribute to item discrimination. The 48 items were balanced across the six sub-skills of the PACIER framework (8 items per sub-skill) (Shin et al., 2025). As a baseline model, we specified author type (human vs. AI) as a fixed effect and PACIER sub-skills as a random intercept to account for domain clustering. We then estimated the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979), which provides an estimate of the proportion of variance in item discrimination attributable to differences among PACIER.

Next, we specified a more comprehensive multi-level model as a main model in which item discrimination was regressed on author type, a set of linguistic predictors, and their interactions with author type, while still accounting for PACIER domain as a random effect. This specification enables us to evaluate the main effects of linguistic characteristics of item content and whether their associations with discrimination varied across author types. This main model is given in Equation 1.

$$\text{Disc}_{ij} = \beta_0 + \beta_1 A_i + \sum_{k=1}^K \beta_k X_{ik} + \sum_{k=1}^K \beta_k^{(A)} (A_i \times X_{ik}) + u_{0j} + \varepsilon_{ij} \quad (1)$$

where  $\text{Disc}_{ij}$  denotes the discrimination parameter of item  $i$  in PACIER domain  $j$ ,  $A_i$  indicates author type (0 = AI, 1 = Human), and  $X_{ik}$  represents the  $k$ -th linguistic feature. The coefficients  $\beta_0$ ,  $\beta_1$ ,  $\beta_k$ , and  $\beta_k^{(A)}$  denote the intercept, the main effect of author type, the fixed effects of linguistic features, and their interactions with author type, respectively. Random intercepts for each PACIER domain are denoted by  $u_{0j}$ , and residual errors by  $\varepsilon_{ij}$ .

## 5 Results

### 5.1 Linguistic and Semantic Feature Analysis

To examine **Claim 1**, we compared the linguistic features of AI-generated items and HAIs. First, as summarized in Table 2, the K-S test results revealed statistically significant differences at the passage level in RTTR, MDP, and lexical density.

A similar pattern was observed in the statements that AI-generated statements differed significantly in RTTR, ASL, MDP, and lexical density.

Table 2: Results of K-S tests for linguistic features comparing HAIs ( $N=24$ ) and AI-generated items ( $N=24$ ).

| Level     | Feature             | $D$    | $p$     | Sig. |
|-----------|---------------------|--------|---------|------|
| Passage   | RTTR                | 0.6250 | < .0001 | ***  |
|           | ASL                 | 0.2917 | 0.2469  |      |
|           | MDP                 | 0.4167 | 0.0129  | *    |
|           | Lexical Density     | 0.4583 | 0.0114  | *    |
|           | Dale-Chall Index    | 0.3333 | 0.1332  |      |
|           | FKGL                | 0.1667 | 0.8525  |      |
|           | Sentiment Score     | 0.3750 | 0.0638  | .    |
|           | Semantic Similarity | 0.3333 | 0.1398  |      |
| Statement | RTTR                | 0.7917 | < .0001 | ***  |
|           | ASL                 | 0.7500 | < .0001 | ***  |
|           | MDP                 | 0.5417 | 0.0005  | ***  |
|           | Lexical Density     | 0.4167 | 0.0299  | *    |
|           | Dale-Chall Index    | 0.2500 | 0.4490  |      |
|           | FKGL                | 0.2917 | 0.2599  |      |
|           | Sentiment Score     | 0.2917 | 0.2628  |      |
|           | Semantic Similarity | 0.2500 | 0.4490  |      |

Note. \*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$ ; .  $p < .10$ .

Subsequent analysis focused on comparing the values of linguistic features that exhibited statistically significant differences, as reported in Table 3.

Table 3: Descriptive statistics of linguistic features showing significant differences by author type.

| Level     | Feature         | Group | M     | SD   | Min  | Max   |
|-----------|-----------------|-------|-------|------|------|-------|
| Passage   | RTTR            | AI    | 5.99  | 0.81 | 4.25 | 7.18  |
|           |                 | Human | 5.01  | 0.92 | 3.33 | 6.87  |
|           | MDP             | AI    | 6.79  | 1.72 | 4.00 | 12.00 |
|           |                 | Human | 5.71  | 1.78 | 4.00 | 12.00 |
|           | Lexical Density | AI    | 0.50  | 0.05 | 0.41 | 0.58  |
|           |                 | Human | 0.46  | 0.06 | 0.35 | 0.62  |
| Statement | RTTR            | AI    | 3.59  | 0.26 | 2.95 | 3.97  |
|           |                 | Human | 2.93  | 0.32 | 2.45 | 3.61  |
|           | ASL             | AI    | 14.04 | 2.24 | 8.75 | 17.50 |
|           |                 | Human | 9.51  | 2.23 | 6.50 | 14.75 |
|           | MDP             | AI    | 4.55  | 0.80 | 3.00 | 6.50  |
|           |                 | Human | 3.69  | 0.76 | 2.50 | 5.50  |
|           | Lexical Density | AI    | 0.56  | 0.06 | 0.40 | 0.69  |
|           |                 | Human | 0.53  | 0.09 | 0.38 | 0.70  |

**Assumption 1** was only partially supported: the results indicate that AI-generated items are not distributionally identical to HAIs across several linguistic indices, showing higher lexical diversity, longer sentences with complex grammatical structure, and higher lexical density. In contrast, **Assumption 2** was supported, as no statistically significant differences were found in semantic similarity or readability indices.

Overall, for **Claim 1**, the results indicate that AI-generated items exhibit stylistic differences from

HAI while maintaining comparable semantic content and grade-level readability. Thus, these features should be viewed as structured stylistic tendencies of AI-generated outputs, warranting further evaluation as the argument progresses.

## 5.2 Construct Validity Review by Experts

Two subject matter experts in CT education evaluated the AI-generated items based on the criteria of **Claim 2**. While the expert review identified several concerns, the reviewers concluded that the items were suitable for use without substantial revision.

First, regarding **Assumption 3**, experts noted that while the items generally mapped to the target construct, certain items sometimes exhibited syntactic complexity that may exceed the intended grade level. However, consistent with the findings of **Claim 1**, these were interpreted as acceptable stylistic tendencies rather than severe failures in construct representation.

Second, with respect to **Assumption 4**, the review confirmed that the majority of items were free from critical bias or fairness concerns. Even though experts noted that some passages tended to emphasize a single perspective, this was deemed acceptable in the context of assessing CT, where the focus lies on the critical and evaluative analysis of the provided information. Overall, all AI-generated items were found to adequately represent the intended construct, thus supporting **Claim 2**. Note that because the HAIs had been validated through prior expert review and pilot testing (Shin et al., 2025), Claim 2 in this case study rests on expert judgment of the AI-generated items against that established benchmark rather than on a parallel within-study review.

## 5.3 Examining Item Functioning through Empirical Data Analysis

For the examination of **Claim 3**, we fit a series of mixed-effects linear models incorporating interactions between author type and linguistic features. To ensure model stability, several features were excluded due to multicollinearity (ASL for passage and statements; MDP for statements).

We first analyzed a baseline model with author type as the sole predictor and PACIER domain as a random intercept. The low ICC (.093) and insignificant fixed-effect of author type ( $b = -0.128, p = .080$ ) were identified, suggesting that author type does not sufficiently explain variance in item discrimination parameters regardless of PACIER sub-

skills. Given the small number of items, this result indicates no detected difference in mean discrimination across author types.

We further analyzed the main model including interactions with linguistic features. The PACIER random-intercept variance was essentially zero, and the model effectively reduced to a fixed-effects model. Regarding fixed effects, the set of predictors explained substantial variance ( $R^2 = .73$ ), suggesting that the interaction between linguistic properties and author type is a significant predictor of item discrimination.

As shown in Table 4, the main interaction model revealed that while both author types achieved similar discrimination levels, they relied on distinct mechanisms. For AI-generated items, higher lexical density ( $b = 0.138, p = .008$ ) and Dale–Chall scores ( $b = 0.093, p = .043$ ) were associated with greater item discrimination. In contrast, these relationships were inverted in HAIs, particularly regarding lexical density ( $b = -0.191, p = .007$ ) at the passage level. This divergence suggests that linguistic characteristics are associated with item discrimination in different ways across author types.

These findings provide support for **Claim 3** and **Assumption 5** by showing that even when AI-generated items diverge from human stylistic norms, they exhibit psychometric performance comparable to that of HAIs while exhibiting distinct linguistic–psychometric relationships. Overall, whereas the evidence for **Claim 1** and **Claim 2** alone would be insufficient to support the validity of AI-generated items, the evidence for **Claim 3** complements it by addressing their psychometric functioning.

## 6 Discussion

### 6.1 Key Findings and Implications

As Woodworth and Solas (2026) argued, the use of AI-generated items is fundamentally a measurement problem rather than merely a technological one. Despite this perspective, prior research on AIG has largely relied on fragmented sources of validity evidence such as expert judgments or psychometric performance. While informative, these indicators risk treating a limited set of “measurable signals” (Woodworth and Solas, 2026) as sufficient evidence of validity and may not, on their own, provide sufficient justification for the intended interpretations and uses of AI-generated items.

Alternatively, this study framed the comprehen-

Table 4: Fixed effects estimates from the mixed-effects main model of item discrimination.

| Predictor                                  | Estimate | Std. Error | <i>t</i> | <i>p</i> | Sig. |
|--------------------------------------------|----------|------------|----------|----------|------|
| (Intercept)                                | 0.667    | 0.065      | 10.29    | < .001   | ***  |
| Human (vs. AI)                             | -0.083   | 0.098      | -0.84    | .405     |      |
| RTTR                                       | 0.002    | 0.077      | 0.02     | .982     |      |
| MDP                                        | 0.037    | 0.062      | 0.61     | .548     |      |
| Lexical Density                            | 0.138    | 0.050      | 2.76     | .008     | **   |
| Dale–Chall                                 | 0.112    | 0.066      | 1.69     | .098     | .    |
| FKGL                                       | -0.050   | 0.057      | -0.88    | .384     |      |
| Sentiment Score                            | -0.067   | 0.031      | -2.20    | .032     | *    |
| Semantic Sim. (passage–statements)         | -0.100   | 0.050      | -1.99    | .053     | .    |
| RTTR (statement)                           | -0.025   | 0.065      | -0.39    | .701     |      |
| Lexical Density (statement)                | -0.005   | 0.055      | -0.10    | .923     |      |
| Dale–Chall (statement)                     | 0.093    | 0.045      | 2.08     | .043     | *    |
| FKGL (statement)                           | 0.003    | 0.076      | 0.04     | .969     |      |
| Sentiment Score (statement)                | 0.031    | 0.036      | 0.87     | .390     |      |
| Semantic Sim. (statement)                  | -0.077   | 0.055      | -1.41    | .166     |      |
| Human × RTTR (passage)                     | -0.019   | 0.096      | -0.20    | .843     |      |
| Human × MDP (passage)                      | 0.019    | 0.079      | 0.24     | .815     |      |
| Human × Lexical Density (passage)          | -0.191   | 0.068      | -2.82    | .007     | **   |
| Human × Dale–Chall (passage)               | -0.070   | 0.076      | -0.93    | .359     |      |
| Human × FKGL (passage)                     | -0.121   | 0.079      | -1.54    | .131     |      |
| Human × Sentiment Score (passage)          | 0.074    | 0.063      | 1.18     | .244     |      |
| Human × Semantic Sim. (passage–statements) | 0.060    | 0.065      | 0.93     | .360     |      |
| Human × RTTR (statement)                   | 0.054    | 0.093      | 0.58     | .566     |      |
| Human × Lexical Density (statement)        | 0.038    | 0.067      | 0.57     | .574     |      |
| Human × Dale–Chall (statement)             | -0.147   | 0.073      | -2.01    | .050     |      |
| Human × FKGL (statement)                   | 0.148    | 0.092      | 1.60     | .117     |      |
| Human × Sentiment Score (statement)        | 0.002    | 0.063      | 0.04     | .972     |      |
| Human × Semantic Sim. (statement)          | 0.116    | 0.066      | 1.76     | .086     | .    |

Note. \*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$ ; .  $p < .10$

sive validation of AIG systems as an argumentation process and introduced an evidence-centered validity framework that evaluates AI-generated items in relation to so-called gold-standard HAIs. With the framework, we examined whether AI-generated items can establish sufficient validity evidence prior to operational administration by integrating computational linguistic evidence, subject matter expert judgments, and psychometric functioning indices derived from pilot-test data.

The application of this framework led to two key conclusions. First, AI-generated items exhibited linguistic characteristics that differed from those of HAIs. Although these linguistic differences might have altered the test-taking experience, they did not result in semantic divergence or introduce psychometric instability into item performance. Specifically, AI-generated items showed greater lexical diversity, linguistic complexity, and lexical density, with lexical density positively associated with item discrimination. In contrast, this positive association was not observed for HAIs, suggesting that AI-generated items may achieve comparable psychometric functioning through different linguistic–psychometric association patterns.

One implication is that automated item parameter prediction models trained on HAI pools may not transfer directly to AI-generated pools, which warrants investigation in future research. Relatedly, features such as lexical density that are associated with item discrimination could be treated as design targets during generation rather than as characteristics to be filtered out afterward.

Second, our findings underscore the need to integrate multiple sources of validity evidence. Although experts identified distinctive stylistic tendencies and construct-related concerns in AI-generated items, these characteristics were not necessarily indicators of poor item quality, but were associated with effective psychometric functioning. This further suggests that systematic comparisons with HAIs may provide a more informative basis for establishing the validity of AIG systems.

In summary, rather than claiming that AI-generated items are or can be valid, this study addresses the fundamental question of how their validity can be established, offering a rigorous methodological framework for accumulating and

integrating diverse validity evidence. From this perspective, the validation of AIG systems must move beyond superficial concerns regarding how closely AI-generated items resemble traditionally validated HAIs, or how extensively human experts intervene in their generation pipeline. Instead, the focus should be directed toward whether the resulting items substantively support the intended interpretations and operational uses of assessment scores, thus strengthening the validity of AIG.

Overall, our findings suggest that establishing the validity of AI-generated items through the systematic accumulation and integration of evidence represents an important step toward human-centered AI, where human expertise is not merely controlling AI outputs, but critically evaluating and arguing their meanings in educational assessment.

## 6.2 Limitations and Future Directions

Despite the implications of this study, several limitations should be acknowledged. First, the small number of items restricts the extent to which the findings can be generalized. Although nonparametric methods were used to avoid distributional assumptions, the small item pool limits statistical power for group comparisons and means that the main mixed-effects model should be regarded as exploratory. Further validation of the framework across assessment domains with larger item pools would strengthen the evidence for each claim.

Additionally, evidence for **Claim 2** in this case study was qualitative and limited to the AI-generated items, with the HAIs serving as a previously validated benchmark. Quantitative construct validity evidence across author types would provide stronger support for the construct and psychometric claims. In this regard, future work should also examine other components of the framework not addressed in this case study, including extrapolation inferences and analyses using computational psychometrics.

Furthermore, because the proposed framework is applied to CT assessments as an illustrative case study, it does not prescribe universal quantitative thresholds, selection criteria, or rigid decision rules for retaining or excluding AI-generated items. Consequently, it would be interesting to further investigate how these proposed evidence-centered, argument-based evaluations can be translated into practical, domain-specific decision-making criteria for operational assessment contexts across diverse measurement constructs.

Similarly, our framework can be examined across a broader range of measurement and evaluation contexts. As Kane (1992, 2001, 2013) suggested, the structure and supporting evidence of a validity argument depend on the specific interpretive and use claims being made. Therefore, future research should explore how the framework can be adapted to incorporate context-specific claims and evidence, particularly regarding downstream issues such as fairness, consequential validity, and operational performance in large-scale assessments.

Ultimately, the contribution of this study lies in making the validation of AIG items and systems more actionable, shifting the focus from fragmented sources of validity evidence to their stepwise integration within a coherent validity argument. We hope that the proposed framework will continue to be refined, extended, and advanced, while fostering both theoretical and practical efforts to establish evidence-centered validity in AI-based educational measurement and evaluation.

## Declaration on Generative AI

The authors used *Writefull* via Overleaf and *ChatGPT* only for language editing to check grammar and improve readability. The authors reviewed and edited all AI-generated suggestions, and take full responsibility for this manuscript.

## References

- Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. [The interactive reading task: Transformer-based automatic item generation](#). *Frontiers in Artificial Intelligence*, 5.
- Shreya Bhandari, Yunting Liu, Yerin Kwak, and Zachary A. Pardos. 2024. [Evaluating the psychometric properties of ChatGPT-generated questions](#). *Computers and Education: Artificial Intelligence*, 7:100284.
- Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.
- Jill Burstein, Geoffrey T. LaFlair, Antony J. Kunnan, and Alina A. von Davier. 2022. [A theoretical assessment ecosystem for a digital-first assessment—the Duolingo English Test](#). Research report, Duolingo English Test.
- Naomi Care and Bryan Maddox. 2021. [Improving test validity and accessibility with Digital-First assessments](#). Technical report, Duolingo, Inc.

- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.
- Shengnan Chen, Qifang Liu, and Bin He. 2023. *A generative AI-based teaching material system using a human-in-the-loop model*. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–8, Wuhan, China. IEEE.
- David A. Cook, Ryan Brydges, Shiphra Ginsburg, and Rose Hatala. 2015. *A contemporary approach to validity arguments: A practical guide to Kane's framework*. *Medical Education*, 49(6):560–575.
- Edith Aurora Graf, Hyo Jeong Shin, Duanli Yan, and Alina A. von Davier. 2026. *Automated item generation: The promises and the challenges*. In *Reference Module in Social Sciences*. Elsevier.
- Pierre Guiraud. 1960. *Problèmes et méthodes de la statistique linguistique*. Presses Universitaires de France.
- Jiangang Hao, Alina A. von Davier, Victoria Yaneva, Susan Lottridge, Matthias von Davier, and Deborah J. Harris. 2024. *Transforming assessment: The impacts and implications of large language models and generative AI*. *Educational Measurement: Issues and Practice*, 43(2):16–29.
- Clayton J. Hutto and Eric Gilbert. 2014. *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225.
- Hui Jin, Hyo Jeong Shin, and Dante Cisterna. 2024. *Systematic validation in science learning progression research*. *International Journal of Science and Mathematics Education*, 22:189–209.
- Michael T. Kane. 1992. *An argument-based approach to validity*. *Psychological Bulletin*, 112(3):527–535.
- Michael T. Kane. 2001. *Current concerns in validity theory*. *Journal of Educational Measurement*, 38(4):319–342.
- Michael T. Kane. 2013. *Validating the interpretations and uses of test scores*. *Journal of Educational Measurement*, 50(1):1–73.
- Euigyum Kim, Salah Khalil, and Hyo Jeong Shin. 2025a. *Comparing human and LLM evaluations on AI-generated critical thinking items: Implications for valid applications of automatic item generation*. In *Proceedings of the 2nd Workshop on Automated Evaluation of Learning and Assessment Content (Eval-LAC 2025)*, volume 4006 of *CEUR Workshop Proceedings*. CEUR-WS.
- Euigyum Kim, Seewoo Li, Salah Khalil, and Hyo Jeong Shin. 2025b. *STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment*. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 920–930, Vienna, Austria. Association for Computational Linguistics.
- J. Peter Kincaid, Robert P. Fishburne, Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Research Branch Report 8-75, Chief of Naval Technical Training, Millington, TN.
- Jaclyn Martin Kowal, Kenzie Hurley Bryant, Dan Segall, and Tracy Kantrowitz. 2025. *Harnessing generative AI for assessment item development: Comparing AI-generated and human-authored items*. *International Journal of Selection and Assessment*, 33(3):e70021.
- Frank J. Massey. 1951. *The Kolmogorov-Smirnov test for goodness of fit*. *Journal of the American Statistical Association*, 46(253):68–78.
- Bahar Memarian and Tenzin Doleck. 2024. *Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis*. *Computers in Human Behavior: Artificial Humans*, 2(1):100053.
- Samuel Messick. 1989. *Validity*. In Robert L. Linn, editor, *Educational Measurement*, 3rd edition, pages 13–103. American Council on Education and Macmillan, New York, NY.
- Fengchun Miao and Wayne Holmes. 2023. *Guidance for Generative AI in Education and Research*. UNESCO Publishing, Paris, France.
- Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. *explosion/spaCy: v3.7.2: Fixes for APIs and requirements*. Computer software, Version v3.7.2.
- Eiji Muraki. 1992. *A generalized partial credit model: Application of an EM algorithm*. *Applied psychological measurement*, 16(2):159–176.
- Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. 2004. *Generalized multilevel structural equation modeling*. *Psychometrika*, 69(2):167–190.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence embeddings using siamese BERT-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.
- Hyo Jeong Shin, Seewoo Li, Ji Hoon Ryoo, Alina von Davier, Todd Lubart, and Salah Khalil. 2025. *The nature and measure of critical thinking: The PACIER framework and assessment*. *Journal of Intelligence*, 13(9):113.
- Ben Shneiderman. 2022. *Human-Centered AI*. Oxford University Press.

- Patrick E. Shrout and Joseph L. Fleiss. 1979. [Intra-class correlations: Uses in assessing rater reliability](#). *Psychological Bulletin*, 86(2):420–428.
- Alistair Van Moere and Ryan Downey. 2016. [Technology and artificial intelligence in language assessment](#). In Dina Tsagari and Jayanti Banerjee, editors, *Handbook of Second Language Assessment*, pages 341–358. De Gruyter Mouton, Berlin, Boston.
- Alina A. von Davier and Jill Burstein. 2024. [AI in the assessment ecosystem: A human-centered AI perspective](#). In Peter Ilic, Imogen Casebourne, and Rupert Wegerif, editors, *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, pages 93–109. Springer Nature Switzerland, Cham.
- Matthias von Davier. 2018. [Automated item generation with recurrent neural networks](#). *Psychometrika*, 83(4):847–857.
- Matthias von Davier and Alina A. von Davier. 2007. [A unified approach to IRT scale linking and scale transformations](#). *Methodology*, 3(3):115–124.
- Johanathan Woodworth and Eddia Solas. 2026. [When validity becomes the measurement problem: Generative AI and the fragility of assessment reasoning](#). *Postdigital Science and Education*.