TPM Vol. 32, No. S2, 2025
ISSN: 1972-6325
https://www.tpmap.org/

ASSESSING AI-GENERATED MATH ITEMS: EVIDENCE FROM EIGHTH-GRADE TRIANGLE CONGRUENCE

MAJED M. ALJODEH
DEPARTMENT OF EDUCATION AND PSYCHOLOGY, UNIVERSITY OF TABUK, TABUK, SAUDI ARABIA
EMAIL: majed_jodeh@hotmail.com, ORCID: HTTPS://ORCID.ORG/0009-0003-1530-930X

Abstract. The rapid advancement of generative AI has prompted new inquiries into its role in educational assessment. This study evaluates the quality of a mathematics achievement test on triangle congruence generated by ChatGPT, aligned with the eighth-grade Jordanian curriculum. The objective was to assess the validity, curricular relevance, clarity, and grade-level appropriateness of AI-generated test items. A quantitative design was employed, with 72 educational professionals including supervisors and mathematics teachers rating 20 multiple-choice items using a five-point Likert scale across four evaluation criteria. Findings showed that 80% of the items met or exceeded expert expectations, 85% scored as having high content validity, and 75% matched the expected difficulty level concerning cognitive demands. However, three items raised red flags on language clarity or contextual drift, indicating an ongoing requirement for human moderation in educational contexts that are specific to culture. These findings suggest that human-AI co-design models can enhance assessment efficiency while safeguarding pedagogical standards. Implications for integrating generative AI into curriculum-based assessment frameworks and recommendations for educator training, ethical oversight, and prompt refinement are discussed.

Keywords: AI-generated assessment; ChatGPT; curriculum alignment; human-AI collaboration; mathematics education.

INTRODUCTION

The integration of artificial intelligence (AI) into educational systems marks a critical turning point in the evolution of teaching, learning, and assessment design. No longer confined to peripheral tasks, AI technologies now underpin adaptive learning platforms, intelligent tutoring systems, automated grading, and increasingly, the autonomous generation of assessment content. Among the most transformative developments is the rise of generative language models, such as OpenAI's ChatGPT, which offer rapid and linguistically fluent item creation potentially alleviating the time-intensive burden of test construction for educators. Yet, this growing utility raises foundational concerns regarding construct validity, cognitive alignment, and cultural appropriateness in context-specific educational settings (Messick, 1995).

While generative AI systems can mimic the syntactic structures of human-authored items, assessment quality requires more than linguistic fluency. Effective assessment design must align with intended learning outcomes, reflect domain-specific content knowledge, and scaffold cognitive complexity. These challenges are particularly pronounced in mathematics education, where standardized assessments are expected to probe not only procedural fluency but also higher-order reasoning, spatial visualization, and abstraction (NCTM, 2014; Bloom, 1956). As such, mathematics offers a fertile domain to examine the capabilities and limitations of AI in generating valid and developmentally appropriate test items.

In localized educational systems such as Jordan's, assessment practices are governed by centralized standards that encode both pedagogical expectations and sociocultural values. Test items must reflect curricular objectives, language norms or terminologies in the area and cultural representations. AI models primarily trained with Western or generalized corpora may hence generate items that, though structurally valid, may be linguistically unfamiliar or conceptually different from local curricula (Zhai, 2022; Abedi, 2006). Such disjunctures may question the fairness of the test and coherence with instruction and accentuate achievement gaps when in high-stakes assessment settings.

In that respect, scholars have emphasized the value of responsible AI in education, noting the need for transparent validation procedures, human oversight, and application with consciousness of equity (Holmes et al., 2021; Luckin et al., 2016). Accordingly, the assessment forms generated by AI should be empirically scrutinized on a rigorous basis so as to meet the standards of validity, reliability, and fairness, as well as being culturally and curricular authentic.

This study erstwhile addresses this growing challenge by empirically analyzing the quality of a mathematics achievement test generated by ChatGPT on the concept of triangle congruence. The twenty multiple-choice items in the test have been judged by seventy-two mathematics teachers and supervisors in terms of four

essential dimensions: item clarity, curricular relevance, developmental suitability, and cognitive difficulty. The purpose is not only to determine whether the AI-generated items are technically sound but whether they are pedagogically meaningful and culturally appropriate for the intended context.

More specifically, the study sets out to answer the following research questions:

To what extent do expert evaluators judge the AI-generated test items as clear, curriculum-aligned, and suitable for eighth-grade learners in Jordan?

What is the perceived difficulty level of the items, and how does this perception vary among evaluator subgroups?

By promoting local educator expertise and contextual relevance, the investigation contributes to an increasingly relevant discourse on human-AI collaboration in assessment design, putting forth empirical findings and practical implications. These grouped results will potentially assist in designing future strategies for AI integration within different national education systems, especially in situations where curricular alignment, cultural fidelity, and assessment equity are paramount.

LITERATURE REVIEW

THE INTEGRATION OF ARTIFICIAL INTELLIGENCE IN EDUCATIONAL ASSESSMENT

The incorporation of Artificial Intelligence (AI) into educational assessment has catalyzed a shift in the design, administration, and analysis of assessments. Nowadays, AI tools are used to automate grading, generate assessment content, and offer personalized feedback functions that were traditionally performed by human teachers (Luckin et al., 2016; Holmes et al., 2021). This literature review seeks to provide a critical synthesis of recent research on AI in assessment and focuses on the four main areas: psychometric validity, curriculum alignment, cultural sensitivity, and ethical scrutiny.

THE EMERGENCE OF AI IN EDUCATIONAL ASSESSMENT

The use of AI technologies such as automated essay scoring (AES) and item generation systems have radically altered the landscape of summative and formative assessments. AES evaluates parameters such as coherence, effectiveness of argument, and grammar on platforms like e-rater and IntelliMetric (Shermis & Burstein, 2013). Apart from automated essay scoring, a large number of generative language models- e.g. GPT-3 and GPT-4- can be completely independent in generating MCQs and math problems now (Kasneci et al., 2023). They raise questions, however, concerning their educational quality and whether questions correlate to learning outcomes, although their production of items would be syntactically fluent.

ENSURING VALIDITY AND RELIABILITY IN AI-GENERATED ASSESSMENTS

Particularly with respect to construct validity and reliability, one of the foremost concerns surrounding the implementation of assessments generated by AI is psychometric rigor. Messick (1995) stated that validity would not be a property of the test but of the inferences raised from test scores. Mislevy et al. (2003) later extended this in the context of technology-enhanced assessments, urging alignment between cognitive models, task design, and interpretation.

Recent studies have noted that while LLMs like ChatGPT can simulate item structures, they often fail to maintain fidelity to the intended construct (Zhai, 2022). For example, Lai et al. (2023) found that AI-generated mathematics questions, though grammatically correct, sometimes introduced conceptual distortions or lacked cognitive depth. Hybrid validation approaches combining automated generation with expert review have therefore been recommended as best practice (Chatterji & Kar, 2023).

CURRICULUM ALIGNMENT IN AI-DRIVEN ASSESSMENTS

Curriculum alignment refers to the degree to which assessment items reflect prescribed learning objectives. Poor alignment can result in construct underrepresentation or construct-irrelevant variance (Pellegrino et al., 2001). From research comes their enforceability of domain-specific ontologies for AI systems in curricular fidelity. In mathematics education, this means fidelity to content strands (geometry, number theory, etc.) and cognitive processes like abstraction and proof (NCTM, 2014). Collaboration with content experts is crucial to minimizing misalignment (Holmes et al., 2021).

CULTURAL RELEVANCE AND SENSITIVITY IN AI ASSESSMENTS

Culturally responsive assessment design is vital in multilingual or multicultural contexts. AI models typically trained on generalized, English-centric datasets risk embedding Western-centric norms and idioms into item

content, reducing accessibility for learners from other regions (Sambasivan et al., 2021). Research has shown that assessments which fail to consider students' linguistic and cultural backgrounds can lead to lower engagement and misinterpretation of tasks (Abedi, 2006)

In Order to counter such challenges, models should be trained using region-specific corpora, and testing should also involve teachers from target populations. This principle is especially valid within the context of Jordan, where national standards emphasize Arabic terminology, symbols relevant to the local cultural context, and examples rooted in the culture.

ETHICAL CONSIDERATIONS AND HUMAN OVERSIGHT

The ethical integration of AI into education requires safeguards on transparency, accountability, and mitigation of bias. Key frameworks, especially the principles of Responsible AI, indicate that explainability and human oversight in high-stakes applications are crucial (Holmes et al., 2021; Cowie et al., 2023). AI-generated assessments may unintentionally reinforce biased training data and deepen inequalities if used without scrutiny.

Consequently, most researchers advocate for co-piloted systems, wherein AI generates initial drafts and human educators refine and validate the items (Luckin et al., 2016; Kasneci et al., 2023). This approach blends efficiency with contextual sensitivity.

FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

While AI's role in education continues to expand, several research avenues remain underexplored. There is a pressing need to:

- Evaluate the long-term cognitive and affective impact of AI-generated assessments;
- Investigate AI’s role in formative assessment and real-time feedback;
- Address equity concerns in the distribution of AI tools across resource-limited settings;
- Explore disciplinary boundaries, extending AI integration into non-STEM fields.

Moreover, longitudinal research is needed to understand how AI-informed assessment affects learning trajectories, motivation, and test-taker trust.

METHODS

This study employed a convergent parallel mixed methods design (Creswell & Plano Clark, 2017) to evaluate the psychometric quality and contextual relevance of AI-generated mathematics assessment items. The approach combined quantitative analysis of student performance and test metrics with qualitative expert evaluations, enabling triangulation between statistical item properties and practitioner insights.

RESEARCH DESIGN

The study integrated three complementary strands of analysis:

- Quantitative Item Validation: Statistical evaluation of AI-generated items using student performance metrics, including item difficulty and discrimination indices.
- Expert Judgment Study: Evaluation of item quality based on educator reviews using a structured rubric aligned with curricular standards.
- Qualitative Feedback Analysis: Collection and thematic coding of narrative comments from educators and students to interpret item clarity, relevance, and fairness.

This design supports instrumental convergence, allowing comparison of quantitative test statistics with professional pedagogical judgments.

DATA COLLECTION INSTRUMENTS

3.2.1 AI-Generated Achievement Test

A 20-item multiple-choice test was generated using OpenAI’s ChatGPT based on the Jordanian eighth-grade mathematics curriculum, specifically targeting the topic of triangle congruence. Each item contained four alternatives and covered key congruence criteria (SAS, ASA, SSS). Items were designed to vary in cognitive demand from basic identification to application and synthesis. While the source language was Arabic, items were translated into English for publication and cross-validation.

3.2.2 Test Judging Tool (for Expert Validation)

A structured evaluation form was developed based on prior frameworks in assessment design (Shute & Ventura, 2013; Mislevy et al., 2003). Experts rated each item on four dimensions using a 5-point Likert scale:

Clarity of wording
• Relevance to triangle congruence
• Grade-level appropriateness (eighth grade)
• Perceived difficulty
• The instrument was distributed via Google Forms to 72 participants (see Table 1). Judges also provided open-ended qualitative feedback.
Student Performance Assessment
To estimate item-level psychometrics, the AI-generated test was administered to 100 undergraduate students enrolled in a geometry course at Hashemite University. The test was administered in a controlled setting over 30 minutes. Students' item responses were used to compute:
• Item Difficulty Index (P-value): proportion answering correctly
• Discrimination Index (point-biserial correlation with total score)
• Test Reliability: Cronbach's alpha
Note: While the test was designed for Grade 8, undergraduate participants were used to simulate item difficulty and response behavior. This was acknowledged as a limitation, and items were cross-compared with human-generated counterparts for benchmarking. Ere applicable).

DATA ANALYSIS

3.3.1 Expert Validation Analysis
Descriptive statistics (means and standard deviations) were computed for each rubric criterion. Internal consistency of expert ratings was assessed using Cronbach's alpha, and Fleiss' kappa was calculated for inter-rater agreement where multiple raters reviewed the same items.

3.3.2 Analysis of Student Performance
Quantitative performance data were analyzed using SPSS v26:
• Item difficulty and discrimination indices were computed.
• Test reliability was evaluated using Cronbach's alpha.
• A comparative analysis between AI- and human-generated items was conducted using independent-samples t-tests.

3.3.3 Qualitative Feedback Analysis
Open-ended responses from experts and students were subjected to inductive thematic analysis (Braun & Clarke, 2006). Two coders independently reviewed responses, developed a coding scheme, and resolved discrepancies through consensus. Emerging themes related to linguistic ambiguity, conceptual alignment, and visual support in item design were synthesized.

3.3.4 Pilot Testing & Refinements
A pilot study involving 10 mathematics educators was conducted to pre-test the AI-generated items. Feedback on clarity, length, and content alignment led to minor item revisions prior to full deployment. This pilot phase enhanced both face validity and linguistic accuracy.

TABLE 1. DISTRIBUTION OF THE SAMPLE OF JUDGES BASED ON THEIR DEMOGRAPHIC CHARACTERISTICS

Demographic characteristic	**Type**	**count**	**%**
Gender	male	32	44.4
	Female	40	55.6
	Total	**72**	**100.0**
Position	Teacher	58	80.6
	Educational supervisor	14	19.4
	Total	**72**	**100.0**
Academic qualification	Intermediate diploma	1	1.4
	Bachelor's	24	33.3
	Higher diploma	22	30.6
	Master's	18	25.0
	Ph.D.	7	9.7
	Total	**72**	**100.0**
Experience	1- 5	20	27.8
	More than 5 less than 10	19	26.4

	10 and more	33	45.8
	Total	**72**	**100.0**

RESULTS

To facilitate interpretation, the rating scale for evaluating item quality was divided into five classes. The scale ranged from 1 to 5, with a total span of 4 points. Accordingly, each class covers an interval of 0.8, as shown in Table 2.

TABLE 2. THE CLASSES OF THE ARITHMETIC MEAN FOR THE JUDGES' RESPONSES TO THE AI-TEST' ITEMS

scale	**Mean class**
1-less than 1.8	Very Low
1.8- less than 2.6	Low
2.6- less than 3.4	Medium
3.4- less than 4.2	High
4.2- 5	Very High

The Judges' responses file was obtained in Excel file format from Google Drive and then converted to an SPSS data file. The responses were analyzed according to the test quality criteria. Below is a detailed presentation of the results of applying the standards.

CLARITY OF TEST ITEMS

Regarding the first study question, arithmetic means and standard deviations were calculated for the judges' responses to the item clarity criterion. The means were classified according to the previously mentioned scaling classes, and the results were recorded in Table 3.

TABLE 3. ARITHMETIC MEANS, AND STANDARD DEVIATIONS FOR THE JUDGES' RESPONSES TO THE ITEM CLARITY CRITERION

Item No.	**Mean**	**Mean class**	**Std. Deviation**
1	4.097	Clear	1.3548
2	4.056	Clear	1.2547
3	4.375	Very Clear	1.1188
4	3.903	Clear	1.4838
5	3.833	Clear	1.4241
6	4.389	Very Clear	1.1452
7	4.056	Clear	1.1855
8	4.347	Very Clear	1.2005
9	3.833	Clear	1.2782
10	1.931	Not Clear	1.1047
11	3.667	Clear	1.2333
12	2.958	Medium clarity	1.2154
13	3.708	Clear	1.1920
14	4.111	Clear	1.0949
15	4.264	Very Clear	1.1382
16	1.861	Not Clear	1.2705
17	3.486	Clear	1.3531
18	3.306	Medium clarity	1.1462
19	4.458	Very Clear	1.0200
20	4.347	Very Clear	1.0768

By observing the arithmetic means in Table 3 of the judges' responses to the clarity of the test items, 16 items fell within the "clear" to "very clear" classification. Items 12 and 18 were categorized as having medium clarity, while items 10 and 16 were classified as not clear.

To investigate the significance of the differences between the judges' responses on the clarity of the test items according to the judges' roles variable (mathematics teacher, mathematics educational supervisor), T-test was used for two independent samples, and the results are recorded in Table 4.
Table 4. Results for the difference between the judges' responses on the clarity of the test items according to the judges' roles variable (mathematics teacher, mathematics educational supervisor)

TABLE 4. RESULTS FOR THE DIFFERENCE BETWEEN THE JUDGES' RESPONSES ON THE CLARITY OF THE TEST ITEMS ACCORDING TO THE JUDGES' ROLES VARIABLE (MATHEMATICS TEACHER, MATHEMATICS EDUCATIONAL SUPERVISOR)

Item No.	t	df	Sig.
1	.007	70	.994
2	-.641-	70	.524
3	.101	70	.920
4	-1.797-	70	.077
5	-.662-	70	.510
6	-.863-	70	.391
7	.714	70	.478
8	.395	70	.694
9	1.113	70	.269
10	.121	70	.904
11	-.313-	70	.755
12	.007	70	.994
13	-.641-	70	.524
14	.101	70	.920
15	-1.797-	70	.077
16	-.662-	70	.510
17	-.863-	70	.391
18	.714	70	.478
19	.395	70	.694
20	1.113	70	.269

All statistical significance values exceeded 0.05, indicating that the judges reached a consensus, despite their differing roles.
To investigate the potential existence of statistically significant differences at the 0.05 significance level among the judges based on the experience variable regarding the clarity of the test items, a One-Way ANOVA (Analysis of Variance) was employed. This method was chosen due to the judges being categorized into three distinct experience groups. The results of the One-Way ANOVA concerning the clarity of the test items among the judges, based on years of experience, are presented in Table 5.

TABLE 5. ONE-WAY ANOVA RESULTS ON THE DEGREE OF CLARITY OF THE TEST ITEMS AMONG THE JUDGES ACCORDING TO THE VARIABLE OF EXPERIENCE IN YEARS

Item No.	df	Mean Square	F	Sig.
1	2	1.075	0.578	0.563
2	2	1.973	1.263	0.289
3	2	1.443	1.158	0.32
4	2	2.586	1.18	0.313
5	2	0.467	0.225	0.799
6	2	1.629	1.251	0.293
7	2	2.399	1.743	0.183
8	2	1.722	1.202	0.307
9	2	0.705	0.424	0.656
10	2	3.319	2.862	0.064
11	2	1.586	1.044	0.358
12	2	0.668	0.445	0.643
13	2	3.716	2.744	0.071
14	2	0.984	0.817	0.446
15	2	1.647	1.281	0.284

16	2	4.237	2.755	0.071
17	2	1.752	0.956	0.39
18	2	0.638	0.478	0.622
19	2	1.164	1.122	0.331
20	2	1.126	0.97	0.384

No statistically significant differences appeared between the judges' responses on the degree of clarity of the test items based on the experience variable. All statistical significance values in Table 5 are greater than the 0.05 level of significance.

TABLE 6. ARITHMETIC MEANS, AND STANDARD DEVIATIONS FOR THE JUDGES' RESPONSES TO THE ITEM RELEVANCE CRITERION

Item No.	**Mean**	**Mean class**	**Std. Deviation**
1	4.236	Very High	1.0941
2	4.167	High	1.1749
3	4.417	Very High	.9750
4	4.333	Very High	1.0615
5	3.986	High	1.2160
6	4.431	Very High	1.0592
7	4.319	Very High	.9905
8	4.444	Very High	.9914
9	3.889	High	1.1934
10	2.778	Medium	1.3761
11	3.694	High	1.0699
12	2.931	Medium	1.1788
13	3.833	High	1.0747
14	2.292	weak	1.5239
15	4.389	Very High	1.0285
16	2.931	Medium	1.4175
17	3.694	High	1.1214
18	3.569	High	1.1237
19	4.403	Very High	1.0570
20	4.375	Very High	1.0934

Out of the (20) items in the test, (16) items were deemed to have high to very high relevance to the topic of triangle congruence, according to the judges' assessments, representing 80% of the total test items. Furthermore, the standard deviation value for the judges' responses to this item indicates a notable level of disagreement among them regarding its relevance to triangle congruence.

Item No. 12 was rated as moderately clear and moderately relevance based on the judges' responses. The text of the item suggests it pertains to triangles in general rather than specifically addressing triangle congruence. There appears to be significant disagreement among the judges regarding the relevance of the item to the topic. This is further supported by the standard deviation and coefficient of variation, which reached 41% among the judges.

The differences between the judges' responses on the relevance of the test items to triangle's congruence according to the judges' roles variable (mathematics teacher, mathematics educational supervisor), were investigated and the results are recorded in Table 7.

TABLE 7. RESULTS FOR THE DIFFERENCE BETWEEN THE JUDGES' RESPONSES ON THE RELEVANCE OF THE TEST ITEMS TO TRIANGLE'S CONGRUENCE ACCORDING TO THE JUDGE'S ROLES VARIABLE (MATHEMATICS TEACHER, MATHEMATICS EDUCATIONAL SUPERVISOR)

Item No.	**t**	**df**	**Sig.**
1	-1.566-	70	0.122
2	-1.186-	70	0.24
3	-.354-	70	0.724
4	-.372-	70	0.711
5	-2.320-	70	0.023*

6	-.272-	70	0.787
7	-.457-	70	0.649
8	-.232-	70	0.817
9	-.386-	70	0.701
10	-.454-	70	0.651
11	-.911-	70	0.365
12	1.276	70	0.206
13	-.644-	70	0.522
14	0.6	70	0.551
15	-.448-	70	0.656
16	0.214	70	0.831
17	0.988	70	0.326
18	1.053	70	0.296
19	-.101-	70	0.92
20	-.203-	70	0.84

(*) significance at the level of significance 0.05

The only difference noted between teachers and supervisors was in their assessment of the relevance of item No. 5 to the topic of triangle congruence.

To analyze the significant differences among the judges based on their experience regarding the relevance of the test items with triangle congruence, a One-Way ANOVA (Analysis of Variance) was conducted. The results of this analysis are presented in Table 8.

TABLE 8. ONE-WAY ANOVA RESULTS ON THE DEGREE OF RELEVANCE OF THE TEST ITEMS WITH TRIANGLE CONGRUENCE AMONG THE JUDGES ACCORDING TO THE VARIABLE OF EXPERIENCE IN YEARS

Item No.	**df**	**Mean Square**	**F**	**Sig.**
1	2	3.986	3.572	0.033*
2	2	2.054	1.509	0.228
3	2	0.985	1.038	0.36
4	2	2.654	2.452	0.094
5	2	4.926	3.572	0.033*
6	2	2.317	2.131	0.126
7	2	2.289	2.427	0.096
8	2	1.44	1.485	0.234
9	2	0.377	0.259	0.773
10	2	0.137	0.07	0.932
11	2	0.759	0.657	0.522
12	2	2.2	1.611	0.207
13	2	1.561	1.365	0.262
14	2	2.147	0.922	0.402
15	2	1.995	1.936	0.152
16	2	0.068	0.033	0.968
17	2	0.056	0.043	0.958
18	2	0.268	0.208	0.813
19	2	1.978	1.811	0.171
20	2	1.388	1.166	0.318

(*) significance at the level of significance 0.05

It also appears that item No. 5 showed a difference between the judges depending on the experience variable, and through making post-hoc comparisons between levels of experience, it was found that the less experienced (less than 5 years) judges see this item as more closely relevance to the subject of triangle congruence than their counterparts of judges with high experience.

SUITABILITY OF THE ITEM FOR EIGHTH GRADE STUDENTS

Regarding the third study question, arithmetic means and standard deviations were calculated for the judges' responses to the item's suitability criterion. The results were recorded in table 9

TABLE 9. ARITHMETIC MEANS, AND STANDARD DEVIATIONS FOR THE JUDGES' RESPONSES TO THE ITEM'S SUITABILITY CRITERION

Item No.	Mean	Mean class	Std. Deviation
1	4.181	suitable	1.0658
2	4.097	suitable	1.1403
3	4.403	Very suitable	1.0162
4	4.125	suitable	1.1741
5	3.819	suitable	1.2596
6	4.389	Very suitable	1.0285
7	4.264	Very suitable	.9640
8	4.306	Very suitable	1.0297
9	3.792	suitable	1.2096
10	2.097	Not suitable	1.1647
11	3.750	suitable	1.0845
12	3.056	Medium	1.1615
13	3.792	suitable	1.0473
14	2.431	Not suitable	1.3513
15	4.403	Very suitable	.9737
16	2.083	Not suitable	1.2644
17	3.528	suitable	1.1745
18	3.597	suitable	1.0570
19	4.500	Very suitable	1.0209
20	4.306	Very suitable	1.1462

It is noted from the values of the standard deviations of the judges' responses to the test items, specifically those that show problems in the assessment's criteria, that there is a difference between their opinions.
The differences between the judges' responses on the suitability criterion of the test items to eighth-grade students according to the judges' roles variable (mathematics teacher, mathematics educational supervisor), were investigated and the results are recorded in Table 10.

TABLE 10. RESULTS FOR THE DIFFERENCE BETWEEN THE JUDGES' RESPONSES ON THE SUITABILITY OF THE TEST ITEMS TO EIGHTH-GRADE STUDENTS ACCORDING TO THE JUDGES' ROLES VARIABLE (MATHEMATICS TEACHER, MATHEMATICS EDUCATIONAL SUPERVISOR)

Item No.	**t**	**df**	**Sig.**
1	-.970-	70	0.336
2	-1.485-	70	0.142
3	-.985-	70	0.328
4	-.822-	70	0.414
5	-1.808-	70	0.075
6	-.448-	70	0.656
7	-.710-	70	0.48
8	-1.078-	70	0.285
9	-.964-	70	0.338
10	0.858	70	0.394
11	-1.240-	70	0.219
12	0.71	70	0.48
13	-1.408-	70	0.164
14	0.006	70	0.995
15	-.414-	70	0.68
16	0.508	70	0.613
17	-.154-	70	0.878
18	0.381	70	0.704
19	0	70	1

20	-.186-	70	0.853

To further examine the impact of the judges' experience levels on the suitability criteria for the items, One-Way ANOVA analysis was conducted, with the results displayed in Table 11.

TABLE 11. ONE-WAY ANOVA RESULTS REGARDING THE SUITABILITY OF THE TEST ITEMS FOR EIGHTH-GRADE STUDENTS ACCORDING TO THE JUDGES' YEARS OF EXPERIENCE

Item No.	df	Mean Square	F	Sig.
1	2	4.671	4.52	0.014
2	2	2.211	1.736	0.184
3	2	2.312	2.323	0.106
4	2	2.597	1.933	0.152
5	2	4.512	3.004	0.056
6	2	1.131	1.071	0.348
7	2	2.277	2.558	0.085
8	2	2.03	1.967	0.148
9	2	1.347	0.918	0.404
10	2	4.837	3.852	0.026
11	2	0.913	0.772	0.466
12	2	1.28	0.947	0.393
13	2	1.868	1.738	0.183
14	2	2.593	1.437	0.245
15	2	2.312	2.545	0.086
16	2	1.55	0.969	0.385
17	2	0.073	0.052	0.95
18	2	0.847	0.753	0.475
19	2	2.341	2.33	0.105
20	2	0.335	0.25	0.78

(*) significance at the level of significance 0.05

LEVELS OF TEST ITEMS DIFFICULTY

In this section, we present the findings related to the level of difficulty of the test items generated by AI, as assessed by the judges. The judges assessed the difficulty of each item, and their responses were quantified through arithmetic means and standard deviations. The results were recorded in table 12.

TABLE 12. ARITHMETIC MEANS, AND STANDARD DEVIATIONS FOR THE JUDGES' RESPONSES TO THE ITEM'S DIFFICULTY CRITERION

Item No.	Mean	Mean class	Std. Deviation
1	1.833	Easy	1.1006
2	2.167	Easy	1.0481
3	2.722	Medium	1.0776
4	3.222	Medium	1.1775
5	3.069	Medium	0.9976
6	2.931	Medium	1.1668
7	3.194	Medium	1.0297
8	3.597	Difficult	1.3496
9	3.542	Difficult	1.0607
10	3.681	Difficult	1.0185
11	3.181	Medium	0.9689
12	2.806	Medium	0.929
13	2.764	Medium	0.9999
14	2.542	Medium	0.9632
15	3.514	Difficult	1.3634
16	3.236	Medium	1.0413
17	3.278	Medium	0.9961
18	2.694	Medium	0.8498
19	3.375	Medium	1.5694

20	3.528	Difficult	1.2779

The differences between the judges' responses on the difficulty criterion of the test items according to the judges' roles variable (mathematics teacher, mathematics educational supervisor), were investigated and the results are recorded in Table 13.

TABLE 13. RESULTS FOR THE DIFFERENCE BETWEEN THE JUDGES' RESPONSES ON THE DIFFICULTY OF THE TEST ITEMS ACCORDING TO THE JUDGES' VARIABLE (MATHEMATICS TEACHER, MATHEMATICS EDUCATIONAL SUPERVISOR)

Item No.	t	df	Sig.
1	1.268	70	0.209
2	-.471-	70	0.639
3	-1.076-	70	0.286
4	-1.502-	70	0.138
5	-.305-	70	0.761
6	-1.539-	70	0.128
7	-1.846-	70	0.069
8	-1.249-	70	0.216
9	-1.831-	70	0.071
10	-1.015-	70	0.313
11	-.450-	70	0.654
12	0.088	70	0.93
13	1.408	70	0.164
14	-.435-	70	0.665
15	-1.728-	70	0.088
16	0.945	70	0.348
17	-.929-	70	0.356
18	-.097-	70	0.923
19	-1.483-	70	0.143
20	-1.314-	70	0.193

By examining the statistical significance values in Table 13, we notice that there are no statistically significant differences between teachers and supervisors in scaling of the difficulty of the test items. All statistical significance values exceeded 0.05, and this means that the judges agreed despite their different roles.

To analyze the significant differences among judges based on their experience with the difficulty of the test items, a One-Way ANOVA was conducted. The results are presented in Table 14.

TABLE 14. ONE-WAY ANOVA RESULTS ON THE DIFFICULTY OF THE TEST ITEMS AMONG THE JUDGES ACCORDING TO THE VARIABLE OF EXPERIENCE IN YEARS.

Item No.	df	Mean Square	F	Sig.
1	2	3.915	3.456	0.037
2	2	0.566	0.508	0.604
3	2	3.81	3.513	0.035
4	2	2.472	1.824	0.169
5	2	2.672	2.823	0.066
6	2	3.676	2.841	0.065
7	2	0.837	0.785	0.46
8	2	6.154	3.629	0.032
9	2	3.215	3.02	0.055
10	2	1.588	1.555	0.219
11	2	1.033	1.104	0.337
12	2	0.582	0.668	0.516
13	2	0.25	0.245	0.784
14	2	0.9	0.969	0.384
15	2	4.902	2.768	0.07
16	2	2.209	2.1	0.13
17	2	3.223	3.475	0.036

18	2	1.113	1.565	0.216
19	2	6.178	2.623	0.08
20	2	5.481	3.603	0.032

(*) significance at the level of significance 0.05

Post-hoc comparisons indicated statistically significant differences between judges with medium experience (5 to less than 10 years) and those with high experience (10 years or more).

A SUMMARY OF THE DESCRIPTION OF THE TEST ITEMS ACCORDING TO THE FOUR CRITERIA

This section contains a summary and description of all test items according to the four criteria, based on the judges' assessments. Table 15. summarizes this.

TABLE 15. A SUMMARY AND DESCRIPTION OF ALL TEST ITEMS ACCORDING TO THE FOUR CRITERIA.

Item No.	Clarity	Relevance	Suitability	Difficulty
1	Clear	Very High	suitable	Easy
2	Clear	High	suitable	Easy
3	Very Clear	Very High	Very suitable	Medium
4	Clear	Very High	suitable	Medium
5	Clear	High	suitable	Medium
6	Very Clear	Very High	Very suitable	Medium
7	Clear	Very High	Very suitable	Medium
8	Very Clear	Very High	Very suitable	Difficult
9	Clear	High	suitable	Difficult
10	Not Clear	Medium	Not suitable	Difficult
11	Clear	High	suitable	Medium
12	Medium clarity	Medium	Medium	Medium
13	Clear	High	suitable	Medium
14	Clear	weak	Not suitable	Medium
15	Very Clear	Very High	Very suitable	Difficult
16	Not Clear	Medium	Not suitable	Medium
17	Clear	High	suitable	Medium
18	Medium clarity	High	suitable	Medium
19	Very Clear	Very High	Very suitable	Medium
20	Very Clear	Very High	Very suitable	Difficult

When reviewing the summary of the results (table 15) of judging the AI- test' items according to the four criteria and based on the judges' responses, we conclude the following:

1. (7) items out of (20), at a rate of 35%, had a high degree of clarity of wording and relevance to the topic of triangle congruence, and were very suitable for the level of eighth grade students, with different levels of difficulty.
2. (16) items out of (20), at a rate of (80%) with different levels of difficulty can be considered appropriate for use in the mathematics achievement test about congruence of triangles for the eighth grade, due to their approval by the judges according to the criteria.
3. Only one item, and despite its clear wording, it was not accepted by the judges due to its weak relevance to the topic of congruent triangles and its unsuitability for eighth-grade students.
4. Two items, although moderately relevance to the topic of triangle congruence, did not receive the approval of the judges due to the lack of clarity in their linguistic wording.
5. One item had a medium score in all criteria.

DISCUSSION

The findings paint a more nuanced picture of the pedagogical potential of such items, weighing both the strengths and limitations of their uses within a localized curricular framework. This research thereby contributes to the growing debate surrounding the integration of artificial intelligence into educational assessment by investigating the four evaluative dimensions.

Eighty percent of the items were judged to be clear-to-very clear by the panel of judges, thus suggesting that large portions of the AI-generated content follow conventional linguistic expectations in math education. Hence, it seems very plausible that generative models, such as ChatGPT, can be used to generate syntactically correct and contextually accurate content in most cases. However, since a small number were found to be unclear, AI systems occasionally might use terminology unfamiliar to local learners or teachers. This outcome resonates with the warnings from Abedi (2006) and Sambasivan et al. (2021), who raised the risk of linguistic and cultural misalignment in automated assessment. Judges' unease over terms like "correspondence of triangles" represents an instance of conceptual mismatches and non-standard terminology in AI-generated math questions reported in Lai et al. (2023).

With relevance, the large majority of the items were perceived as appropriate to the topic of triangle congruence, while others, such as item 14, stood out on the negative side of relevance evaluations. These evaluations were mostly due to the concern that the item focused on general geometric properties instead of congruence-specific reasoning. Zhai (2022) also noticed that AI-generated content may face some difficulties in somewhat accurately capturing the intended constructs without human guidance. Different levels of agreement on item relevance reinforce the call by Chatterji and Kar (2023) for hybrid methods that merge automated generation with expert validation, especially for subject-matter item construction.

The statistical analyses do not find much difference in judges' ratings of clarity and relevance according to judges' professional roles or levels of experience, implying rather strong inter-rater reliability. However, differences due to experience are observed with items 1 and 5. Experienced educators may apply more stringent criteria in interpreting construct alignment and pedagogical usefulness. This finding resonates with Mislevy et al. (2003) in implying an interpretation of tasks within validity constructs may be influenced by professional training.

Another important criterion for evaluation that emerged was item appropriateness for eighth-grade students. Most items were considered appropriate, whereas items 10, 14, and 16 were inappropriate due to the items being unclearly worded, low in relevance, or offensive to curricular expectations. This echoes Holmes et al. (2021) and Pellegrino et al. (2001) emphasizing the need for developmental and curricular alignment in item construction. Differences in suitability ratings also point to the need for cultural and linguistic contextualization, especially in an area like Jordan with somewhat different educational standards.

Regarding item difficulty, it appeared that most test items presented a moderate difficulty level, with some rated as easy and some as harder, thus allowing for a balanced evaluation that may discriminate student proficiency levels. This outcome also suggests that AI systems have the power to some extent to replicate test difficulty hierarchies. Yet some items recorded statistically significant differential item display in perceived difficulty across the levels of judges' experience. For instance, judges of experience reckoned item 1 as more difficult than those less experienced. Similar patterns emerged in the rating for item 8. Thus, such discrepancies validate Luckin et al.'s (2016) proposition that the presence of human moderation is indispensable to ensuring that the assessments are properly standardized across various learner populations.

In general, this study affirms that AI tools can provide a base pool of assessment items that fulfill many of the pedagogical and psychometric requirements, though, ironically, their use still calls for heavy human intervention. Without the review of educators, one runs the risk of items that do not fit local curriculum standards, use unfamiliar terminologies, or even exhibit construct underrepresentation issues. Therefore, a hybrid approach is strongly recommended, where AI-generated drafts undergo expert refinement and contextual adaptation. This approach is aligned with best practices in Responsible AI use as outlined by Cowie et al. (2023) and Kasneci et al. (2023), who highlight the importance of transparency, accountability, and fairness in AI-enhanced educational systems.

CONCLUSION

This study examined the pedagogical and psychometric quality of AI-generated mathematics assessment items on the topic of triangle congruence, using expert evaluations from educators in Jordan. The majority of items were rated highly for clarity, relevance, difficulty, and grade-level appropriateness, indicating that generative AI tools particularly ChatGPT hold strong potential in producing curriculum-aligned assessment content. These findings affirm the value of AI in streamlining test development, particularly in time-sensitive or resource-limited educational settings.

However, the study also reveals that AI-generated content, while syntactically coherent, may suffer from linguistic ambiguity and cultural misalignment. Items containing uncommon terminology or conceptually vague language weakened both content relevance and construct validity. Psychometric analysis showed consistent evaluations across most clarity dimensions, but statistically significant differences in difficulty and

relevance judgments especially between teacher and supervisor roles highlight the subjective influence of evaluator experience and professional expectations. No items were found to be quite easy or somewhat difficult, further accounting for a balanced distribution allowing for individual differences: an indication that this could be an excellent assessment.

The ecological validity and developmental appropriateness of the findings are restricted by using university students for the school population of eighth-grade learners: one important limitation. Nonetheless, this study offers a rich addition to the growing literature modeling AI-human collaboration in educational design. It points to the need for an expert review and localized validation, especially in culturally defined educational systems at the linguistic, cognitive, and curricular levels.

In effect, generative AI is a powerful supporting instrument for assessment design, the outputs of which must be moderated by humans and calibrated to their local context. Hence the future of AI in education will not lie in replacement but in the responsible integration based upon psychometric integrity, pedagogical relevance, and cultural sensitivity.

RECOMMENDATIONS

Several key points can be derived from the findings to suggest recommendations for a responsible integration of generative AI in educational assessment: first, prompt engineering should be localized to meet the linguistic, curricular, and cultural context of the target learners. This entails training AI models on region-specific educational data and using terminology according to the given context.

Second, a more rigorous validation process should be adopted, with expert review, psychometric evaluation, and piloting with actual students to reinforce the reliability and relevance of AI-generated items before their classroom use.

Third, teachers should be trained in AI literacy so they can critically analyze AI-generated content, identify bias, and adjust items to guarantee pedagogical clarity and alignment with learning objectives.

Fourth, future research should involve field-testing of products with the target population in order to improve developmental appropriateness; additionally, comparative studies of AI-only versus co-designed items will offer additional perspectives to best inform the practice.

Finally, collaboration among educators, AI developers, and psychometric experts is essential to produce assessment tools that are both innovative and contextually sound. With such safeguards, AI can enhance not replace human judgment in educational measurement.

LIMITATIONS

Despite its contributions, this study is subject to several limitations that warrant consideration. First, the evaluation relied exclusively on the judgments of Jordanian educators, which may constrain the generalizability of the findings to broader educational systems with different curricular and cultural contexts. Second, while structured rubrics were employed, the use of Likert-scale ratings inherently introduces an element of subjectivity that may affect the consistency of the assessments. Third, the study focused solely on the topic of triangle congruence, limiting the applicability of the results to other areas of mathematics or educational domains. Fourth, the test items were generated using a single AI model (ChatGPT), and outcomes may vary with different generative systems or prompt configurations. Fifth, although the test was designed for eighth-grade students, performance data were collected from university students, which may affect the developmental validity of the findings. Lastly, the absence of a direct, item-level comparison with teacher-generated assessments precludes definitive conclusions regarding the comparative effectiveness of AI-authored test content.

REFERENCES

1. Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 377–398). Routledge.
2. Barrett, M., Branson, L., Carter, S., DeLeon, F., Ellis, J., Gundlach, C., & Lee, D. (2019). Using artificial intelligence to enhance educational opportunities and student services in higher education. Inquiry: The Journal of the Virginia Community Colleges, 22(1), 11.
3. Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. Longmans, Green.
4. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
5. Chatterji, M., & Kar, N. (2023). Human in the loop systems for AI generated educational content. AI in Education Review, 9(2), 112–130.

6. Cowie, B., Jones, A., Harlow, A., McGee, C., & Cooper, B. (2023). Ethical and practical considerations of artificial intelligence in assessment: A New Zealand perspective. Assessment in Education: Principles, Policy & Practice, 30(2), 182–198. https://doi.org/10.1080/0969594X.2023.2179301
7. Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). Sage Publications.
8. Holmes, W., Bialik, M., & Fadel, C. (2021). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.
9. Jeon, Y., & Kim, T. (2018). The development and application of a responsive web-based smart learning system for the cyber project learning of elementary informatics gifted students. Journal of Theoretical and Applied Information Technology, 96(5), 1397.
10. Kasneci, E., Sessler, K., Kübler, A., & Sailer, M. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
11. Lai, M., Zhu, X., & Zhang, Q. (2023). Evaluating the pedagogical soundness of LLM generated math problems. International Journal of Educational Technology in Higher Education, 20(1), 22–35.
12. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson.
13. Malik, G., Tayal, D., & Vij, S. (2019). An analysis of the role of artificial intelligence in education and teaching. In Recent findings in intelligent computing techniques (pp. 407–417). Springer.
14. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.
15. Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02
16. National Council of Teachers of Mathematics. (2014). Principles to actions: Ensuring mathematical success for all. NCTM.
17. Neto, A. J. M., & Fernandes, M. A. (2019). Chatbot and conversational analysis to promote collaborative learning in distance education. In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT) (pp. 324–326). IEEE.
18. Pereira, J., Fernández Raga, M., Osuna Acedo, S., Roura Redondo, M., Almazán López, O., & Buldón Olalla, A. (2019). Promoting learners' voice productions using chatbots as a tool for improving the learning process in a MOOC. Technology, Knowledge and Learning, 24(6), 807–820.
19. Rane, N. (2024). Enhancing the quality of teaching and learning through Gemini, ChatGPT, and similar generative artificial intelligence: Challenges, future prospects, and ethical considerations in education. TESOL and Technology Studies, 5(1), 1–6. https://doi.org/10.48185/tts.v5i1.1000
20. Sambasivan, N., Kapania, S., Grevet, C., Paritosh, P., & Aoki, P. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–15). https://doi.org/10.1145/3411764.3445518
21. Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.
22. Shukhman, A. E., Bolodurina, I. P., Polezhaev, P. N., Ushakov, Y. A., & Legashev, L. V. (2018). Adaptive technology to support talented secondary school students with the educational IT infrastructure. In Global Engineering Education Conference (EDUCON). IEEE.
23. Shute, V. J., & Ventura, M. (2013). Stealth assessment: Measuring and supporting learning in video games. MIT Press. https://mitpress.mit.edu/9780262518813/stealth_assessment/
24. Vázquez Cano, E., Mengual Andrés, S., & López Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. International Journal of Educational Technology in Higher Education, 18(1), 1–20.
25. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 16, 39. https://doi.org/10.1186/s41239-019-0171-0
26. Zhai, X. (2022). Can AI generate good educational assessments? Evaluating GPT-based item generation in science. Educational Technology Research and Development, 70(6), 3235–3252. https://doi.org/10.1007/s11423-022-10132-3