

# A Comprehensive Evaluation of GenAI-based Items in a Medical Examination

**Yanlin Jiang, Marcus Walker, Andrew Dallas,  
Aquia Richburg, Nikole Gregg, Brittany Corrigan**

National Commission on Certification of Physician Assistants

## Abstract

This study evaluated GenAI-based medical assessment items against subject matter expert developed items. Analyses of item content and response data from a high-stakes examination showed comparable content quality and psychometric characteristics across the two item types. These findings provide empirical support for the quality of GenAI-based items and their potential use in educational and professional assessment programs.

medical assessments in a high-stakes testing context. The study was guided by two research questions: (1) Are these items comparable to human-developed items in terms of content quality? and (2) Do they exhibit psychometric characteristics comparable to those of human-developed items? The evaluation focused on item content quality and key psychometric properties, including the quality of item stems and item response options, classical item statistics, item response theory (IRT)-based item parameter estimates, and item flagging rates.

## 1 Introduction

As generative artificial intelligence (GenAI) technologies continue to advance, GenAI-based items are expected to play a significant role in the item development process within educational and professional testing programs. In this study, GenAI-based items refer to items initially generated by large language models (LLMs) using a predefined prompting strategy and subsequently undergo human review prior to administration. Prior studies have examined AI-generated items. Concerns regarding their bias, clarity and overall quality have been addressed in literature and have challenged their broader use. Empirical evidence regarding GenAI-based items that incorporate human review and refinement under actual testing conditions remains limited. Such investigation is particularly important in high-stakes testing, where item quality directly affects score reliability and test outcome.

## 2 Aims

The present study seeks to evaluate GenAI-based multiple-choice questions (MCQs) developed for

## 3 Methods

The study integrates pre-administration content review with post-administration psychometric analyses to provide a comprehensive evaluation of item quality and the comparability of the GenAI-based and subject matter experts (SME)-developed items.

### 3.1 Item Development

The GenAI-based items examined in this study were developed in 2025 using GPT-4.1. A typical MCQ consists of an item stem and a set of response options. In medical assessments, item stems typically present clinical scenarios that incorporate patient symptoms, medical history, physical examination findings, and/or laboratory results. The prompting strategy was evaluated in prior studies (Walker et al., 2025A, 2025B; Makinde, 2025) and includes the following components: 1). question stem requirements following established medical examination pattern and difficulty; 2). clinical scenario templates containing patient demographics, presenting symptoms, examination findings, and laboratory results; 3). five-option

multiple-choice format with detailed rationales for each option for SME review; 4). relevant organization style rules; 5). chain-of-thought reasoning to ensure clinical coherence; 6). a three-item quiz format for each clinical topic to produce item drafts that assessed the same topic in different ways.

### 3.2 Item Content Evaluation

The evaluation included a systematic content review and revision process conducted by six experienced item writers (i.e., SMEs) and two Exam Content Developers. For each clinical topic, the SMEs panel jointly reviewed GenAI-based items until selecting one to use (typically in the first set of three items). Necessary revisions were made with the help of the Exam Development team members and tracked.

Afterwards, the number and types of changes were categorized to assess item content quality. The revisions were assigned to one of seven categories – laboratory value formatting, unit formatting, text changes for style and clarity, distractor revision/replacement, terminology modification, clinical focus, and patient details.

### 3.3 Psychometric Analysis

#### 3.3.1 Item response options

Response options were evaluated at both the item and test-section levels, with particular attention to distractor functioning. Distractors were classified as functioning or non-functioning using two criteria (Ebel & Frisbie, 1991; Lakens, 2017; Loudon & Macias-Muñoz, 2018; Osterlind, 1998). First is selection proportion ( $P_{ij}$ ). A distractor  $j$  in item  $i$  was classified as a non-functioning distractor (NFD) if it was selected by fewer than 5% of examinees, i.e.,  $P_{ij} = \frac{n_{ij}}{N} < 0.05$ , where  $N$  denotes the number of examinees responding to item  $i$  and  $n_{ij}$  denotes the number selecting distractor  $j$  on item  $i$ . Second is discrimination index  $D_j$ . Distractor discrimination was evaluated by comparing selection rates between upper- and lower-ability groups. A distractor was classified as non-functioning when its selection rate was higher among the upper-ability group than among the lower-ability group. That is,  $D_j = P_{Uj} - P_{Lj} > 0$ , where  $P_{Uj} = \frac{n_{Uj}}{N_U}$  and  $P_{Lj} = \frac{n_{Lj}}{N_L}$ .

#### 3.3.2 Statistical flagging rate:

Statistical flagging was based on classical item statistics, including item percent correct, point-biserial correlation, or response time. Although specific criteria may vary across testing programs, common flagging rules include low or negative point-biserial correlation, extreme item difficulty or response time, and combinations of these indicators. In the present study, the flagging criteria used in the operational review process for a certification examination were applied consistently to both SME-developed and AI-assisted items. The resulting flagging rates were then compared to assessing the overall psychometric comparability of the two item types.

#### 3.3.3 Item Statistics and Item Parameters

Item statistics and model-based parameters are central to evaluating item functioning and informing test scoring. Accordingly, examining these characteristics across item types provides important evidence regarding the feasibility of using GenAI-based items. Specifically, the distribution of response time, percent correct (i.e., classical item difficulty), point-biserial correlation coefficients, and IRT item difficulty were compared between the two item types.

## 4 Results

### 4.1 Data

The study sample included 3,514 examinees who completed a high-stake medical certification examination during the first half of 2026. The examination was administered as a timed, computer-based test consisting of 300 MCQs designed to assess medical and surgical knowledge and skills. Both SME-developed and GenAI-based pretest items were embedded within the examination. Four test forms were administered, with each form associated with three 50-item pretest sections, yielding 12 unique pretest sections in total. Examinees were randomly assigned to one of these pretest sections, and each pretest item was administered to approximately 282 to 305 examinees).

### 4.2 Item Content

Sixty-two GenAI-based items were reviewed and revised prior to administration. The 62 items contained 50 items for a complete pretest section and the remaining 12 were retained as backup

items in case content issues were identified during Forms Review (although none were subsequently chosen for replacement by two pairs of SMEs who were blinded to the AI assisted nature of the items). Content revisions were classified into seven categories described previously. Across categories (Appendix Table 1), the number of changes per item ranged from 0 to 18, with the average number of changes per item ranging from 0.27 (less than one change) to 9.48 (more than 9 changes). Revisions were most frequent for stem style or clarity, distractor content, and patient details (Categories 3, 4, and 7), with approximately 85% to 98% of items receiving revisions in these areas. In contrast, relatively few revisions were made to clinical focus and lab formatting (Categories 1 and 6).

To account for potential variation across content areas, GenAI-based and SME-developed items were compared within the same content blueprint and clinical topic. Fifteen matched pairs were identified based on their clinical topics. GenAI-based items required fewer revisions than SME-developed items across most categories, with the exception of distractor revision or replacement (Category 4). For both item types, revisions remained most frequent in Categories 3, 4, and 7 (Figure 1). Overall, the revision pattern for GenAI-based items after matching content was consistent with that observed in the unadjusted condition.

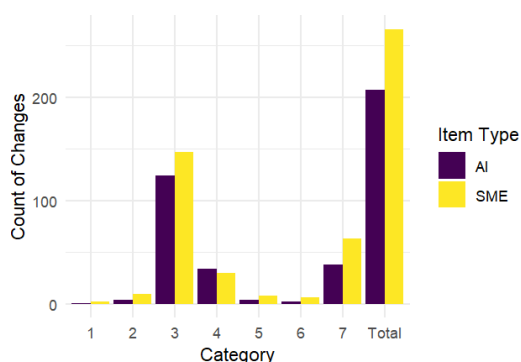


Figure 1: Count of changes across 7 categories for both item types

### 4.3 Psychometric analyses

Item quality was further evaluated using psychometric analyses, with SME-developed items serving as the reference group. Item analysis was conducted in R to obtain response time (RT), percent correct (p+), and point-biserial correlation

( $r_{pb}$ ). A unidimensional Rasch model was implemented in the computer software WINSTEPS to estimate item difficulty and person ability for both item types. A concurrent calibration was carried out to ensure model parameters at different test forms are on the same scale. Mean ability estimates ranged from 1.68 to 1.76, with standard deviation (SD) ranging from 0.44 to 0.55. Prior to comparing item statistics across the 12 pretest sections (4 forms x 3 pretest sections), mean ability was examined for the corresponding examinee groups. Equivalence across groups was assessed using ‘two one-sided t-tests’ (TOST) procedure with an equivalence margin of  $\pm 0.15$  logits. All pairwise comparisons satisfied the equivalence criterion ( $p < .05$ ), supporting the comparability of examinee ability across forms and pretest sections.

#### 4.3.1 Item Statistics

Pretest sections were constructed to achieve balanced content coverage across test forms. Response data was obtained from 12 unique sections. Section 1 contained 50 AI-assisted pretest items, whereas Sections 2-12 contained SME-developed items. Given the aligned ability distributions across examinee groups and balanced content coverage, consistent item performance was expected across item types. As shown in Table 1, mean response time (RT), percent correct (P+), and point-biserial correlation ( $r_{pb}$ ) are generally similar across sections. The largest differences were 7 seconds in RT, 0.08 in P, and 0.04 in  $r_{pb}$ .

| Section | N   | RT    | P+   | $r_{pb}$ | $b$  |
|---------|-----|-------|------|----------|------|
| 1       | 297 | 69.02 | 0.72 | 0.12     | 0.28 |
| 2       | 287 | 69.01 | 0.69 | 0.14     | 0.57 |
| 3       | 302 | 65.42 | 0.72 | 0.12     | 0.21 |
| 4       | 299 | 66.85 | 0.74 | 0.11     | 0.25 |
| 5       | 305 | 66.15 | 0.70 | 0.12     | 0.42 |
| 6       | 286 | 64.17 | 0.69 | 0.14     | 0.41 |
| 7       | 282 | 63.02 | 0.70 | 0.11     | 0.48 |
| 8       | 294 | 62.73 | 0.73 | 0.10     | 0.20 |
| 9       | 292 | 65.86 | 0.72 | 0.12     | 0.28 |
| 10      | 289 | 61.85 | 0.74 | 0.11     | 0.01 |
| 11      | 297 | 66.04 | 0.66 | 0.13     | 0.69 |
| 12      | 284 | 67.03 | 0.68 | 0.12     | 0.39 |

Note: N = Number of examinees in each section

Table 1: Average Item Statistics across Pretest Sections

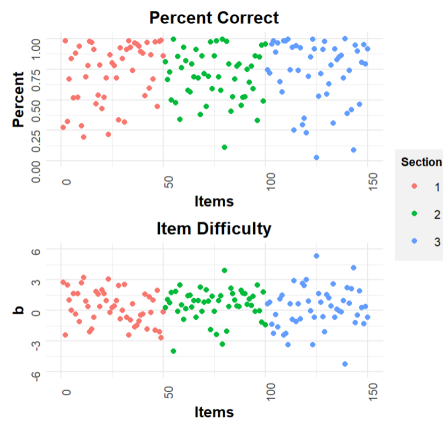


Figure 2: Percent correct and b-parameter estimates for items in Sections 1-3

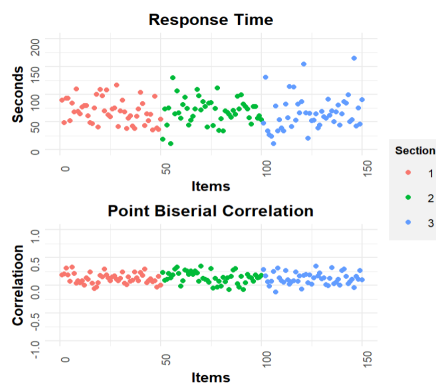


Figure 3: Response time and point-biserial correlation for items in Sections 1-3

Response time was slightly longer in Section 1 (AI-assisted) than in other sections, possibly reflecting its higher stem and response option word count. Median stem and response option lengths were 117 and 21 words, respectively, compared with approximately 105 (stem) and 18 (response option) words in most other sections (Appendix Table 2).

Figure 2 and 3 present the distributions of item statistics and Rasch b-parameter estimates for Sections 1–3, which were administered with the same operational test form. Section 1 contained AI-assisted items, whereas Sections 2 and 3 contained SME-developed items, allowing their distributional patterns to be examined under the same operation form. The distributions appeared visually similar across sections in terms of range, density, and spread. Section 2 showed a slightly higher mean b-parameter estimate, likely

influenced by a greater concentration of more difficult items.

Given the importance of Rasch item difficulty b-parameters in test scoring, additional analyses were conducted to assess their distributions for AI-assisted and SME-developed items. Wasserstein two-sample tests (Ramdas et al., 2017; Panaretos & Zemel, 2019) were performed at both the section and item-type levels. Section-level comparisons contrasted Section 1 (GenAI-based items) with each of Sections 2–12 (SME-developed items); all tests were nonsignificant ( $p=.1622$  to  $.9256$ ). At the item-type level, the Wasserstein distance was  $0.2608$  ( $p=.6838$ ). Collectively, these analyses detected no statistically significant differences in the distributions of b-parameter estimates between the two item types. Consistent with these results, the Q-Q plot (Figure 4) showed quantiles closely aligned with the identity line, providing visual support for similar distributional patterns.

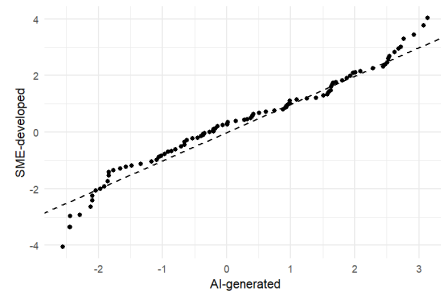


Figure 4: Q-Q plot of b-parameter estimates for GenAI-based and SME-developed items

### 4.3.2 Statistical Flag Rate

Four operational flagging criteria were applied to evaluate pretest item statistics. As shown in Figure 5, AI-assisted items had relatively high flag

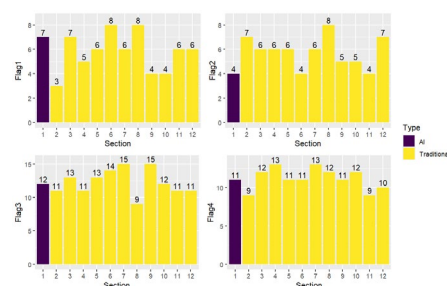


Figure 5: Number of flags across sections based on 4 rules

counts under criterion 1 (i.e., hard item) and low counts under criterion 2 (low point-biserial correlation), but no unusual pattern evident across the four criteria overall. We also examined the number of items flagged within each section. The GenAI section had 21 flagged items (42%), the highest rate observed, whereas the SME-developed sections ranged from 14 to 20 items (28%-40%). Although the GenAI section showed the highest number of flagged items, several SME-developed sections had similar counts (Appendix Table 3).

### 4.3.3 Item response options

Response data on item distractors from all 600 pretest items were analyzed. During content review, some item response options were removed, resulting in a small number of four-option items (Appendix Table 4). The final analysis therefore included 2,384 distractors rather than 2,400. To compute the discrimination index, examinees at the top 33% of the ability distribution were classified as the upper group, and those in the bottom 33% as the lower group. Using the selection-frequency criteria, Table 2 showed that the proportion of functioning distractors (FD) across SME-developed sections ranged from 34% to 46%, with most sections near 37%. Section 1, which contained GenAI-based items, had an FD rate of 36%, close to the median of the SME-developed sections. Under the discrimination-index criterion, FD rate ranged from 76% to 84%, substantially higher than those based on selection frequency, with the GenAI section at the lower end of this range. Pearson's Chi-Squared test was conducted to examine the association between item type (GenAI vs. SME) and distractor functioning. No statistically significant association was found under either criterion,  $\chi^2(1) = 0.19316$ ,  $p = .6603$  for selection frequency, or  $\chi^2(1) = 2.5542$ ,  $p = .110$  for discrimination index.

## 5 Conclusions

This study evaluated the quality of GenAI-based items using content-review data and examinee responses from a high-stakes medical certification examination. Under content-matched conditions, GenAI-based items required fewer revisions than SME-developed items, suggesting comparable

| Section | S.F. | D. I. | Both |
|---------|------|-------|------|
| 1       | 0.36 | 0.76  | 0.28 |
| 2       | 0.43 | 0.81  | 0.31 |
| 3       | 0.34 | 0.84  | 0.23 |
| 4       | 0.35 | 0.84  | 0.27 |
| 5       | 0.36 | 0.76  | 0.25 |
| 6       | 0.36 | 0.82  | 0.29 |
| 7       | 0.37 | 0.82  | 0.30 |
| 8       | 0.39 | 0.83  | 0.31 |
| 9       | 0.38 | 0.80  | 0.29 |
| 10      | 0.35 | 0.78  | 0.25 |
| 11      | 0.46 | 0.78  | 0.33 |
| 12      | 0.37 | 0.79  | 0.27 |
| Total   | 0.38 | 0.80  | 0.28 |

Note: S.F. = Selection frequency criterion  
D.I. = Discrimination index criterion  
Both = both criteria are considered

Table 2: Functioning Distractor Rate across Sections

content quality and potential gains in item-development efficiency.

Psychometric analyses showed that GenAI-based items performed comparably to SME-developed items across classical item statistics, Rasch b-parameter estimates, distractor functioning, and operational flagging indicators. Collectively, these findings provide empirical support for the feasibility of incorporating GenAI-based items into operational assessment development.

Both GenAI-based and SME-developed items underwent rigorous human review and revision before administration to meet the requirement for operational use; therefore, the GenAI-based items evaluated in this study reflect a combined AI-human development process rather than fully automated item generation.

Several limitations should be noted. First, the study included 50 GenAI-based items, and replication with larger and more diverse item samples is needed to strengthen generalizability. Second, the Wasserstein analyses assessed whether distributional differences could be detected but did not formally establish distributional equivalence. Future research should apply explicit similarity or equivalence frameworks. Third, item parameters were examined only under a Rasch model. Future studies should evaluate item performance under additional IRT models, including the two- and

three-parameter logistic models, to extend the relevance of the findings to testing programs that use these models.

## 6 References

Ebel, R. L. & Frisbie, D.A. (1991). *Essentials of educational measurement*. 5th edition. Englewood Cliffs, N.J.: Prentice Hall, Englewood Cliffs.

Lakens, D. (2017). *Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses*. Social Psychological and Personality Science, 8(4), 355–362.  
<https://doi.org/10.1177/1948550617697177>

Loudon, C. & Macias-Muñoz, A. (2018). *Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design*. Advances in Physiology Education 42 (4): 565–575. <https://doi.org/10.1152/advan.00186.2016>

Makinde, H., Walker, M., & Dallas, A. (2025, April 23-26). AI Generation of Medical specialty Items with LLM. Annual Meeting of the National Council on Measurement in Education Conference, Denver, CO, United States.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Kluwer Academic Publishers.

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6, 405–431.  
DOI: 10.1146/annurev-statistics-030718-104938.

Ramdas, A., García Trillos, N., & Cuturi, M. (2017). *On Wasserstein two-sample testing and related families of nonparametric tests*. Entropy, 19(2), 47.  
DOI: 10.3390/e19020047

Schuirman, D. J. (1987). *A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability*. Journal of Pharmacokinetics and Biopharmaceutics, 15(6), 657–680.  
<https://doi.org/10.1007/BF01068419>

Walker, M., Richburg, A., Makinde, H., Dallas, A., Jiang, Y., & Gregg, N. (2025, October 27-29). *SME comparison of PubMed RAG vs non-RAG items and fact-checking*. [Paper presentation]. Annual Meeting of the Artificial Intelligence in Measurement and Education Conference, Pittsburgh, PA, United States.

Walker, M., Makinde, H., Richburg, A., Dallas, A., & Goodman, J. (2025, October 27-29). *From theory to practice, high-stakes item development and review with Generative AI*. [Paper presentation]. Annual Meeting of the Artificial Intelligence in Measurement and Education Conference, Pittsburgh, PA, United States

## A Appendices

|                     | Cat. 1 | Cat. 2 | Cat. 3 | Cat. 4 |
|---------------------|--------|--------|--------|--------|
| Items with a change | 19.4%  | 38.7%  | 98.4%  | 87.1%  |
| N. changes per item | 0.27   | 0.56   | 9.48   | 2.35   |
| Minimum changes     | 0      | 0      | 0      | 0      |
| Maximum changes     | 3      | 4      | 18     | 5      |
|                     | Cat. 5 | Cat. 6 | Cat. 7 |        |
| Items with a change | 21.0%  | 12.9%  | 85.5%  |        |
| N. changes per item | 0.34   | 0.13   | 2.63   |        |
| Minimum changes     | 0      | 0      | 0      |        |
| Maximum changes     | 4      | 1      | 8      |        |

Note: Category 1: Lab formatting; Category 2: Unit formatting; Category 3: Text changes for style and/or clarity; Category 4: Distractor content (revising/replacing); Category 5: Medical terminology (revising or replacing); Category 6: Clinical focus (changing the lead-in to meet the assigned task); Category 7: Patient details:

Appendix Table 1: Summary of Item Revisions in GenAI-based Items

| Section | Stem | Option |
|---------|------|--------|
| 1       | 117  | 21     |
| 2       | 114  | 20     |
| 3       | 114  | 21     |
| 4       | 93.5 | 20     |
| 5       | 104  | 17.5   |
| 6       | 102  | 17     |
| 7       | 104  | 14     |
| 8       | 104  | 17.5   |
| 9       | 112  | 17     |
| 10      | 109  | 16.5   |
| 11      | 108  | 19.5   |
| 12      | 106  | 18     |

Appendix Table 2: Median of Word Counts per Item in Each Section

| Section | N  | N Flagged<br>Items | % Flagged<br>items |
|---------|----|--------------------|--------------------|
| 1       | 50 | 21                 | 0.42               |
| 2       | 50 | 14                 | 0.28               |
| 3       | 50 | 17                 | 0.34               |
| 4       | 50 | 18                 | 0.36               |
| 5       | 50 | 18                 | 0.36               |
| 6       | 50 | 20                 | 0.4                |
| 7       | 50 | 20                 | 0.4                |
| 8       | 50 | 19                 | 0.38               |
| 9       | 50 | 18                 | 0.36               |
| 10      | 50 | 16                 | 0.32               |
| 11      | 50 | 17                 | 0.34               |
| 12      | 50 | 14                 | 0.28               |

Appendix Table 4: Frequency and % of Flagged Items across Sections

| Section | N. items (4-<br>option) | N. items<br>(5-option) | N.<br>Distractors |
|---------|-------------------------|------------------------|-------------------|
| 1       | 1                       | 49                     | 199               |
| 2       | 3                       | 47                     | 197               |
| 3       | 0                       | 50                     | 200               |
| 4       | 1                       | 49                     | 199               |
| 5       | 2                       | 48                     | 198               |
| 6       | 1                       | 49                     | 199               |
| 7       | 2                       | 48                     | 198               |
| 8       | 2                       | 48                     | 198               |
| 9       | 1                       | 49                     | 199               |
| 10      | 0                       | 50                     | 200               |
| 11      | 2                       | 48                     | 198               |
| 12      | 1                       | 49                     | 199               |
| Total   | 16                      | 584                    | 2384              |

Appendix Table 3: Frequency of Items and Distractors across Sections