

Do Cognitive Classifications Predict Empirical Item Behaviour? Aligning Bloom's Taxonomy with the Psychometric Properties of National Centralized Mathematics Items across Grades 4–12, with an AI-Assisted Classification Benchmark

Rafat Altaweel^{1,*} & Moustafa Kamal Moussa^{2,3}

¹PPDT Department, City University Ajman, Ajman, UAE

²Education Department, City University Ajman, Ajman, UAE

³Faculty of Specific Education, Ain Shams University, Cairo, Egypt

*Correspondence: PPDT Department, City University Ajman, Ajman, UAE. E-mail: y.rafat@cu.ac.ae

Received: July 15, 2026

Accepted: September 18, 2026

Online Published: October 9, 2026

doi:10.5430/jct.v15n4p53

URL: <https://doi.org/10.5430/jct.v15n4p53>

Abstract

Centralized testing systems routinely build test blueprints on the assumption that an item's targeted cognitive level, coded with Bloom's revised taxonomy, is monotonically related to its empirical difficulty and discrimination: lower-order items should be easier, and higher-order items harder and more discriminating. This assumption is rarely tested empirically in K–12 national assessment, and almost never at scale in mathematics. Using item-analysis records for 265 four-option multiple-choice items drawn from centralized mathematics examinations (End-of-Term 1, 2025–2026) across Grades 4–12 and both the General and Advanced tracks, this study examined (a) the alignment between the expert human cognitive classification and the empirical psychometric behaviour of items, and (b) the agreement between the expert classification and an independent large-language-model (LLM) classification produced by Claude (Anthropic).

Difficulty (p-value) differed significantly across cognitive bands (Kruskal–Wallis $H = 11.26$, $p = .004$), with higher-order items being the hardest; the effect, however, was weak ($\epsilon^2 = .043$; Spearman $\rho = -.16$), and a Jonckheere–Terpstra test confirmed a significant decreasing trend across the ordered bands ($J = 7943$, $z = -2.56$, $p = .006$). Discrimination did not differ across cognitive bands ($H = 0.54$, $p = .77$), and difficulty and discrimination were essentially uncorrelated ($\rho = -.05$). Human–AI agreement was only fair (Cohen's $\kappa = .22$ at the six-level resolution and $\kappa = .17$ at the three-band level; raw agreement was 51.7% and 58.1%, respectively), with the model systematically over-assigning the “apply/analyse” band. No significant item-level difference emerged between the General and Advanced tracks. The findings offer partial, weak empirical support for the cognitive-difficulty assumption, no support for a cognitive–discrimination link, and a clear caution against replacing expert human coders with current LLMs in high-stakes item classification. Implications for blueprint validation, item-writer training, and AI-assisted assessment workflows are discussed.

Keywords: Bloom's taxonomy, item analysis, classical test theory, test blueprint, large language models, cognitive level, difficulty, discrimination, centralized assessment, mathematics education

1. Introduction

Large-scale centralized assessment systems anchor their construction on a test blueprint (a table of specifications) that allocates items to cognitive levels according to Bloom's taxonomy (Bloom, 1956) and its revision by Anderson and Krathwohl (2001). In the United Arab Emirates national mathematics programme, as in most large national systems, the blueprint fixes target proportions in advance—approximately 40% of items at the remember/understand level, 50% at apply/analyse, and 10% at higher-order (evaluate/create) skills. The blueprint therefore encodes a strong implicit measurement assumption: that the intended cognitive demand of an item is monotonically related to how the item

actually behaves once administered to students—that lower-order items will be statistically easier, and that higher-order items will be both harder and more discriminating between low- and high-ability students.

This assumption is intuitively appealing and operationally convenient, but it is not self-evidently true. Answering a multiple-choice item involves multiple cognitive and contextual factors, and item writers routinely misjudge how difficult students will find an item (Kibble & Johnson, 2011; Sydorenko, 2011). Classic and contemporary work has repeatedly questioned whether a taxonomic label attached to an item survives contact with real response data—memorably framed by Cunningham, Norman, Blake, Dauphinee, and Blackmore (1997) as the question of whether “a fact is an artefact.”

Empirical results are mixed. Some studies report that lower-order items are answered more easily than application-level items, broadly consistent with the blueprint assumption (e.g., a Turkish open-and-distance-learning analysis of 905 items; Koçdar, Karadağ, & Şahin, 2016), while others find that difficulty tracks cognitive level only weakly and that discrimination is largely independent of the intended cognitive level (Kim, Patel, Uchizono, & Beck, 2012; Zaidi et al., 2018). Kim et al. (2012), in particular, found that remember/understand/apply items were easier than analysis and evaluation items, yet the discrimination of application and evaluation items exceeded that of remember/understand items—suggesting that difficulty and discrimination need not move together with cognitive level. Consistent with this, Testa, Toscano, and Rosato (2018) found only a limited relationship between distractor efficiency and the Bloom level of statistics items, underscoring that taxonomic level is a coarse predictor of empirical item behaviour.

A second, newer, source of interest is generative artificial intelligence. Large language models (LLMs) now promise to automate parts of the item-development pipeline: classifying items by cognitive level and generating plausible distractors (Duong-Trung et al., 2024; Elkins & Kochmar, 2024; Scaria, Chenna, & Subramani, 2024). Recent work has begun to evaluate LLMs against Bloom’s taxonomy both as classifiers and as generators (Huber & Niklaus, 2025), and dedicated systems such as Bloom LLM have been fine-tuned to generate level-appropriate questions (Duong-Trung et al., 2024). Yet the existing literature is concentrated in medical and higher education, evaluates AI output chiefly through subjective expert judgement, and rarely benchmarks the AI against large-scale, real field-response data. The reliability of LLM-generated multiple-choice items and classifications remains contested (Kiyak & Emekli, 2024). Established guidance on principled item development (Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013) treats cognitive-level coding as one component of content validity rather than as a guarantee of empirical behaviour, a distinction the present study makes testable.

The present study exploits an uncommon research position: access to item-level psychometric records (difficulty and discrimination computed on real student responses) for an operational national mathematics item bank spanning Grades 4–12 and two tracks, together with the expert human cognitive classification assigned at item-writing time. We use these data to test how well the theoretical cognitive classification aligns with the empirical psychometric behaviour of items, and to benchmark an independent LLM classification against the expert coding. To our knowledge this is among the first studies to combine all three elements—expert taxonomy, real psychometric behaviour, and an LLM classifier—in a K–12 mathematics context.

2. Research Problem, Questions, and Hypotheses

The research problem concerns a potential gap between what is planned in the blueprint (the intended cognitive level) and what happens when items reach students (their empirical difficulty and discrimination). If empirical difficulty does not track the intended cognitive level, the construct validity of the blueprint is weakened, and item-writing procedures may need revision. A parallel, practical question is whether current LLMs can be trusted as an auxiliary—not a replacement—for expert cognitive classification.

The study addressed five questions:

RQ1. How well does the expert human cognitive classification agree with an independent LLM classification of the same items?

RQ2. Does this agreement vary by grade or track (General vs. Advanced)?

RQ3. Do empirical difficulty values order monotonically with the intended cognitive level (remember/understand < apply/analyse < higher order)?

RQ4. Does item discrimination differ by cognitive level or by track?

RQ5. Is there evidence of item-level differences between the General and Advanced tracks?

Corresponding null hypotheses ($\alpha = .05$), aligned one-to-one with the research questions:

H0₁ (RQ3): Empirical difficulty does not differ across intended Bloom bands.

H0₂ (RQ4): There is no relationship between intended Bloom level and item discrimination.

H0₃ (RQ5): There is no item-level difference in performance between the General and Advanced tracks.

H0₄ (RQ1): Human and AI cognitive classifications agree no better than chance.

3. Method

3.1 Design

The study used a descriptive, psychometric, secondary-data design. No experimental manipulation of testing conditions occurred; the analysis relies on archival item-analysis reports from administered centralized examinations.

3.2 Sample of Items

The item sample comprised 265 four-option multiple-choice items drawn from the End-of-Term 1 (2025–2026) centralized mathematics examinations. Thirteen test forms were included: Grades 4–8 in the General track, and Grades 9–12 in both the General and Advanced tracks. Each form contributed 20 items, except the Grade 5 form (25 items). The composition of the thirteen forms is summarized in Table 1. Items were classified by the assessment team, at construction, into the six revised-taxonomy levels; for analysis these were collapsed into three blueprint bands: B1 remember/understand, B2 apply/analyse, and B3 higher-order (evaluate/create).

Table 1. Composition of the Thirteen Centralized Examination Forms (N = 265 Items)

Grade (Form)	Track	No. of items
Grade 4	General	20
Grade 5	General	25
Grade 6	General	20
Grade 7	General	20
Grade 8	General	20
Grade 9	General	20
Grade 9	Advanced	20
Grade 10	General	20
Grade 10	Advanced	20
Grade 11	General	20
Grade 11	Advanced	20
Grade 12	General	20
Grade 12	Advanced	20
Total — 13 forms	—	265

Note. Grades 4–8 were offered in the General track only; the Advanced track operates from Grade 9. All forms carried 20 items except Grade 5 (25).

3.3 Data and Measures

For every item the following were extracted from the official item-analysis reports: grade, track, mathematical domain, the intended Bloom level, the empirical difficulty index (p-value: proportion of students answering correctly, so that higher = easier), the point-biserial discrimination index, the high–low (27%) discrimination index, and an item-reliability index (Crocker & Algina, 2008; Haladyna & Rodriguez, 2013). Because the reports are aggregated, person-level response vectors were not available; consequently option-level distractor analysis, item-response-theory calibration (de Ayala, 2022), and person-level differential item functioning (DIF) (Osterlind & Everson, 2009) were outside the scope of the present data and are identified as future work.

3.4 AI Classification Procedure

An independent cognitive classification was produced by a large language model (Claude, Anthropic). The model classified each item's learning-outcome statement into the six revised-taxonomy levels by applying the same operational definitions used in the official blueprint, without access to the human classification. The model's output was then compared with the expert coding. This design treats the LLM as a single independent rater and evaluates human–AI convergence; it does not present the AI as ground truth.

To support reproducibility, the following generation details are reported: the classification was produced with Claude (Anthropic), using a fixed system prompt (reproduced verbatim in Appendix A); the generation temperature was set to 0 to minimise stochastic variation; and each item was classified in a single, independent request with no conversation history carried across items.

Intra-rater consistency of the model. We note that the present classification treated the model as a single independent rater and did not include repeated classification of the same items. To further characterise the LLM as a classifier, we recommend an intra-rater reliability check in which each learning-outcome statement is re-classified across two or three independent runs under identical settings and run-to-run agreement is quantified (e.g., Fleiss' κ); such a check would separate genuine human–AI disagreement from run-to-run instability of the model. We flag this as a priority for the multi-model extension identified in the Limitations, rather than reporting it on the present, single-pass data.

3.5 Statistical Analysis

Analyses were conducted in Python (SciPy, scikit-learn, statsmodels). Human–AI agreement was quantified with raw agreement and Cohen's κ (unweighted, and linearly and quadratically weighted to respect the ordinal structure), overall and perform. The observed band distribution was compared with the 40/50/10 blueprint target using a chi-square goodness-of-fit test. Because difficulty and discrimination departed from normality and cell sizes were unequal, the Bloom–difficulty and Bloom–discrimination relationships were tested with the Kruskal–Wallis test, complemented by one-way ANOVA and by Spearman rank correlations between the ordinal cognitive rank and each psychometric index; ε^2 and η^2 were reported as effect sizes. Because RQ3 specifically concerns an ordered (monotonic) trend across the three cognitive bands, the Jonckheere–Terpstra trend test was adopted as the primary test for RQ3 and is reported alongside the Kruskal–Wallis test. Track differences (Grades 9–12) were tested with the Mann–Whitney U test. The interpretive benchmarks of Landis and Koch (1977) were used for κ .

3.6 Ethics and Data Use

This study analysed secondary, aggregate item-analysis records generated in the ordinary course of an operational national assessment programme. The data contained no personally identifiable student information: only item-level statistics (difficulty, discrimination, reliability, cognitive level, grade, track, and domain) were used, and no individual response vectors, names, or identifiers were accessed at any point. The lead author, in his official capacity as senior specialist and head of the mathematics item-construction team, is an authorised custodian of these operational assessment records, and institutional permission was in place for the secondary, non-identifiable use of the aggregated statistics for methodological research. Because the analysis used only de-identified, aggregate operational data and involved no intervention with human participants, it did not require individual-participant informed consent. The authors confirm that the work was conducted in accordance with the ethical standards of the responsible institution and with the journal's ethical guidelines for authors.

4. Results

4.1 Blueprint Fidelity: Item Distribution across Cognitive Bands

Across the 265 items, 41.5% fell in the remember/understand band, 52.8% in apply/analyse, and 5.7% in the higher-order band, against blueprint targets of 40/50/10. The higher-order band was thus underrepresented by roughly four percentage points. A chi-square goodness-of-fit test indicated that the departure from the target proportions approached but did not reach significance, $\chi^2(2) = 5.57$, $p = .062$ (Figure 1). The under-representation of higher-order items is the single clearest blueprint-fidelity concern.

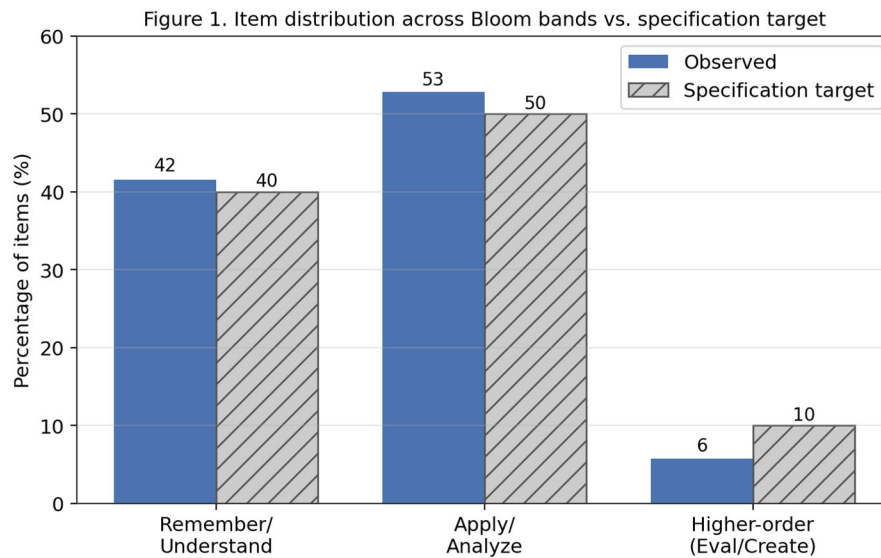


Figure 1. Observed Item Distribution across Cognitive Bands Versus the 40/50/10 Blueprint Target

4.2 RQ1–RQ2: Human–AI Classification Agreement

Human and AI classifications agreed on 51.7% of items at the six-level resolution and 58.1% at the band level. Cohen's κ was .22 (unweighted), .21 (linear-weighted) and .19 (quadratic-weighted) at six levels, and .17 at band level—values that fall in the “fair” range of Landis and Koch (1977). The quadratically weighted κ (.19) fell slightly below the unweighted κ (.22). This is not anomalous once the pattern of errors is considered: quadratic weights penalise disagreements in proportion to the square of their distance on the ordinal scale, so a classifier is rewarded only when its errors are near diagonal (one level off). The sub-unweighted value therefore indicates that the residual disagreements were not confined to adjacent levels—a small number of items were displaced across two or more levels (for example, higher-order items re-assigned to remember/understand)—and these few large errors are penalised heavily enough under quadratic weighting to outweigh the credit earned by the many near-diagonal cases. Agreement varied widely by form (κ from $-.24$ for Grade 9 General to $.38$ for Grade 4 General), with no systematic advantage for either track (RQ2).

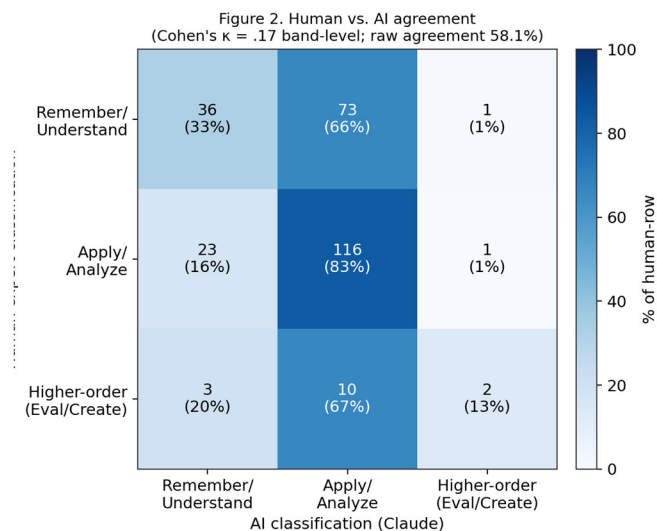


Figure 2. Human (Rows) Versus AI (Columns) Band Classifications; Counts and Row Percentages

The confusion matrix (Figure 2) shows the source of disagreement: the model systematically re-assigned items toward the apply/analyse band. Of 110 items the experts coded as remember/understand, the model moved 73 into apply/analyse; of 15 higher-order items, the model recovered only 2 and placed 10 in apply/analyse. In other words, the LLM compressed the cognitive scale toward the middle band and rarely recognized higher-order demand. H0₄ was rejected (agreement exceeded chance), but the level of agreement is too low to support substitution of human judgement.

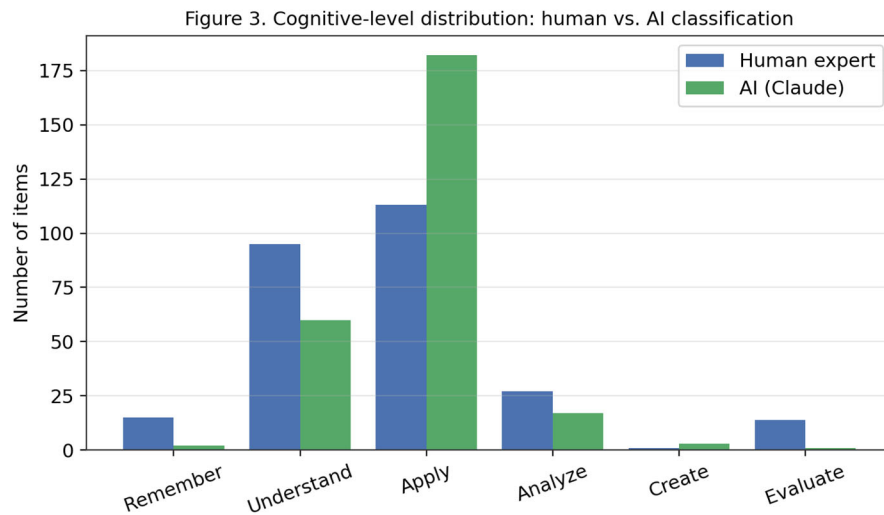


Figure 3. Six-Level Cognitive Distribution: Expert Versus AI Classification. The model over-produces “Apply” and under-produces higher-order levels

4.3 RQ3: Cognitive Level and Empirical Difficulty

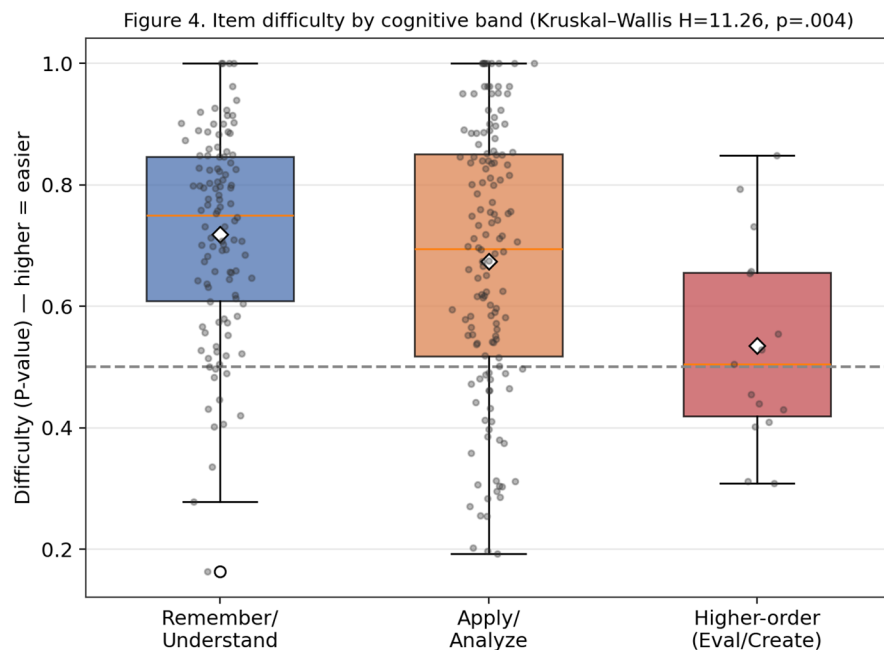


Figure 4. Item Difficulty (P-Value) by Cognitive Band. Diamonds Mark Means; the Dashed Line Marks P = .50.

Mean difficulty declined monotonically across bands—B1 remember/understand $p = .718$ ($SD = .17$), B2 apply/analyse $p = .674$ ($SD = .22$), B3 higher-order $p = .535$ ($SD = .17$)—so that higher-order items were, on average, the hardest. The difference was statistically significant (Kruskal–Wallis $H = 11.26$, $p = .004$; ANOVA $F(2, 262) = 6.12$, $p = .003$), and the Spearman correlation between cognitive rank and difficulty was negative and significant ($\rho = -.16$, $p = .009$), confirming the hypothesized direction. The Jonckheere–Terpstra test, adopted as the primary ordered-alternative test for RQ3, confirmed a significant decreasing trend in difficulty across the ordered bands ($J = 7943$, $z = -2.56$, one-sided $p = .006$), in agreement with the Kruskal–Wallis and Spearman results. H_{01} was therefore rejected. Crucially, however, the effect was weak: $\varepsilon^2 = .043$ and $\eta^2 = .045$, meaning intended cognitive level explained under 5% of the variance in difficulty. The blueprint assumption is thus directionally supported but of limited practical magnitude (Figure 4).

4.4 RQ4: Cognitive Level and Discrimination

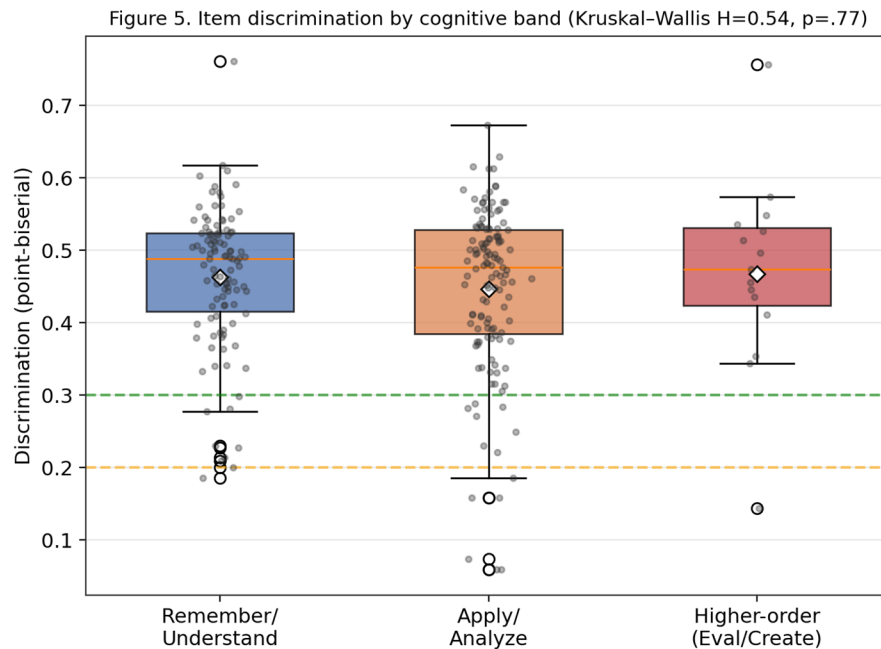


Figure 5. Item Discrimination (Point-Biserial) by Cognitive Band; the Difference is Non-Significant

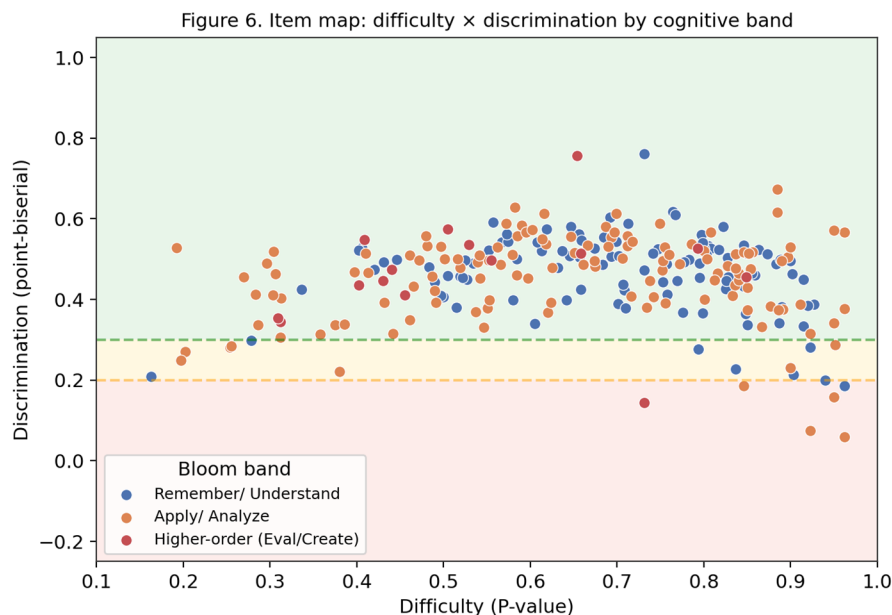


Figure 6. Item Map of Difficulty $\times$ Discrimination, Coloured by Cognitive Band; Shaded Zones Mark Discrimination-Quality Thresholds

In contrast to difficulty, discrimination was flat across cognitive bands: mean point-biserial was .463, .447 and .468 for B1, B2 and B3 respectively. Neither the Kruskal–Wallis test ($H = 0.54$, $p = .77$) nor the Spearman correlation ($\rho = -.03$, $p = .65$) was significant, so H_{02} was retained. Intended cognitive level carried no information about how well an item separated high- from low-ability students. Moreover, difficulty and discrimination were themselves essentially uncorrelated across the pool ($\rho = -.05$, $p = .43$), and most items discriminated acceptably (the majority above the .30 benchmark), as shown in the item map (Figures 5 and 6).

4.5 RQ5: Track Comparison (Grades 9–12)

Restricting attention to Grades 9–12, where both tracks operate, Advanced-track items were slightly easier on average than General-track items ($p = .693$ vs. $.639$), but the difference was not significant (Mann–Whitney $U = 2710$, $p = .095$). Discrimination was virtually identical (.475 vs. .470; $U = 3216$, $p = .96$). At the item level, then, the two tracks did not differ reliably in psychometric behaviour, and H_{03} was retained. Because grade and track are partially confounded in this design—the Advanced track is offered only in Grades 9–12, so track cannot be separated from grade-level content—this comparison is necessarily exploratory and descriptive: the observed track differences cannot be given a causal interpretation and are reported as such. Per-form difficulty and discrimination profiles appear in Figure 7, which also highlights ceiling effects in the Grade 4 and Grade 5 forms (mean $p = .96$ and $.88$), where several items had $p = 1.00$ and therefore no computable discrimination.

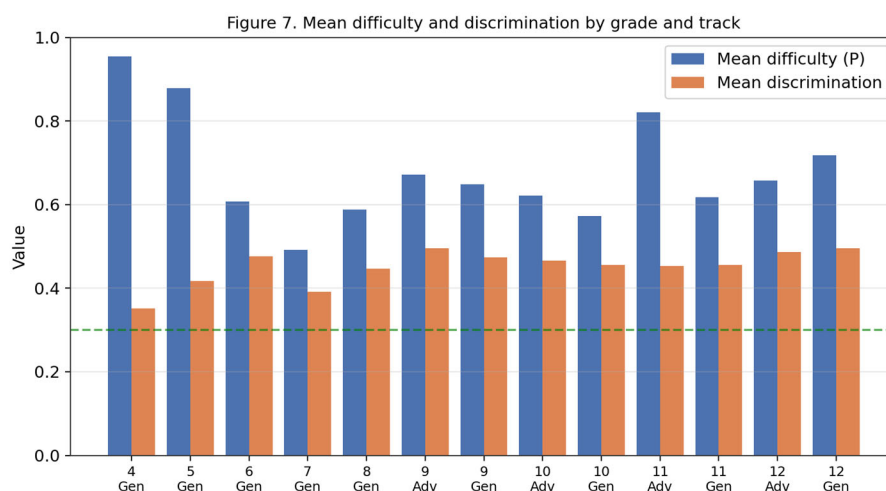


Figure 7. Mean Difficulty and Discrimination by Grade and Track; the Dashed Line Marks the .30 Discrimination Benchmark

5. Discussion

Three findings stand out. First, the blueprint's core assumption—that higher cognitive demand implies greater empirical difficulty—was supported in direction but not in strength. Higher-order items were reliably the hardest, yet intended cognitive level accounted for under 5% of the variance in difficulty. This pattern echoes the mixed picture in the medical-education literature, where difficulty tracks taxonomy weakly and item writers' expectations often diverge from students' experience (Kibble & Johnson, 2011; Zaidi et al., 2018). For an operational blueprint, the practical implication is that cognitive level is a coarse and unreliable proxy for difficulty; difficulty targets should be managed directly through pretesting rather than inferred from taxonomic coding.

Second, discrimination was unrelated to cognitive level. The intuition that higher-order items discriminate better received no support—consistent with the possibility, noted by Kim et al. (2012), that difficulty and discrimination decouple from cognitive level in different ways. The near-zero difficulty–discrimination correlation observed here ($\rho = -.05$, n.s.) reinforces this reading; within this pool the two indices behave as largely independent item properties rather than as two expressions of a single underlying trait. Practically, this means an item cannot be assumed to discriminate well merely because it is difficult (or poorly because it is easy); each property must be evaluated on its own field evidence. Because discrimination is the property most directly tied to an item's measurement value, this result reassures on one point (higher-order items are not psychometrically inferior) while cautioning on another (they are not superior either, or cannot be justified on discrimination grounds alone).

Third, the LLM classifier agreed with expert coders only fairly, and did so in a specific, informative way: it compressed the cognitive scale toward apply/analyse and rarely detected higher-order demand. This central-tendency behaviour is best described as a majority-/modal-class bias rather than as a literal property of model “compression.” Two mechanisms plausibly contribute. First, apply/analyse is the modal category in typical instructional and training corpora, and next-token likelihood favours high-frequency labels, so a decision made under uncertainty is pulled toward the middle band. Second, the classifier operated on brief learning-outcome statements whose surface wording (verbs such as “solve,” “calculate,” “apply”) is highly similar across bands; when the discriminating cues are thin, the model defaults to the most probable label. On this account the compression is a calibration artefact of label priors and truncated input, not evidence that the model cannot in principle represent higher-order demand. For a national system, this means an LLM is presently unsuitable as an autonomous classifier for high-stakes blueprints but may still be useful as a first pass flagging tool whose disagreements with human coders identify items worth a second expert look. The model’s reluctance to assign higher-order labels also parallels the human under-production of higher-order items (5.7% vs. a 10% target). Given the small higher-order band ($n = 15$), we advance this only as a suggestive observation—rather than a confirmed finding—that higher-order demand may be comparatively difficult to specify and to recognise for humans and machines alike.

Taken together, the results support a measured stance: Bloom’s taxonomy remains valuable for communicating intended cognitive coverage, but it is a weak predictor of empirical item behaviour and should not be treated as a substitute for direct psychometric evidence. AI tools can assist, but not replace, expert judgement.

6. Limitations

Several limitations should temper interpretation. (a) The data were aggregate item-analysis records; without person-level responses we could not conduct option-level distractor analysis, IRT calibration, or person-level Mantel–Haenszel DIF—analyses we recommend as the natural next step. (b) The higher-order band was small (15 items), limiting power for that band and widening its confidence intervals. (c) Grade 4 and Grade 5 forms showed ceiling effects, with zero-variance items whose difficulty and discrimination are uninformative. (d) A single LLM served as the AI rater; multi-model comparison (e.g., additional independent systems) would strengthen the AI-benchmark component. (e) Item content was analysed through official learning-outcome statements rather than full item stems and answer options. This is a substantive constraint on the human–AI comparison in RQ1: the low agreement may be driven in part by the reduced, paraphrastic input given to the model rather than by an inherent limitation of the LLM, and a robustness re-run using full stems and options may materially change the reported agreement. We therefore interpret the RQ1 agreement figures as a lower bound conditioned on outcome-statement input. (f) Grade and track are partially confounded (the Advanced track exists only in Grades 9–12), so the RQ5 track comparison is exploratory and descriptive rather than causal. (g) The study is confined to one subject, one national system, and one testing window, so generalization beyond these bounds should be cautious.

7. Conclusion

Using 265 operational national mathematics items across Grades 4–12, this study found that intended cognitive level, coded with Bloom’s taxonomy, is a weak (though directionally correct) predictor of empirical difficulty, is unrelated to discrimination, and is only fairly reproduced by a current large language model. National assessment programmes should therefore continue to use the taxonomy for coverage planning while validating difficulty and discrimination directly through pretesting, invest in item-writer training for higher-order items (which are both under-produced and hard to classify), and treat AI classification as a decision-support flag rather than an authoritative judgement. Future work should repeat the analysis with person-level data to enable distractor analysis, IRT, and formal DIF, and should benchmark multiple AI systems against expert coders.

References

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Longman.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage.

- Cunnington, J. P. W., Norman, G. R., Blake, J. M., Dauphinee, W. D., & Blackmore, D. E. (1997). Applying learning taxonomies to test items: Is a fact an artefact? In A. J. J. A. Scherpbier et al. (Eds.), *Advances in medical education* (pp. 89–92). Springer.
- de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Routledge.
- Duong-Trung, N., et al. (2024). Bloom LLM: Large language models-based question generation combining supervised fine-tuning and Bloom's taxonomy. In *Proceedings of the International Conference on Intelligent Tutoring Systems*.
- Elkins, S., & Kochmar, E. (2024). How teachers can use large language models and Bloom's taxonomy to create educational quizzes. *arXiv*. Retrieved from <https://arxiv.org/abs/2401.05914>
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Huber, T., & Niklaus, C. (2025). LLMs meet Bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5211–5246). Association for Computational Linguistics.
- Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education*, 35(4), 396–401. <https://doi.org/10.1152/advan.00024.2011>
- Kim, M.-K., Patel, R. A., Uchizono, J. A., & Beck, L. (2012). Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American Journal of Pharmaceutical Education*, 76(6), Article 114. <https://doi.org/10.5688/ajpe766114>
- Kıyak, Y. S., & Emekli, E. (2024). *Reviewing the reliability of ChatGPT for generating multiple-choice questions: A systematic review*. Postgraduate Medical Journal. Advance online publication.
- Koçdar, S., Karadağ, N., & Şahin, M. D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology*, 15(4), 16–25.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Sage.
- Scaria, N., Chenna, S. D., & Subramani, D. (2024). Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation. In *International Conference on Artificial Intelligence in Education*.
- Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8(1), 34–52. <https://doi.org/10.1080/15434303.2010.536924>
- Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relationship with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology*, 9, Article 1585. <https://doi.org/10.3389/fpsyg.2018.01585>
- Zaidi, N. L. B., Grob, K. L., Monrad, S. M., et al. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's taxonomy work? *Academic Medicine*, 93(6), 856–859. <https://doi.org/10.1097/ACM.0000000000002087>

Appendix A. LLM classification prompt (reproducibility)

The following system prompt was used to elicit an independent cognitive classification from the model. It reproduces the operational definitions of the six revised-taxonomy levels used in the official blueprint and constrains the output to a single level per item. For blueprint reporting, the six levels map onto the three bands used throughout this study: B1 remember/understand (target 40%), B2 apply/analyse (target 50%), and B3 higher order = evaluate/create (target 10%).

“You are an expert in educational assessment. Classify the following mathematics learning-outcome statement into exactly one of the six levels of the revised Bloom’s taxonomy, using these operational definitions: (1) Remember — recall or recognise facts, terms, symbols, or previously learned procedures without transformation (verbs: define, list, identify, recall, name). (2) Understand — explain, interpret, or represent a mathematical idea in another form; grasp meaning (verbs: explain, interpret, summarise, classify, represent). (3) Apply — carry out or use a known procedure, rule, or algorithm in a routine or slightly varied situation (verbs: solve, compute, calculate, use, execute). (4) Analyse — break material into parts, examine relationships, distinguish relevant from irrelevant information, or compare structures (verbs: differentiate, organise, compare, deconstruct, relate). (5) Evaluate — make and justify judgements against criteria, check reasoning, or select the best of several approaches (verbs: judge, justify, critique, verify, defend). (6) Create — combine elements into a new coherent whole, design a procedure, or generate an original solution or model (verbs: design, construct, formulate, generate, produce). Return only the single level name, with no explanation. Learning outcome: <text>.”

Author Statements

Author(s) declared the statement below:

Acknowledgments

The authors gratefully acknowledge the administration of City University Ajman for its continuous support and encouragement, which made this work possible.

Author Contributions

Author contributions follow the CRediT (Contributor Roles Taxonomy). Use only the contributor roles applicable to this manuscript.

Conceptualization: Dr. Rafat Altaweel, <https://orcid.org/0000-0003-2216-3911>

Methodology: Dr. Rafat Altaweel, <https://orcid.org/0000-0003-2216-3911>; Dr. Moustafa Kamal Moussa, <https://orcid.org/0000-0002-3151-8836>

Formal analysis: Dr. Rafat Altaweel, <https://orcid.org/0000-0003-2216-3911>

Investigation: Dr. Rafat Altaweel, <https://orcid.org/0000-0003-2216-3911>

Writing – original draft: Dr. Rafat Altaweel, <https://orcid.org/0000-0003-2216-3911>

Writing – review & editing: Dr. Moustafa Kamal Moussa, <https://orcid.org/0000-0002-3151-8836>

Supervision: Dr. Moustafa Kamal Moussa, <https://orcid.org/0000-0002-3151-8836>

Consent for Publication

All authors have read and approved the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Informed Consent

Not applicable. The study used only secondary, de-identified, aggregate item-analysis records and involved no human participants.

Ethics Approval

The study analysed de-identified, aggregate item-analysis records from an operational national assessment programme; no personally identifiable student data were accessed and no intervention with human participants took place. The lead author is an authorised custodian of these operational records, and institutional permission was in place for their secondary, non-identifiable use for methodological research (see Section 3.6).

Data Availability Statement

The aggregate item-analysis statistics supporting the findings may be made available from the corresponding author on reasonable request, subject to the data-governance rules of the national assessment programme.

Use of Artificial Intelligence

The core academic content of this manuscript—its research questions, design, analysis, interpretation, and writing—was created entirely by the authors. A large language model (Claude, Anthropic) was used solely as an object of study, that is, as one independent rater whose classifications were benchmarked against expert human coding; it was not used to generate the scientific content, arguments, or text of the manuscript. Any incidental language assistance did not extend to the creation of research content.

Supplementary Materials

No supplementary materials are available.

Journal Information

The Editorial Board declared the below statement on behalf of the journal and publisher.

Ethics Statement

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Preservation Policy

This journal is archived by Library and Archives Canada (LAC).

Open Access

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0). Please find details at <https://creativecommons.org/licenses/by/4.0>.

Copyright © The Author(s)

Copyright remains with the author(s). The author(s) grant the journal the right of first publication, and the article is published under the terms of the Creative Commons Attribution (CC BY 4.0) License.