

Human-in-the-Loop Gemini Item Generation for Large-Scale MCQ Banks

Melchor Sánchez-Mendiola, Milton García-Lima, Elibidú Ortega-Sánchez,
Manuel García-Minjares, and Enrique Buzo-Casanova

Coordination of Evaluation, Innovation and Educational Development (CEIDE), National Autonomous University of Mexico (UNAM), Mexico City, Mexico

melchorsm@unam.mx; milton_garcia@ceide.unam.mx; elibidu_ortega@ceide.unam.mx;
manuel_garcia@ceide.unam.mx; enrique_buzo@ceide.unam.mx

Abstract

Large-scale assessment programs need secure, content-aligned item banks, yet faculty item writing is resource intensive and variable. We report a work-in-progress, naturalistic comparison of 2,405 Spanish-language multiple-choice questions developed at the National Autonomous University of Mexico: 986 drafted with ten customized Gemini Gems and 1,419 drafted by faculty. A Gem is a reusable Gemini configuration with persistent instructions and optional knowledge files. The subject-specific Gems embedded assessment blueprints, diagnostic gatekeeping, item-writing constraints, and a standardized output schema. Three-member faculty committees reviewed every candidate item using institutional criteria. We summarized review activity with a validation-friction index (VFI; recorded comments divided by submitted items) and analyzed recurring comment patterns. Overall VFI was 1.84 for Gemini-drafted items and 2.30 for faculty-drafted items; the AI source had lower VFI in three of four knowledge areas. Gemini drafts were more consistent in blueprint adherence and structure, whereas faculty drafts often offered richer contextualization and misconception-based distractors. Because comments mixed confirmations and corrections and no examinee data were available, VFI is a process indicator, not evidence of psychometric equivalence. Findings support a human-in-the-loop workflow combining structured generation, automated checks, expert review, and post-administration validation.

1 Introduction

Large-scale assessment programs depend on item banks that are sufficiently broad, secure, and aligned with intended content and cognitive demands. Producing those banks requires much more than subject-matter expertise, each multiple-choice question (MCQ) must represent

the intended construct, use clear language, contain a defensible key, and include plausible distractors without grammatical, lexical, or visual clues. Weaknesses at this stage can introduce construct-irrelevant variance before an item is ever administered (American Educational Research Association et al., 2014; Haladyna et al., 2002; Haladyna and Rodriguez, 2013).

Traditional faculty authoring supports disciplinary authenticity and can draw on knowledge of common student misconceptions. It is also labor intensive, difficult to standardize across many authors, and challenging to scale when programs need multiple test forms or frequent bank renewal. Earlier automatic item generation (AIG) approaches addressed this problem through cognitive models, item models, templates, and content coding. These approaches demonstrate that human judgment and computational generation can be combined to improve production capacity and test security (Circi et al., 2023; Gierl et al., 2022).

Large language models (LLMs) extend AIG by producing fluent items from natural-language instructions and curricular materials. The resulting flexibility is attractive, but fluency is not equivalent to measurement quality. Reviews and empirical comparisons have identified usable AI-generated items alongside ambiguous keys, weak or implausible distractors, inappropriate difficulty, and content-alignment problems (Artsi et al., 2024; Grévisse et al., 2024; Kaldaras et al., 2024). Recent studies using examinee data have also reported mixed results: AI-generated items may be comparable on some indices yet easier, differently discriminating, or more readily detectable than human-authored items (Kaya et al., 2025; Law et al., 2025; Linde et al., 2026). Across this literature, expert review remains a central safeguard.

This work-in-progress paper examines a different but complementary stage of the validity

process: pre-operational item review in a large, cross-disciplinary, Spanish-language item-bank project. We compare items drafted with customized Google Gemini Gems and items written by faculty at the National Autonomous University of Mexico (UNAM). The study asks three questions: How did the two item sources differ in recorded expert-review activity across four knowledge areas? What validity-relevant strengths and weaknesses recurred in reviewer comments? What workflow implications follow for responsible, scalable, human-in-the-loop item development?

2 Institutional Context

UNAM's Coordination of Evaluation, Innovation and Educational Development (CEIDE) developed the project to support a large institutional MCQ bank spanning four knowledge areas: Physical-Mathematical Sciences and Engineering (PMSE), Biological, Chemical, and Health Sciences (BCHS), Social Sciences (SS), and Humanities and Arts (HA). The assessment blueprint specified 120 items per area for each of five versions, or approximately 2,400 items. The realized comparison dataset contained 2,405 candidate items across ten subjects: Mathematics, Spanish, Physics, Chemistry, Biology, Literature, Geography, Philosophy, Mexican History, and World History.

Two authoring routes contributed items to the same institutional development process. Faculty authors wrote items using the applicable table of specifications. In parallel, ten subject-specific Gemini Gems generated drafts from the same types of inputs: subject, topic, expected learning outcome, intended cognitive level, item format, and media requirements. The comparison was therefore a retrospective, naturalistic process evaluation rather than a randomized experiment. Item source was not assigned within matched blueprint cells, and the number of AI- and faculty-drafted items differed by area.

3 Gemini Gem Item-Generation Architecture

Google Gemini 3.0 Pro was used. Google describes a custom Gem as a saved configuration that can retain instructions and optional knowledge files for repeated use (Google, n.d.). In this project, a Gem differed from a one-off prompt in three practical respects: the same instructions were reused across sessions, curricular and item-writing documents could be

attached as persistent knowledge, and changes to the central instruction set could be propagated to subsequent generations. The Gems were configured rather than fine-tuned; no new foundation model was trained.

The available project records identify the application as Gemini Gems but do not preserve an auditable foundation-model version identifier. Accordingly, this paper evaluates the configured workflow as implemented, not a frozen-model benchmark. Future replications should log the model label, generation dates, account tier, prompt version, knowledge-file versions, and any accessible inference settings.

The architecture combined four design components. First, each Gem received a subject and assessment-expert persona. Second, a diagnostic gate required the user to specify the relevant blueprint information before generation; the explicit instruction was not to generate an item until the diagnostic phase was complete. Third, the Gems embedded institutional item-writing constraints, including Mexican Spanish, positive wording, one defensible key, balanced options, plausible distractors, structured metadata, and LaTeX-compatible notation for scientific and mathematical content. For database standardization, the draft key was placed in option A and could be randomized after import. Fourth, when an item required a figure, the Gem produced the editable text item and a separate image-generation specification rather than rendering the entire item as an image.

3.1 Prompt Scaffold and Calibration

Table 1 presents an abridged English translation of the documented prompt scaffold. It does not reproduce confidential operational content or any item. The design sought to make curricular grounding and output constraints explicit before the model generated a draft. This is consistent with recommendations that AI-assisted test development document the construct, intended use, prompt design, semantic alignment, and human validation steps (Kaldaras et al., 2024; Suárez-Álvarez et al., 2026).

The architecture also specified an iterative calibration loop. When reviewers identified a recurring methodological or stylistic problem, the project team could revise the central instructions by adding a restriction or a counterexample. The first production audit identified priorities for subsequent prompt revisions, including shorter stems, more symmetric option length, less lexical overlap

between stem and key, and distractors grounded more directly in common misconceptions. Because the available records describe the calibration mechanism but do not contain a

complete prompt-version history, the present analysis does not estimate improvement across prompt iterations.

Component	Illustrative translated instruction	Quality-control function
Persona and scope	You are a [subject] instructor and an expert in educational assessment. Work only within the specified curriculum and assessment standards.	Stabilizes disciplinary role, register, and intended task.
Diagnostic gate	Do not generate an item until the user supplies the topic, expected learning outcome, cognitive level, item type, and media requirements.	Reduces premature generation and forces blueprint completion.
Item constraints	Use Mexican Spanish and positive wording; provide one defensible key and three plausible distractors based on common student errors; avoid grammatical, lexical, and visual clues.	Operationalizes institutional item-writing rules.
Technical format	Use a fixed metadata schema; place the draft key in option A for import and later randomization; use LaTeX-compatible notation for mathematics and science.	Supports parsing, database transfer, and reproducible metadata.
Multimodal handling	When a visual is required, return the text item plus a separate, precise image specification rather than rendering the complete item as an image.	Preserves editable item structure and separates content from rendering.
Calibration	When a recurrent flaw is found, add a specific restriction or a counterexample to the central instructions and retain a prompt revision record.	Enables centralized, auditable refinement across future generations.

Table 1: Abridged and translated Gemini Gem instruction scaffold. Bracketed terms were replaced with the applicable subject and blueprint information.

4 Data and Methods

The unit of analysis was a candidate MCQ submitted to the institutional review process. Of the 2,405 items, 986 were drafted with Gemini Gems and 1,419 were drafted by faculty. Items were distributed across the four areas as shown in Table 3. The project records contained item-source counts, the total number of recorded committee comments, and area-level qualitative audits of those comments.

4.1 Faculty Review Process

Every candidate item underwent external content review by a committee of three faculty experts in the relevant subject. Reviewers assessed the item against institutional criteria covering blueprint alignment, cognitive level, clarity, technical

correctness, a unique defensible key, distractor plausibility, and absence of unintended clues. Committees recorded one or more comments in the item-management workflow. Importantly, the comment field included both positive confirmations (for example, that an item met a criterion) and requests for revision. Consequently, a comment count cannot be interpreted as an error count.

The review criteria are summarized in Table 2. They functioned as pre-operational content and technical quality checks. They did not establish score reliability, difficulty, discrimination, differential item functioning, or validity evidence based on examinee response processes or relationships with external variables.

Review dimension	Primary question for the faculty committee
Blueprint alignment	Does the item assess the specified subject, topic, and expected learning outcome?
Cognitive demand	Does the item elicit the intended level of knowledge, comprehension, or application?
Clarity and language	Is the stem understandable, concise, grammatically sound, and appropriate in Mexican Spanish?
Key validity	Is there one clearly defensible correct response?
Distractor quality	Are incorrect options plausible, parallel, and related to likely misconceptions?
Technical accuracy	Are terminology, notation, units, formulas, and supporting media correct?
Absence of clues	Do option length, wording, grammar, or lexical overlap avoid revealing the key?

Table 2: Institutional criteria used by subject-matter faculty committees during pre-operational item review.

4.2 Measures and Analysis

We calculated a validation-friction index (VFI) for each area and item source: $VFI = \text{total recorded validation comments} / \text{total submitted items}$. Lower values denote fewer recorded reviewer comments per item and therefore lower comment density within the workflow. VFI is used here as an administrative process indicator, not as a direct measure of item quality or validity. We also report the faculty-to-AI VFI ratio to describe relative comment density.

For the qualitative component, the project team organized recurring comments deductively using the institutional review criteria and compared patterns across knowledge areas and item sources. The available materials contained aggregate counts and area-level summaries rather than a complete item-by-comment analytic file. Accordingly, the current paper reports descriptive results without hypothesis tests, formal inter-rater agreement estimates, or item-level regression models.

5 Results

Across all areas, Gemini-drafted items received 1,810 recorded comments for 986 items ($VFI = 1.84$), whereas faculty-drafted items received 3,258 comments for 1,419 items ($VFI = 2.30$). The faculty-to-AI VFI ratio was 1.25, indicating 25.1% more recorded comments per faculty item in the aggregate. This overall comparison is descriptive and does not adjust for area, subject, cognitive level, or item format.

The direction and magnitude of the difference varied by area. In BCHS, VFI was 2.25 for Gemini drafts and 2.90 for faculty drafts; the faculty rate was 28.9% higher. In SS, the corresponding values were 1.57 and 2.16, a 37.6% higher faculty rate. HA showed the largest contrast: 1.13 for Gemini and 2.36 for faculty, or a 108.8% higher faculty rate. PMSE was the exception, with VFI of 2.01 for Gemini and 1.78 for faculty; the Gemini rate was 12.9% higher. Table 3 presents the underlying counts.

Area	AI items	AI comments	AI VFI	Faculty items	Faculty comments	Faculty VFI
PMSE	298	598	2.01	301	536	1.78
BCHS	307	691	2.25	299	866	2.90
SS	206	324	1.57	394	853	2.16
HA	175	197	1.13	425	1,003	2.36
Overall	986	1,810	1.84	1,419	3,258	2.30

Table 3: Candidate items, recorded faculty-review comments, and validation-friction index (VFI) by knowledge area and authoring source.

5.1 Recurring Qualitative Patterns

Across areas, the most consistent strength of Gemini drafts was adherence to the table of specifications. Review summaries described standardized syntax, stable metadata, and more consistent matching of expected learning outcomes and requested cognitive levels. The principal AI-related concerns were not uniform across disciplines. In PMSE, reviewers identified visual asymmetry when the key was longer or more detailed than the distractors, as well as lexical clues created by repeating words from the stem. In BCHS, some distractors were thematically distant, allowing superficial elimination rather than discrimination based on domain knowledge. In SS, comments focused on micro-format conventions and occasional elevation of cognitive complexity. In HA, most comments on Gemini items concerned fluency or

stylistic refinement rather than substantive content.

Faculty-authored items showed some complementary strengths. Reviewers noted more natural disciplinary context, particularly clinical or pedagogical framing in BCHS, and stronger use of authentic misconceptions in distractor construction in some PMSE items. At the same time, faculty drafts were more heterogeneous. Recurring concerns included drift from the specified content or cognitive level, overly complex stems, implicit missing conditions, ambiguous instructions, grammatical clues, implausible distractors, and occasional incorrect or debatable keys. These patterns suggest that structured AI generation and faculty expertise address different sources of item-development risk.

5.2 Human-Platform Interface Errors

The PMSE audit highlighted a separate source of quality loss: transfer from generated text into the item-management platform. Review records traced multiple problems in exponents, subscripts, equation symbols, and graph images to manual transcription or platform entry rather than to the original Gemini output. Similar notation-transfer issues were noted in other technical subjects. This distinction matters because a workflow-level evaluation that assigns every downstream defect to the content generator will misidentify the source of error. It also suggests that parsers, direct database upload, and automated notation checks may be as important as prompt refinement.

6 Discussion

This work-in-progress study provides preliminary evidence that a persistent, blueprint-grounded Gem configuration can support large-scale item drafting with manageable expert-review burden. Gemini-drafted items had lower comment density in three of four knowledge areas and in the aggregate. The result should not be read as proof that the AI items were psychometrically superior. Rather, it suggests that explicit constraints and standardized output may reduce some forms of pre-operational variation that faculty committees otherwise need to identify and correct.

The findings are consistent with prior literature in two respects. First, LLMs can produce plausible MCQ drafts efficiently when prompts specify content and format. Second, human review remains necessary because recurring problems with keys, distractors, difficulty, and alignment do not disappear (Artsi et al., 2024; Grévisse et al., 2024; Law et al., 2025; Kaya et al., 2025). The contribution of the present case is its scale, its inclusion of four broad knowledge areas and ten subjects, its use of Mexican Spanish, and its focus on an institutional workflow rather than isolated zero-shot prompts.

The area differences are also instructive. The lowest Gemini VFI occurred in Humanities and Arts, despite the interpretive nature of those subjects. Conversely, PMSE was the only area in which Gemini drafts received more comments per item. The qualitative evidence suggests that this difference combined model-related issues (option asymmetry and lexical clues) with workflow-related notation and transfer errors. In BCHS, faculty contextual sensitivity and AI

consistency were particularly complementary. These contrasts argue against a single claim that either AI or faculty authors are universally better. The appropriate question is which tasks should be standardized, automated, or reserved for expert judgment in each domain.

A defensible human-in-the-loop workflow would include at least five layers. The first is specification: the construct, blueprint cell, learning outcome, cognitive demand, item format, and intended use must be explicit. The second is controlled generation using a versioned prompt, curated source documents, and a standardized output schema. The third is automated linting for prohibited wording, option-length imbalance, key-position rules, lexical overlap, duplicate content, metadata completeness, and notation integrity. The fourth is expert review for substantive correctness, curricular meaning, bias, contextual authenticity, and distractor quality. The fifth is empirical validation after pilot or operational administration, including difficulty, discrimination, point-biserial correlations, distractor functioning, reliability, and fairness analyses. This sequence locates GenAI within an evidence-building system rather than treating generation itself as evidence of validity (American Educational Research Association et al., 2014; Kaldaras et al., 2024).

The study also suggests how VFI could mature into a more useful process measure. Future versions should separate positive confirmations from required corrections, distinguish content, cognitive-level, style, formatting, and platform errors, and record the time spent at each review stage. Item-level data would permit analyses that adjust for subject, cognitive level, format, and reviewer committee. Blinded source evaluation and matched or randomized blueprint cells would reduce expectation and selection effects. Inter-rater agreement should be estimated, and the relationship between pre-operational review indicators and post-administration item statistics should be examined.

Finally, reproducibility requires governance of the AI system itself. Institutions should retain the full system instructions, prompt revision history, knowledge-file versions, model label, generation date, original output, human edits, reviewer decisions, and platform transformations. Such provenance is necessary for auditing unexpected items and for determining whether a problem originated in the blueprint, the model, the prompt, the human editor, or the technical

interface. The project's centralized Gem architecture offers a practical starting point for this form of assessment governance.

7 Conclusion

Customized Gemini Gems produced structured MCQ drafts that, in this naturalistic institutional comparison, generated fewer recorded review comments than faculty drafts in three of four knowledge areas. Their main advantages were consistency and blueprint adherence; their recurring weaknesses involved distractor quality, verbosity, option asymmetry, and unintended clues. Faculty authors supplied disciplinary context and knowledge of authentic misconceptions, but their drafts showed greater variability in alignment and item-writing form.

The findings support neither full automation nor fully manual production. A more defensible division of labor uses AI for constrained structural drafting, automated tools for formal checks and platform transfer, and faculty committees for substantive validation and adjudication. VFI may help monitor this workflow, but only after comment types and reviewer effects are distinguished, and it cannot replace evidence from examinee responses. The next phase will test whether the observed pre-operational patterns predict difficulty, discrimination, distractor functioning, and fairness in actual administration.

Limitations

This study has several limitations. It was a retrospective process evaluation rather than a randomized comparison. AI- and faculty-authored items were not matched within blueprint cells, and source counts differed substantially in Social Sciences and Humanities and Arts. Differences in subjects, item formats, or cognitive levels may therefore partly explain the observed VFI patterns.

VFI is a novel administrative ratio with no established validity evidence. The comment count combined positive confirmations and requests for correction, may reflect reviewer communication style, and does not measure the severity of a problem. The available records did not support item-level modeling, formal qualitative reliability, inter-rater agreement, or confirmation that reviewers were blinded to item source. Actual authoring time, review time, rejection rates, and costs were not available.

No examinee response data were analyzed, so the study cannot establish equivalence in

difficulty, discrimination, reliability, fairness, or score interpretations. The foundation-model version used by the Gemini application was not preserved in the available project documentation, limiting technical reproducibility. Finally, findings from one institution, one language, one prompt architecture, and one item-management workflow may not generalize to other settings or later model versions.

Ethical Considerations

The analysis reports aggregate institutional quality-assurance data and includes no student or examinee records. No operational item content, answer keys, or confidential source documents are reproduced. The prompt scaffold is abridged and translated to illustrate the workflow without compromising test security. Human committees retained final authority over item acceptance and revision. Future operational use should also include systematic review for cultural, linguistic, and construct-related bias and should preserve auditable records of AI output and human decisions.

Acknowledgments

The authors thank the faculty members who served on the subject-specific validation committees and the staff of CEIDE's educational evaluation and systems teams who supported the item-bank workflow.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
- Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: Systematic review. *BMC Medical Education*, 24, Article 354. <https://doi.org/10.1186/s12909-024-05239-y>
- Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, Article 858273. <https://doi.org/10.3389/feduc.2023.858273>
- Gierl, M. J., Shin, J., Firoozi, T., & Lai, H. (2022). Using content coding and automatic item generation to improve test security. *Frontiers in Education*, 7, Article 853578. <https://doi.org/10.3389/feduc.2022.853578>

- Google. (n.d.). *Tips for creating custom Gems*. Gemini Apps Help. Retrieved September 1, 2026, from <https://support.google.com/gemini/answer/15235603>
- Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *SN Computer Science*, 5, Article 636. <https://doi.org/10.1007/s42979-024-02963-6>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. <https://doi.org/10.4324/9780203850381>
- Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, 9, Article 1399377. <https://doi.org/10.3389/feduc.2024.1399377>
- Kaya, M., Sonmez, E., Halici, A., Yildirim, H., & Coskun, A. (2025). Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: A psychometric analysis. *BMC Medical Education*, 25, Article 949. <https://doi.org/10.1186/s12909-025-07528-6>
- Law, A. K. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Hung, K. K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: A cohort study in a high-stakes examination. *BMC Medical Education*, 25, Article 208. <https://doi.org/10.1186/s12909-025-06796-6>
- Linde, P., Fichter, F., Dietlein, M., Sudbrock, F., Afshar, K., Dapper, H., Fokas, E., Hillebrecht, A.-L., Raupach, T., & Laupichler, M. C. (2026). Psychometric properties and detectability of GPT-4o-generated multiple-choice questions compared with human-authored items across imaging specialties. *npj Digital Medicine*, 9, Article 132. <https://doi.org/10.1038/s41746-025-02313-7>
- Suárez-Álvarez, J., He, Q., Guenole, N., & D'Urso, D. (2026). Using artificial intelligence in test construction: A practical guide. *Psicothema*, 38(1), 1-12. <https://doi.org/10.70478/psicothema.2026.38.01>