Item Quality Consideration in Automatic Item Generation

Alexander Hoffman, AleDev Research & Consulting

David Whitcomb, ATLAS, University of Kansas

Marjorie Wine, ATLAS, University of Kansas

Abstract

This study investigates the viability of the new *Low-Bar Item Draft Acceptability Tool* (LBIDAT) to assess the quality of early-stage item drafts produced by AIG/AQG. These tools address overlooked cognitive and content traits, offering a cross-content framework focused on items' potential to elicit evidence of targeted cognition across a test taking population.

1 Introduction

The advancement of AIG/AQG (automatic item & question generation) for use in large scale assessment requires greater attention to item quality than it typically receives. Such research must account for the various types of quality reviews in professional item development processes and clarify which dimensions of quality the project aims to address—while recognizing which dimensions are addressed elsewhere in the item development pipeline. Unfortunately, AIG/AQG research efforts tend to focus only on the most superficial aspects of items quality (i.e., akin to proofreading and formatting), rather than items' suitability for their intended purposes. With rare exception, they apply lower quality standards more appropriate for classroom assessments (with its more numerous assessments and consequential potential for triangulation) than for large scale assessments (Hoffman et al., 2024).

Moreover, disruption theory (Christensen, 1997; Christensen et al., 2015) makes clear that the success of disruptive technological innovations depends on altering a value proposition, by trading off excessive quality or features for dramatically lower costs. Understanding the potential for AIG requires recognizing the quality produced by the new technology so that it can positioned in the old or in new markets.

The Low-Bar Item Draft Acceptability Tool (LBIDAT; Hoffman & Wine, 2025) was designed to assess item quality in the earlier stages of item development. It is a multi-dimensional, non-compensatory rubric built around the idea of *critical issues* (p. 3). Critical issues, as opposed to significant issues or hygiene issues, are those i) whose fixing often creates cascades of additional issues to address, ii) are not guaranteed to be fixable and therefore iii) have the potential to render items unusable or unsalvageable.

The LBIDAT addresses quality considerations precisely where AIG/AQG is most likely to be used—either in place of item writers or in lieu of both item writers and early content development professionals' (CDPs') work (Song et al., 2025; Tan et al., 2025). Therefore, it focuses on early- and mid-process standards for item quality, rather than markers of late-stage refinement. It offers a standardized, cross-content framework for judging the quality of items produced by AIG/AQG projects through the basic lens of how much work it would take a CDP to refine the item to the point that it could be used to produce viable evidence of the targeted cognition on an assessment.

2 Basic Item Development Workflow

Item development is a multi-stage process that conventionally requires 18-30 months from stimulus or item writing assignments through selection of items for inclusion on operational test forms or adaptive item pools (Wine & Hoffman, forthcoming), as shown in Figure 1. The second major step is item writing, in which content experts and/or teachers draft item starters. These item writers are *not* usually assessment professionals, and instead are doing additional work somewhat adjacent to their expertise. Then, CDPs and their expert colleagues review and refine successive item drafts before getting feedback on them from

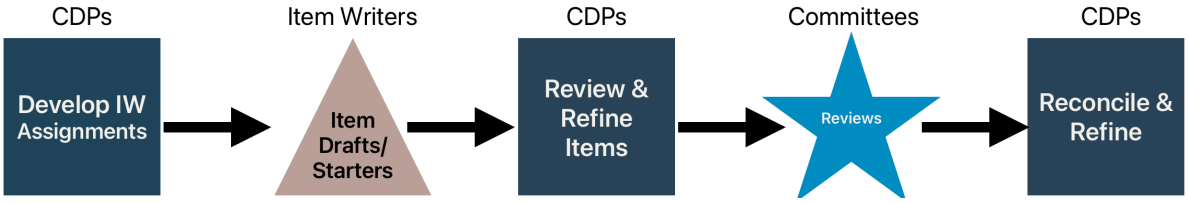

Figure 1: A grossly simplified content development process

expert committee reviews. Before field testing (and associated production work), CDPs do their final reconciliation and item refinement work. Items are then field tested and finally considered for operational use. At any point in this workflow, items may be dropped for being unworkable or unsalvageable.

3 Initial Intake/ Human-in-the-Loop

Initial intake is the critical first step in which *CDPs* examine item starters drafted by item writers and therefore the beginning of the *review and refinement* process. If AIG has replaced item writers, this is the first step in which humans examine the automatically generated items for potential viability in human-in-the-loop workflows.

Expectations for item quality at initial intake are often rather low. The practical questions are i) whether the item starters have the potential to be useful valid items (i.e., item that elicit evidence of the targeted cognition for the range of typical test takers) and ii) how much work it may take to refine them to the point that they can be used.

Pure item intake does not include any editing or refinement work. Instead, it marks a border and an initial inspection point. Some number of critical issues are expected in nearly every item starter, though their number, nature and dispersion in the item vary. (Some organizations fold item intake into a larger process which *does* include some amount of item review and refinement.)

Pure initial intake is a very quick process, with a single CDP examining and considering each item for approximately five minutes. Even when included in a longer first contact process, initial inspection of items is incredibly important.

4 Understanding the LBIDAT

The LBIDAT structures initial intake inspection into five dimensions: stem & task, key, distractors, stimulus, and fairness. Each dimension is scored on a three point ordinal scale [√, ~, !!]. It uses these marks to discourage mistaking it for a compensatory rubric. The number of critical errors that qualify an item for a value varies from dimension to dimension.

A single !! value should raise the question of whether an item is even salvageable, as should many ~ values. That is, the CDP should seriously entertain the question without assuming that the item either can or cannot be salvaged.

4.1 Stem & Task

This is the most important dimension, including both the cognitive task that the item intends test takers to complete and the expression of the stem itself. Critical issues for this dimension include task alignment with its designated alignment reference (e.g., learning standard), (grade) level appropriateness, requirement of additional KSAs and availability of alternative paths to a successful response. This and the other dimension are fully explained in the original item acceptability tools white paper (Hoffman & Wine, 2025).

4.2 Key

This is the simplest dimension. The key i) must be directly responsive to the stem and ii) must be definitively correct. Even one shortcoming here pushed this dimension to the !! category. However, because mismarking the key is easily addressed, it does *not* count as a critical issue and is therefore not a problem through the lens of the LBIDAT.

4.3 Distractors

In the dominant multiple choice item format, distractors are as important to alignment as stems. Therefore, distractors also have many types of potential critical issues. They include failing to be directly responsive to the stem, resulting from errors *not* in the targeted cognition, redundancy and not being definitively incorrect. (We have observed that failing to account for distractor quality is the most obvious flaw in AIG research. This mistakes automatic *item* generation for automatic *question*

Table 1. Selected CCSS Standards

CCSS Standard	Standard Text
Math 8.F.A.2	Compare properties of two functions each represented in a different way (algebraically, graphically, numerically in tables, or by verbal descriptions). For example, given a linear function represented by a table of values and a linear function represented by an algebraic expression, determine which function has the greater rate of change. Answer options must relate the two functions to each other
Math 8.F.B.4	Construct a function to model a linear relationship between two quantities. Determine the rate of change and initial value of the function from a description of a relationship or from two (x, y) values, including reading these from a table or from a graph. Interpret the rate of change and initial value of a linear function in terms of the situation it models, and in terms of its graph or a table of values
Math 8.G.C.9	Know the formulas for the volumes of cones, cylinders, and spheres and use them to solve real-world and mathematical problems
ELA RI 8.2	Determine a central idea of a text and analyze its development over the course of the text, including its relationship to supporting ideas; provide an objective summary of the text
ELA RL 8.1	Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text
ELA RL 8.4	Determine the meaning of words and phrases as they are used in a text, including figurative and connotative meanings; analyze the impact of specific word choices on meaning and tone, including analogies or allusions to other texts

generation, and therefore ignores many critical aspects of items quality.)

4.4 Stimulus

Stimulus critical issues focus on time, construction mistakes, plausibility, and necessity for successful responses.

4.5 Fairness

The LBIDAT does not consider bias issues, as initial intake inspection does not have time to include these issues. Instead, it focuses just on sensitivity issues. Again, the threshold for a !! is a single critical issue.

5 Methodology

For this viability study, we attempted to use the LBIDAT to evaluate math and ELA item starters generated by a range of large language models.

5.1 Item Generation/Data

Item starters/drafts were generated by ChatGPT 5 (auto), ChatGPT 4o, Google Gemini Flash 2.5, Google and Gemini Pro 2.5.

Three representative 8th grade standards were selected from each of the mathematics and English language arts Common Core State Standards, as shown in Table 1. Reading passages for ELA item generation were drawn from practice tests and publicly released reading passages from a diverse range of six states, as shown in Table 2. (A second passage was drawn from New Jersey because the first passage was too low quality to support items aligned to the selected standard. Those items were kept, as it is not uncommon to receive such items from item writers.)

We used an *explicative prompting strategy* that combined instructional prompting (Saleem et al., 2025; Yang et al, 2025) and constraint prompting (Maxwell, 2025) in a new *tiered prompting* approach that began with mere reference to the standard number and culminated in a prompt with 100+ words explaining what was expected from the standard and item and 75+ words explaining what would *not* be appropriate. Each of these six tiers of prompts was presented both with and without the LBIDAT to each LLM for each standard.

This resulted in 336 ELA item starters and 144 mathematics item starters, plus additional items generated with less complete prompts (e.g., from failing to ask that the key be indicated or that math items include rationales).

Table 2. Reading Passages for Item Generation

State	Passage	Standard
NJ	Emerald Ash Borer (adaptation)	RI 8.2
NJ	Elephants Can Lend a Helping Trunk	RI 8.2
UT	James "Jim" Bridger	RI 8.2
MA	The Night Circus (excerpt)	RL 8.1
TX	Mr. Linden's Library (excerpt)	RL 8.1
NM	Dr. Blackwell's New Assistant	RL 8.4
TN	O Pioneers (excerpt)	RL 8.4

5.2 Item Inspection/LBIDAT Use

Items were evaluated with the LBIDAT by at least one content development expert with expertise in the content area, yielding a five dimensional score for each item (e.g., "~ √ ~ √ !!", "!! √ !! √ √").

Item inspectors all reviewed the LBIDAT instructional packet (Hoffman & Wine, 2025) before reviewing items. Initially, items were reviewed synchronously by two-person panels who discussed each starter, *without* a requirement for coming to agreement. This was not a CAT (Consensual Assessment Technique; Amabile, 1982) approach, because final scores of from *each* inspector were recorded. These panels reviewed complete sets of 12 items from an LLM for a standard to ensure a range of item quality would be included. These panels reviewed >1/4 of the items in each content area.

Later reviews were done by individually by the same inspectors, usually in complete 12 items. However, a sample from each set was reviewed by at least one inspector. Collectively ~2/3 of the items were reviewed with the LBIDAT.

We originally intended also to stress test the CBIDAT (*Creativity-Based Item Draft Acceptability Tool*), but the LLMs did not produce novel enough approaches in any item in the pool to merit its application. This was not entirely surprising, given how LLMs base their responses on their training corpuses. Appropriate creativity might result from altered temperatures, but that is beyond the scope of this project.

5.3 Analysis/Compilation of Findings

Item inspectors occasionally broke from the item-focused discussions during their panel meetings to address observations about the LBIDAT tool and its usage. They added additional notes on the LBIDAT and its usage through individual item inspections.

Item inspectors repeatedly met together across content areas to review and discuss usage of the LBIDAT, problems with the LBIDAT and alternations to the LBIDAT that would make it either more useful or easier to use.

This analysis work followed the *RTD Internal Methodology* (Hoffman & Wine, 2017) because it is based upon using the judgments of experts as data. In this case, those judgments were content expert who are also item development experts. The RTD *Internal Methodology* is especially important because the analysis is aimed at developing artifacts of organizational learning—such as a tool to standardize high quality reviews of item drafts.

6 Findings

This viability/validation study has five principal findings, one of which was entirely unsurprising and one of which was *quite* surprising.

6.1 Basic Viability

The LBIDAT proved to be a viable item intake protocol.

- The LBIDAT can be used for quick (i.e., 5-minute) item inspections *and/or* longer and more substantive initial item engagement which include some degree of item refinement work.
- The LBIDAT works both in the context of a traditional item-writer→CDP workflow and a human-in-the-loop AIG system→CDP workflow.
- Though the LBIDAT has a learning curve associated with it, its structure is easily understood and the judgments it requires are no different than those required by best practice in its absence.
- Though use of the LBIDAT is hampered by a lack of answer option rationales—especially with math items involving calculations—this is no more the case than in item evaluation without the LBIDAT.
- The LBIDAT collects information that can be very useful in later review and refinement steps of item development. For example, its results can guide work assignments to more junior vs. more senior CDPs. It can identify the most promising item drafts that are the best targets for further investment.

The LBIDAT is a useful tool for a variety of purposes and workflows—including evaluation of AIG outputs.

6.2 Dependence Upon Expert Judgment

Though the LBIDAT provides a structure for item draft evaluation, it does not eliminate the need for expertise to make appropriate judgments about item draft quality.

Judgements about item draft acceptability still depend on expert knowledge of the content area, of the specific alignment reference (e.g., a single state learning standard), and the relationships between (and distinctions among) different alignment

references—both vertically and horizontally—that collectively make up the domain model. Use of the LBIDAT still requires that expertise.

Judgments about item draft acceptability still depend on expert understanding of how items—especially multiple choice items—work to prompt cognition and collect evidence thereof. The structure of the LBIDAT provides scaffolding to help item inspectors to evaluate parts of an item, but use of the LBIDAT in item intake requires a facility with this understanding in order to meet time constraints. This is somewhat eased by the supportive scaffolding of the LBIDAT, but it is not erased.

Judgments about item draft acceptability still depend on experience with fairness review committees. Item inspectors must be able to anticipate the kinds of sensitivity issues these committees notice, how they will view items and the likelihood of influencing their views on a particular potential sensitivity topic. That is, the political context of psychological empathy and sensitivity cannot be simulated with quantitative tools (e.g., DIF) or those less-than-deeply-familiar with the dynamics of this kind of committee work.

6.3 Need for LBIDAT Refinement

The LBIDAT would nonetheless benefit from a handful of alterations.

First, the LBIDAT is too generous with items whose Stem & Task are simply ill-suited to the alignment reference. Otherwise well-constructed items that simply are not designed to elicit evidence of the *designated & targeted* cognition are not necessarily flagged appropriately. This is because the three critical issues !!-threshold for the Stem & Task dimension is too lenient. Stem & Task are so important that the !!-threshold should be lowered to two critical issues.

Other problems with the LBIDAT are more in the guidance/training documentation than the final tool, itself. For example, the problem of entirely missing stimulus elements is covered by the dimensional guidance question regarding stimulus elements that do not match their description elsewhere, but this could be explained more clearly. Similarly, items based on a misunderstanding of the targeted cognition—as is so often the case with traditional item logics that have not been updated to contemporary domain models—should be flagged for various Stem & Task dimensional guiding question violations, but that basic problem should be highlighted in the guidance documentation.

The issue of mismarked keys was oddly common, and we had to resolve whether this is a critical issue. The guidance documentation should explain *why* it is not a critical issue—though clearly it needs to be addressed before an item can be used.

Though it is a little beyond the LBIDAT itself, it is important that test commissioning organizations (i.e., clients) contribute to discussions about interpreting alignment references and identifying their cores. RTD workflows should record these results in task models long before initial intake—an such information is likely vital to high quality AIG outputs—but not all item development or AIG workflows are grounded in Rigorous Test Development practice. Therefore, the LBIDAT guidance documentation should call out this issue.

6.4 Acceptability, But Not Direction for Item Improvement

Perhaps the least surprising finding was that the LBIDAT does not identify how to fix an item or even where the true sources of issues lie. This is not surprising because it was never designed to do so. This is not surprising because we have long understood that recognizing the true source of an issue often requires considerable expertise with the mechanics of items and how they shape test takers' cognitive paths.

The LBIDAT is designed to support quick determinations of initial item draft quality, not to identify the particular strengths and weaknesses of an item. Tracing apparent problems to their source is much more complex goal.

Nonetheless, it is easy to anticipate that LBIDAT users will want this tool to deliver more than it can. A quick inspection of an item cannot reveal all the connections between the parts of an item and their influence on test taker cognition. While expert CDPs are often able to recognize how apparent issue here is actually caused by a problem over there, this depends on experience with refining hundreds or thousands of items in the past. The LBIDAT is intended to support novice and junior CDPs, in addition to senior CDPs, so that they can deliver item scores or profiles in the same form as their senior colleagues.

This is not a weakness of the LBIDAT any more than the fact that our automobiles are useless for interstellar travel. But the fact is that many without item (or content) expertise are looking for tools to

automate those judgments. The LBIDAT does not match their aspiration.

6.5 LLM Self-Evaluation

This phase of our larger AIG efforts was *not* intended to evaluate the LLMs themselves—not their outputs and not their capabilities. Nonetheless, they forced our hand(s).

When prompts included, "You should write an item that scores well on the LBIDAT, which is detailed below," and were given the full text of the LBIDAT guidance and the tool itself (in markdown format), the LLMs often included their own evaluations of their item on the LBIDAT in their outputs.

These evaluations were universally inflated, only once acknowledging even in a single critical issue in any item. Items quite rarely met the √-√-√-√-√ standard, quite often meriting multiple !!'s. Nonetheless, the LLMs 'thought' quite highly of their own work.

We are surprised that we are able to determine that inclusion of the LBIDAT and its guidance documentation are *not* sufficient for LLMs to use the LBIDAT to evaluate items. We are *not* surprised that LLMs cannot make good use of the LBIDAT, but we did not expect this study to offer sufficient evidence to conclude that.

However, this conclusion is based on zero-shot with rubric (Tian et al. 2024) prompting. It is yet undetermined whether LLMs can be trained or fine-tuned to use the LBIDAT at all accurately. The seemingly obvious fact that the vast vast majority of items in their original training corpuses are simply very low quality presents a considerable obstacle and even fine-tuning (e.g., LoRA, QLoRA) may not be sufficient.

7 Discussion

The LBIDAT is truly a new tool, offering a structure for consistent reports of item draft quality at a key early stage within item development workflows. This is valuable, even without consideration of its application in AIG contexts.

However, within the AIG research context, the LBIDAT is uniquely invaluable. There simply is no other tool available to evaluate the quality of AIG output that considers their potential to contribute meaningfully to assessments whose results are interpretable as indicators of test takers' proficiencies with the KSAs that make up the alignment references in a domain model.

AIG research efforts that do not consider item validity simply are not positioned to inform anyone whether currently available technologies can deliver on the promises made by their proponents. Not only do they not know, they are not even truly asking that most singular important question: *do these generated items have the ability to elicit evidence of the targeted cognition for the range of typical test takers?*

We are sad to say that without the LBIDAT—or some equivalent—AIG research merely advances our ability to generate items with a surface resemblance to high quality items, and therefore items that can be produced at a lower expense. However, those items will fail to provide instructionally useful information to instructors or learners, accurate classifications of test takers to those who depend upon them or bases for comparing test takers. They will never be able to help anyone evaluate learners' needs, aspirants' abilities, instruction, curriculum, educational policy or any other relevant influences. Instead, they will offer the illusion of doing so to those who trust experts—who in such cases have oversold their items or systems.

We do not believe for a second the LBIDAT is irreplaceable. There doubtless are other approaches to evaluating item drafts' potential and issues than the LBIDAT offers. We *do,* however, believe that something akin to the LBIDAT is required for any serious AIG research. If a project has no mechanism to determine how prone an item is to falsely suggesting proficiency or falsely suggesting a lack of proficiency, how can a new method's success be evaluated? Without some determination of the disruptive technological innovation's new value proposition, how can appropriate uses be determined?

Unfortunately, there is not yet any evidence that contemporary AI engines are capable of automating item quality (or even item draft acceptability) judgments. Surely, this would speed up item development and enable improved AIG research and commercial—even educational—use. Neither the hope that they could do such a thing, nor the value of their doing such a thing has any bearing on whether they actually can. Epistemic humility requires that everyone working in the field acknowledge the fact that the evidence does not exist. Disciplinary humility should require everyone who is not an expert in *item* quality to defer those who are.

It is *not* our goal in the project to automate item quality judgments, for a variety of reasons—even though we understand the potential value of doing so. We see far more realizable potential value using AIG to replace item writers to generate item drafts. The LBIDAT is an appropriate tool for that use.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5), 997. https://doi.org/10.1037/0022-3514.43.5.997

Christensen, C. (1997). The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business Review Press.

Christensen, C. M., Raynor, M., & McDonald, R. (2015). What is disruptive innovation. Harvard Business Review, 93(12), 44-53. https://hbr.org/2015/12/what-is-disruptive-innovation

Hoffman, A., Glore, C., Harrison, D. & Wine, M. (2024). *Recognizing the Strengths of Classroom Assessment: What a Principled Large Scale Content Development Practice Has to Offer to Classroom Assessment*. 2024 NCME Special Conference on Classroom Assessment, Chicago, Illinois. https://www.researchgate.net/publication/384065928_Recognizing_the_Strengths_of_Classroom_Assessment_What_a_Principled_Large_Scale_Content_Development_Practice_Has_to_Offer_to_Classroom_Assessment

Hoffman, A. & Wine, M. (2017). The Rigorous Test Development Project: Internal Methodology [White paper]. Rigorous Test Development Project. http://dx.doi.org/10.13140/RG.2.2.13125.46563

Hoffman, A. & Wine, M. (2025). *The Low Bar Item Draft Acceptability Tool & Creativity-Based Item Draft Acceptability Tool* [White paper]. The Rigorous Test Development Project. https://osf.io/preprints/edarxiv/xuvyn_v1

Maxwell, I. A. (2025). *Meta-Cognitive Prompting: A Comparative Framework for Prompt Engineering in Large Language Models*. https://www.researchgate.net/profile/Ian-Maxwell-2/publication/392558190_Meta-Cognitive_Prompting_A_Comparative_Framework_for_Prompt_Engineering_in_Large_Language_Models/links/6858c22993040b17338ca00b/Meta-Cognitive-Prompting-A-Comparative-Framework-for-Prompt-Engineering-in-Large-Language-Models.pdf

Saleem, S., Asim, M. N., Zulfiqar, S., & Dengel, A. (2025). The Evolution of Natural Language Processing: How Prompt Optimization and Language Models are Shaping the Future. arXiv preprint arXiv:2506.17700.

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. Interactive Learning Environments, 1-20. https://www.researchgate.net/publication/390147704_Automatic_item_generation_for_educational_assessments_a_systematic_literature_review

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. International Journal of Assessment Tools in Education, 12(2), 317-340. https://doi.org/10.21449/ijate.1602294

Tian, X., Mannekote, A., Solomon, C. E., Song, Y., Wise, C. F., McKlin, T., Barrett, J., Boyer, K. E., & Israel, M. (2024). Examining LLM prompting strategies for automatic evaluation of learner-created computational artifacts. In *Proceedings of the 17th International Conference on Educational Data Mining* (EDM 2024) (pp. 698–706). International Educational Data Mining Society. https://learndialogue.org/pubs/LearnDialogue_Tian_EDM2024.pdf

Wine, M. & Hoffman, A. (forthcoming). *The test development cycle.*

Yang, H., Zhao, Y., Min, S., Su, B., Yao, C., & Xu, W. (2025). *Instructional Prompt Optimization for Few-Shot LLM-Based Recommendations on Cold-Start Users*. arXiv preprint arXiv:2509.09066

A Appendix

The LBIDAT form is reproduced on the next page Its full explanation and guidance can be found in Hoffman and Wine (20225).

Low Bar Item Draft Acceptability Tool (LBIDAT) Form

Stem & Task

No Apparent *Critical Issues*	1-2 Apparent *Critical Issues*	3+ Apparent *Critical Issues*
✓	~	!!

- The prompted task is based upon a content mistake or misunderstanding that undermines the workability of whole item.
- The targeted cognition is not at a (grade) level appropriate version of the cognition.
- The targeted cognition is not part of the most important core of the alignment reference or standard.
- Alternative Paths: The prompted task does not depend upon the targeted cognition for a significant number of typical test takers.
- Additional KSA: The prompted task also requires some other KSAs outside of the alignment reference that are at the item's (grade) level, above the item's (grade) level, or just one (grade) level below the item.
- The task requires notable learning for a successful response. (May be acceptable for inquiry-based tasks aligned to inquiry-based alignment references.)

The Key

No Apparent *Critical Issues*	1+ Apparent *Critical Issues*
✓	!!

- The key is not directly responsive to the question or command in the stem.
- The key is not *definitively* correct.

Distractors

No Apparent *Critical Issues*	1-2 Apparent *Critical Issues*	3+ Apparent *Critical Issues*
✓	~	!!

- Each distractor that does not appear to directly respond to the question or command in the stem.
- Each distractor that is not the product of an error, misapplication or misconception with the targeted cognition (i.e., is plausible)
- Each distractor that is not *definitively* incorrect.
- Multiple distractors follow from the same error, misapplication or misconception as another distractor.
- Each distractor that is a duplicate of another distractor.

Stimulus

No Apparent *Critical Issues*	1-2 Apparent *Critical Issues*	3+ Apparent *Critical Issues*
✓	~	!!

- The stimulus requires too much time for test takers to make sense of. (Primarily for stand-alone items.)
- The stimulus contains inappropriately implausible or incorrect elements.
- The item does not require the stimulus for a successful response for a significant number of typical test takers.
- Each element of the stimulus that does not match its description or assumptions elsewhere
- Each construction mistake in structured stimuli.

Fairness

No Apparent *Critical Issues*	1+ Apparent *Critical Issues*
✓	!!

- Each inappropriate sensitivity topic raised by the item.