RIGOROUS TEST DEVELOPMENT

Low Bar Item Draft Acceptability Tool (LBIDAT) & Creativity-Based Item Draft Acceptability Tool (CBIDAT)

By: Alexander Hoffman
Marjorie Wine

The LBIDAT and CBIDAT are designed to evaluate the quality of unrefined drafts of items for large scale assessment development, such as might be submitted by item writers and/or automatic item generation. The LBIDAT does not determine whether an item draft is ready for use, but rather the existence and location of *critical issues* they may render an item unusable. The CBIDAT focuses entirely on the item logic of the draft to consider whether it might be a good new basis for future items. Even high scoring item drafts will still require significant additional professional refinement and polish before use.

ALEDEV CONSULTING

The RTD (Rigorous Test Development) project is an attempt to build a professionalized content development practice that focuses on individual item quality, particularly by leaning into the importance of validity throughout the content development process. It assumes that content development professionals develop professional judgment that can be raised, honed and calibrated by providing frameworks and clarifying expectations in ways that account for the constraints and demands of typical practice within test development, today. RTD is a conscious and deliberate attempt to respond to the disparity in status, training and shared knowledgebases between psychometrically oriented professionals and content development professionals.

Table of Contents

Low Bar Item Draft Acceptability Tool (LBIDAT)

Different standards or expectations of quality apply at different points in the item development process for large scale assessment. This is not simply to say that there are lower standards earlier and higher standards later, there are *different* standards that focus on different issues. The LBIDAT (*Low Bar Item Draft Acceptability Tool*) and CBIDAT (*Creativity-Based Item Draft Acceptability Tool*) focus on *Critical Issues* that are most important to address early in an item's development. There are layers of other *Significant Issues* that must be addressed as well, but they can be addressed later.

Critical Issues differ from *Significant Issues* for three main reasons. First, altering an item to address a *Critical Issue* is far likelier to create additional issues than trying to address a significant issues is. That is, the work of fixing a critical issue often turns into a cascade of issues that each need to be fixed. Second, though important, *Significant Issues* can generally be assumed to be fixable, whereas a critical issue might not be fixable. Whether a critical even can be fixed—which necessarily encompasses fixing any cascade that results—is often uncertain until it is done. Therefore, *Critical Issues* should be addressed first, if possible.

Most importantly, *Critical Issues* may prevent an item from being refined to be usable on an assessment because they fairly fundamentally can prevent the responses from test takers from providing clear evidence of whether a test taker does or does not have proficiency with the targeted alignment reference (e.g., a state learning standard) and its associated KSAs (knowledge, skills and/abilities). The LBIDAT focuses exclusively on these issues. The CBIDAT goes even deeper, evaluating item drafts of stems *not* for their implementation, but rather for whether they present a valuable good *new* idea for how to assess some targeted cognition.

Both of these tools rely on the expert judgment of CDPs (content development professionals)—their applied expertise in how items work, how to assess specific content, and how test takers engage in cognitive work to produce a response. While highly experienced and capable CDPs are likely to come to the same results when using either tool to evaluate an item, less experienced CDPs may come to different results. Therefore, the LBIDAT is best used in the hands of at least moderately experienced CDPs, and the CBIDAT should be left to highly experienced CDPs. (However, they also provide scaffolding for early item analysis that can be quite useful to those teaching or learning how to be a highly capable content development professional.)

Each of these tools is multi-dimensional, made up of non-compensatory, interval scales. The are *not* about counting, and certainly not about summing up to quantitative scores. Rather, they each recognize and report on the multiple criteria that relatively early item work addresses.

The Challenges of Large Scale Content Development & the Need for Quality

Understanding content and item development for large scale assessment (i.e., its practices, processes and goals) requires acknowledging the great twin challenges of large scale assessment—challenges that make it so different from classroom assessment.

Low Bar Item Draft Acceptability Tool (LBIDAT)

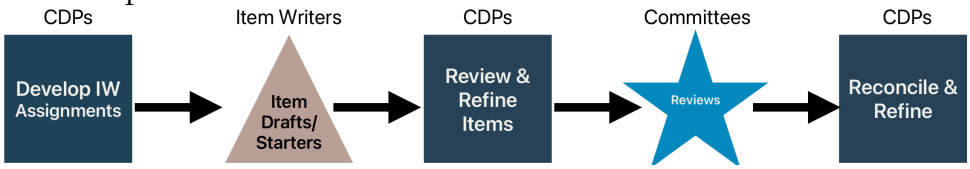
First, large scale assessment has dramatically fewer formal opportunities to assess test takers (e.g., students) than classroom teachers do. If one acknowledges the reality that classroom assessment also includes *informal* assessment, classroom assessment has exponentially more opportunities to collect data. With so much less data, it is all the more important that large scale assessments have more signal and less noise. Moreover, the volume and variety of classroom assessments allow for triangulation to make relatively precise inferences about rather small constructs even when multiple signals come in at the same time. Thus, individual items and complete instruments for large scale assessment *must* be higher quality and better focused than what is used in classroom assessment—dramatically or even exponentially!

Second, there is an enormous distance between the developers of large scale assessment and test takers—virtually by definition. They do not know each other, are not personally accustomed to understanding each other and actually know quite little about each other. Teachers can rely on months of training their students to know what they mean and vice versa. On the other hand, test developers try to make their intentions clear to strangers and then attempt to make inferences about these strangers' cognition, of course based on very little evidence. Moreover, the range of typical tests takers for large scale assessment is far broader than found in a single classroom, school or even district. Thus, it simply is harder to create items of equal quality to that of a typical classroom assessment for the *range of test takers* of a large scale assessment.

Therefore, *item validity*—items' ability to elicit evidence of the targeted cognition for the range of typical test takers—is more important with large scale assessment and more difficult to achieve than in classroom assessment.

The Place of Quality and Acceptability in Item Development

A minimal—though of course grossly simplified—view of the item development process has five steps[1].



First, CDPs develop item writing assignments, based on what is needed for the eventual test. Second, item writers write initial item drafts (or "item starters"). Third, CDPs take those starters an engage in the iterative process of item review and refinement to develop them into items that hopefully are of high enough quality to be used on a test. Fourth, external expert review committees—primarily teachers and/or practitioners in the field—review the items and give feedback on them. Fifth, CDPs examine and consider that feedback and edit the items to improve them further, as needed. All of those steps precede any field testing or psychometric evaluation of items and final decisions about whether to

[1] In fact, the item development process has nearly 100 steps, and far more contributors than shown in this view. See Wine & Hoffman's (2025) *The Item Development Cycle* for a more complete explanation.

include an item in an operational pool or on an actual test form. When CDPs work for a test development vendor, a representative for the client (e.g., a state's department of education) might step in at a number of different points to observe, review, offer feedback or even edit items themselves. Automatic item generation (AIG) efforts—such as with AI or LLM engines—often look to replace that second step (i.e., item writers). However, the quality review and consequent item refinement work remains necessary. AI proponents often call this *human-in-the-loop*.

The complete item development process (i.e., most of which is omitted above) takes well over a year to complete. This is because the quality of those initial item drafts or starters are so far from what is needed for operational large scale assessments. The work of developing suitable items is largely done in the middle step, by CDPs. Their goal is to present items to review committees so free of problems, that little or no work is needed in the last step. Ideally, item writers' drafts would already be that good, but that is never the case. It is far more likely that they require wholesale reworking than that they do not require any work by CDPs, if they are appropriate at all. Most items drafts need serious work by CDPs, even if they do not require wholesale reworking[2].

Highly refined items not only *elicit evidence of the targeted cognition for the range of typical test takers*, but do so very efficiently (i.e., without suggesting Type I or Type II errors). They are polished to the point that they both look professionally developed and do not include little problems that might distract test takers from their work. Therefore, the full process includes various forms of editing reviews, from compliance with the style guide to proofreading and fact checking, in addition to the more substantive issues of content alignments and fairness (e.g., bias and/or sensitivity). CDPs aspire to catch all of those sorts of issues, and this is so important that there are many reviews by experts contributors who act as safety nets in various ways.

Hence, item quality starts very low. Through the early work of CDPs in the third step, *Critical Issues* are addressed and an item becomes more acceptable, with generous allowances given for a lack of various types of polish. *Acceptability* is much like safety inspection for a vehicle, in that just because an item has no *Critical Issues*, that does not mean it is a high quality or desirable item. Rather, like an automobile that passes safety inspection, it merely is free of problems that fundamentally undermine its functioning. Medium or high quality items can *easily* pass this level of inspection, and there are additional factors that move them beyond merely acceptable into higher levels of quality.

The Misunderstood Role of Item Writers

Perhaps the most important thing to know about item writing is that the academic field of Educational Measurement, industry researchers and psychometricians, and even

[2] The rate at which item drafts by item writers are immediately rejected by CDPs is hardly a function of their quality. The primary determining factor is either contractual obligations agreed to between organizations or the personal and political ramifications of rejecting many items. This leaves CDPs with large amounts of work to do on the vast majority of items.

the some leaders in assessment organizations do not understand the role of *item writers*[3]. They wrongly believe that item writers are responsible for most content development work, when they only play a small role in the process for a short time.

One of the greatest values that item writers contribute to many projects is the ability for test proponents to say, "Every item was written by teachers from our state," or "Every item is written by an expert practitioner in our field." Their identities as teachers or expert practitioners in the field is incredibly important. But the fact is that item writers are usually far more expert in teaching and learning, curriculum and instruction, than they are in assessment. A long-standing complaint about teacher preparation programs is the lack of attention paid to assessment literacy—to say nothing of more advanced assessment topics and techniques. Efforts to streamline pre-service paths to the classroom have only limited this important area, further. Moreover, even if they are true experts in classroom assessment, that does not prepare them for the special challenges of large scale assessment. Obviously, expert practitioners in other fields (e.g., medical specialists, architects, interior designers, pharmacy technicians) are even *less* expert in assessment than teachers are. In both cases, their *lack* of expertise is the basis for that great public relations value.

Therefore, there simply is no reason to expect item writers to be able to produce high quality items *for large scale assessment*. They simply are not trained to do so, and do not have the time to stick with items through the content development cycle to learn how to spot and address issues that are not of great importance in their own classrooms or professional practice. Their public relations value as "teachers from our state" or "practicing specialists" stems from their *lack* of expertise in large scale assessment, simply because the public so distrusts standardized tests.

Hence, the term "item writer" is often used to refer to anyone who works to write or refine items, under the assumption that it is actual item writers who do most of this work. CDPs—often called content specialists, item specialists, item editors or some other organizational title—often go unrecognized and their work is largely unacknowledged. Because of the amount of specialized expert work done by CDPs, it is insulting to refer to them as "item writers." More importantly, efforts to reform, streamline or even automate item development that are not based on accurate understandings of the actual workflows of content development are unlikely to succeed.

Replacing item writers with AIG system may yield, in fact, higher quality initial item starters/drafts—though whether those early quality gains are worth lose the rhetorical *every item has been written by someone you should trust* is something that needs to be considered. Even if AIG system produce higher quality item starters than item writers do,

[3] This section assumes the stringent development standards of large scale K-12 assessment, such as that used for state accountability tests. Other kinds of assessments, including large scale assessments, may expand the role of item writers, but this invariably comes at a cost to item validity.

there remains enormous work to refine them to the point they are appropriate for large scale assessment use.[4]

The Vital Contributions of Review Committees

Review committees (e.g., Content/Validity Review, Fairness Review, Sensitivity Review, Accessibility Review) serve a number of invaluable functions in the content development process. Of course, they also allow test proponents to reassure the public that "Every item has been reviewed by multiple teachers from our state" or "…by experienced professionals from our field." But their contributions can vastly exceed that important one.

Review committees are a way to bring experiences and perspectives to the content development process that are not already available within the content development teams in test development vendors and/or their clients. For example, across a review committee, members may have deep experience with students from across the many different regions of a state, deep experience with students from many different cultural backgrounds, experience with teaching a variety of different curricula, experience with different ages of students, with students of different proficiency levels and/or with different kinds of disabilities. Members themselves may bring those different backgrounds and identities themselves. Hence, they are able to apply their substantive expertise through the lens of their experience with different perspectives founds in the test taking populations.

Therefore, review committees are well suited to checking to ensure that items will be understood by *the range of typical test takers* in ways that the content development team cannot. That is, they complement the perspective and experience of CDPs, both contributing their evaluations of how test takers might understand items *and* help CDPs to expand their own understanding of a wider range of potential test takers. Review committees can spot and explain issues in items that some CDPs cannot.

However, review committees are not item editors or CDPs themselves. Writing cannot be done well by committees, nor can editing. That really is a *too many cooks can spoil the soup* situation. Review committees are best used to identify issues, so that CDPs can compile and consider all of their feedback and then determine how to address it. Review committees do not have time to actually fix all the issues they spot, even if they could agree on how to prioritize them. Certainly, because review committees invariably are newly formed at each meeting, they have not worked out the norms and dynamics between their members that would enable them to reach agreement on the challenges of balancing priorities and/or different potential solutions to problems.

Therefore, review committees serve as an invaluable quality control mechanism—truly a critical type of substantive content and fairness validation. Their vision into items and how test takers will understand and think about them is vital to the test development process. However, taking advantage of the depth and breadth of expertise that they bring

[4] We have only seen on organization that subjects their AIG items to professional quality reviews sufficient to determine whether items ought truly ought to be used. This lack of consideration for item validity is the true blind spot *and* Achilles heel of AIG.

to the process requires taking care of all the *Critical Issues* and *Significant Issues* that CDPs can address *before* committees review items. Otherwise, their time and attention are wasted on the more obvious issues, rather than the subtle/more nuanced issues such as the issues impacting various smaller groups within the testing population.

Item Logic, Creativity and *a Good Idea*

Item logic is the basic idea or design of an item, the logic of how its components and contents prompt the intended task and targeted cognition. More complete item logic for selected response items also includes explanations for how various answer options reflect errors, understandings or misapplications of the targeted cognition. Item logic is visible in various sorts of learning and assessment, not just large scale standardized tests. Drills are often based on a sort of item logic, where the exercise includes particular KSAs with which a teacher or coach wants learners to become more practiced. Test prep often includes items that very much resemble the sorts of items found on the test. That *very much resembles* is merely the use of the same item logic.

Unfortunately, predictable items on large scale assessment create multiple problems. They lead to inappropriate test prep—test prep which targets predictable item logics. This undermines the assumptions of sampling that are required to make useful inferences from test taker performance. When test takers are prepared for particular item logics, rather than broader application or deeper understandings the of the targeted KSAs, those sampling assumptions are violated, and tests are no longer useful for their intended—or most any—purpose. More importantly, when large scale assessments are taken seriously by learners (and teachers), they often focus on predictable content and item logics, at the expense of covering the breadth of the domain model or curriculum. That is, predictable item logics have the effect of narrowing the enacted curriculum by distorting priorities about what should be taught and learned.

Therefore, creativity in item logic is invaluable. Variation in the types of item logics that appear on tests, including variation in the type of item logic use to assess a particular alignment reference, are vital to respecting the importance of teaching and learning the official curriculum and/or domain model.

Hence, an item draft that has numerous *Critical Issues*, yet presents a good new idea for new item logic for some cognition is extremely useful, and is better than a nice MLT sandwich[5]. Its value exceeds that of a merely highly acceptable item or even a highly refined high quality item, because it expands the possibilities of what might be included on tests, makes it less predictable and/or less likely to have the negative distorting effects of item predictability. Its value is so great because its value goes beyond that of a single item.

Of course, good new ideas—like true love—are rare. That is why they are so valuable. Hence, while the LBIDAT focuses on the presence and/or absence of *Critical Issues*, the CBIDAT ignores *Critical Issues* to focus on a very different standard of acceptability. That is, the *Creativity-Based Item Acceptability Tool* focuses on whether the

[5] You know, when the mutton is nice and lean and the tomatoes are ripe. They're so perky. We love that.

item draft presents a new *good idea* that could be used as the basis for items that expand variation within the item pool for an alignment reference.

Assumptions

These two item draft acceptability tools are based on a set of assumptions that form the basis for content development work. These assumptions are nearly universally understood across the range of test development organizations, be they a part of test development vendors, test sponsors, professional licensure organizations or governmental organizations.

- Items are written to align to some targeted cognition found in alignment references taken from a designated domain model.
- Items' alignment references are selected to meet a test blueprint's requirements.
- The goal of each item is to elicit evidence of the targeted cognition for the range of typical test takers.
- Items should not falsely suggest that test takers have proficiency or mastery with their targeted cognition (i.e., a Type I error).
- Items should not falsely suggest that test takers lack proficiency or mastery with their targeted cognition (i.e., a Type II error).
- Predictable items are not bad just for their undermining of the assumption of sampling that is fundamental to making inferences from test taker performance. The larger problem with predictable items is how they narrow the enacted set of learning goals (e.g., as in curriculum) from the official set of learning goals, thereby undermining instructional quality, learning and even democratic oversight of schools.

Significant Issues

While the *Critical Issues* that the LBIDAT considers must be addressed even to determine whether the item might be brought to the level of refined quality necessary in large scale assessment, there are many other types of still *Significant Issues* that must be addressed to reach that level. Many of them are generally so obvious to client representatives and review committee members that their inclusion in items that are presented to those audiences distract from any deeper considerations of the items. That is, those audiences are unlikely to have the professional skill and discipline of CDPs to look past those issues when first evaluating an item, and it does a disservice to the entire process to present them with such items.

The classic guidance provided from the work of Haladyna & Downing (1989) and Haladyna, Downing and Rodriguez (2002) contains many categories of *Significant Issues*. Rigorous Test Development (RTD) has distilled some of them into what it terms *item hygiene*. Because item hygiene issues are identifiable even by people who do not have deep expertise with the cognition being assessed or the perspectives of the range of typical takers of the assessment, they receive disproportionate attention. They are easier to spot

quickly, and therefore there is no excuse for not addressing them once the *Critical Issues* are worked out.

Many of RTD's 22 content and cognition traits—some of which are also found in the Haladyna lists—are found in the *Critical Issues* of the LBIDAT, but not all of them. They must also be addressed, as they improve an item's ability to elicit evidence of the targeted cognition for the range of typical test takers beyond the low bar of the LBIDAT.

Significant Issues include, but are not limited to:

- Proofreading issues
- Inappropriate use of regionalism or idioms
- Minor inconsistencies in or across items
- Use of qualifiers, negators and absolutes (e.g., *most, not, always*)
- Confusingly worded item elements
- Lack of frontloading
- Item type
- Poorly formatted items
- Inappropriate sources of difficulty

Many *Significant Issues* relate to improving an item's ability to elicit evidence of the targeted cognition, reducing the likelihood of suggesting Type I or Type II errors and/or reducing problems that give one group an inappropriate advantage or disadvantage of overs in responding to the item successfully.

The Low Bar Item Draft Acceptability Tool (LBIDAT)

The LBIDAT focuses on *Critical Issues* in the item draft, the kinds of problems that render and item unable to elicit evidence of the targeted cognition for the range of typical test takers without being prone to encouraging Type I and or Type II errors. Unlike the *Significant Issues* caused by problems with item hygiene, *Critical Issues* are not assumed to be fixable. Moreover, CDPs cannot be sure whether an issue is fixable until they attempt to do so[6]. Thus, the LBIDAT reports on the foundational and structural quality of an item, but not all of the refinements needed to actually appear on a test.

The LBIDAT reports can be treated as a rough guide to how much work an item requires before it is ready to be presented to a review committee, or perhaps even to a test developers' client. Because addressing *Critical Issues* can beget more *Critical Issues*, its guidance is only rough. Moreover, even an item without *any Critical Issues* may still require hours of work to clean up all the *Significant Issues*.

The Scale

The LBIDAT is a multi-dimensional, non-compensatory rubric. Users should evaluate each dimension independently, with one of three qualitative marks.

- ✓ The top mark, indicative of no *Critical Issues* in this area.
- ~ The middle mark, indicative of a low number of *Critical Issues* in this area, as might be expected initially.
- !! The lowest mark, indicative of a quite concerning number of *Critical Issues* in this area.

These marks can be reported together, but they cannot be added together or averaged, as this is a *non-compensatory* scale. Therefore, the summary of a an LBIDAT evaluation might look like one of the following, if they were to appear in a larger table of item metadata.

~ ~ ~ ~ ~

✓ ~ ~ !! ~

✓ ✓ ✓ ~ !!

The Dimensions

The LBIDAT considers five aspects of an item.

The Stem & Task. The stem of an item directs the test taker to engage in a task, promoting a cognitive path and response. This dimension primarily, though not exclusively, considers whether the task is appropriately aligned to the assigned alignment reference.

[6] Even relatively inexperienced CDPs may have a good sense that a *Critical Issue* in an item is not fixable and experienced CDPs may have a good sense that a particular Critical Issue is fixable—or even quickly and easily fixable. But these senses of a fixability of a *Critical Issue* are not final determinations. Any of them can prove wrong after serious attempts to actually fix them, though of course that is less likely with more experienced CDPs.

The Key. This dimension considers whether the key—the successful response—meets its requirements. As presented, the LBIDAT is focused on selected response items (e.g., multiple choice items), but this dimension can be modified to consider scoring guides for technology enhanced and constructed response items.

Distractors. This dimension considers unsuccessful responses, particularly in the context of selected response items. Because bad distractors can completely undermine the alignment of an otherwise appropriate task, this dimension is nearly as important to evaluating item alignment as the stem & task. The criteria can be modified to consider other kinds of unsuccessful responses for selected response, technology enhanced and constructed response items. Please not that a mismarked or unmarked key is not a *Critical Issue*, so long as the reviewer can identify the real key.

Stimulus. This dimension only examines item stimuli in the context of particular items and therefore takes a blindered view of stimuli issues. It should be evaluated again for each associated item that it is attached to. Thus, it considers whether the stimuli are appropriately/accurately described or linked by other stimuli elements of the item, and whether stimuli are properly constructed. For example, mislabeled graph axes or errors in a data table are such construction problems. As items are supposed to require analysis or manipulation of their stimuli to reach a successful response using the targeted cognition, items that may be answered by test takers without reference to their stimuli have a *Critical Issue* in this area.

Fairness. Fairness is a vast and vastly important topic. However, for LBIDAT purposes, this dimension only considers whether the stimulus and/or item raises inappropriate sensitivity topics. That is, whether it is likely to elicit an emotional response that notably distracts tests takers and thereby make them less likely to produce a successful response. In fact, this dimension should be evaluated through the lens of what the members of a fairness committee *for this assessment project* would conclude. That is, it is not a question of the what a CDP thinks really would or ought to disturb some test takers, but rather how the CDP expects a fairness committee to rule on this issue. This often requires experience with the political and/or empathetic sensitivities of a particular test sponsor/client and the kinds of people they want on their fairness review committees. (Issues that give some groups inappropriate advantage (i.e., bias) over others in the testing populations are perhaps the most important *Significant Issues* and can be the most challenging to address. However, because CDPs cannot always spot them themselves, they fall *outside* of the LBIDAT's criteria.

Critical Issues

The LBIDAT focuses exclusively on *Critical Issues*. This excludes almost all item hygiene issues and even many of the content & cognition traits of high quality items. Those may be *Significant Issues*, but they are issues for later steps in item development. *Critical Issues* in an item are those that may absolutely prevent an item from being usable, because they so undermine an item's ability to elicit evidence of the targeted cognition for the range of typical test takers. *Critical Issues* may be fixable, but it is often unclear until a CDP

attempts to address them whether or not that is the case. Unfortunately, fixing one Critical Issue can create a new—perhaps even more than one—critical issue. Therefore, one cannot simply sum up the number of *Critical Issues* in an item.

The *Critical Issues* that should be considered for each dimension listed on the LBIDAT form (see next page). Of course, any given project may alter those lists, but other issues, be they item hygiene (Wine & Hoffman, 2022), content & cognition traits (Wine, Glore & Hoffman, 2025) or anything else should be identifiable by CDPs without the support of expert external review committee members.

Low Bar Item Draft Acceptability Tool (LBIDAT) Form

Stem & Task

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

- The prompted task is based upon a misunderstanding of the targeted cognition.
- The prompted task is based upon a misunderstanding or misreading of its own stimulus or the work it attempts to prompt.
- 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 an exact 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

The Creativity-Based Item Draft Acceptability Tool (CBIDAT)

The Creativity-Based Item Draft Acceptability Tool (CBIDAT) is fundamentally different than the Low Bar Item Draft Acceptability Tool (LBIDAT). The CBIDAT is built exclusively around the question, *Is this a good idea for an item?* As the name suggests, the CBIDAT focuses on whether the item draft presents a workable *new* idea for an item—its potential to help expand the variety of items in the pool for an alignment reference or standard. That is, whether the item logic, application of the KSAs and/or the context is a) sufficiently unlike items that have come before to make items that test takers encounter less likely to narrow the enacted curriculum and b) whether the idea has the potential to support a complete item that could get through the LBIDAT.

Thus, the CBIDAT is an alternative view of the value of an item draft that is rarely appropriate, yet nonetheless incredibly important when its criteria are met. Therefore, items that do quite poorly on the LBIDAT may do well on the CBIDAT. The harm of predictable items—which both narrow the enacted curriculum and undermine the assumption of sampling—is so great that an item draft that clears the bar of the CBIDAT is worth far more than any item that clears the low bar of the LBIDAT. A *good idea* for an item—a good *new* idea for an item—is invaluable. This is because good ideas are so rare, and yet without them teachers and learners respond to what is on the tests, rather than teachers and assessments responding to what is in the domain model (e.g., state learning standards).

When judged through the lens of the CBIDAT, an item draft must offer some novel item logic for assessing an alignment reference, which could take the form of a novel application, a novel context or even a larger new idea.

The Scale

The CBIDAT is a multi-dimensional, non-compensatory rubric. Users should evaluate each dimension independently, with one of five qualitative marks. Anything less than the highest marks for a dimension may be an unrecoverable failure.

- ✓+ The top mark, compatible with the most valuable sort of new item logic.
- ✓ A top mark, indicating reasonable promise for a good new idea.
- ~ A middle mark, indicative of very low likelihood that the item draft presents a new good idea.
- !? A mixed mark, suggesting while the idea does not apply to the *assigned* cognition, it may be a good idea for some other element of the domain model.
- !! The lowest mark, indicating that that there is virtually no chance that the item presents a *new good idea*.

These marks can be reported together, but they cannot be added together or averaged, as this is a non-compensatory scale. Therefore, the summary of a an CBIDAT evaluation might look like one of the following, if they were to appear in a larger table of item metadata.

Creativity-Based Item Draft Acceptability Tool (CBIDAT)

~ ~ ~

✓+ ~ !!

✓ !? ✓

The Dimensions

The CBIDAT considers three aspects of an item.

Novelty. The entire purpose of the CBIDAT is to identify potential *new* good ideas, so novelty is its primary dimension. Evaluating this dimension depends on the deep knowledge that comes only from considerable and considered experience in assessing a particular content area and/or domain model.

Alignment. Alignment with the targeted cognition or alignment reference is one of the most fundamental aspects of item quality—one that pervades every dimension of the LBIDAT. This dimension considers whether the best-case development of this idea into an item is aligned. That is, it is a professional judgment of the potential for this item.

Suitability. This dimension considers whether the best case implementation of this idea could meet the requirements of the various item types and other constraints in a project. As presented below, this presumes the requirements and constraints on multiple choice items and seat time budgets. However, this dimension can be altered for other item types, test platform and assessment constraints.

Critical Issues

Only the *Suitability* dimension of the CBIDAT considers *Critical Issues* as the LBIDAT does. Because the CBIDAT considers the potential for the item, were it developed into a best case item, any critical issue is almost certainly fatal in fact. The other two dimension focus on scales that are barely ordinal, each explained on the CBIBAT form (see below).

Creativity-Based Item Draft Acceptability Tool (CBIDAT)

Novelty

New item logic for previously under-tested KSAs	New item logic for already tested KSAs	Novel context, but cloned item logic	Virtual clone of existing items	
✓+	✓	~	!!	

- ✓+ The item draft presents new item logic for previously under-tested KSAs in the test blueprint—for which there is interest in adding them to the blueprint.
- ✓ The item draft presents new item logic for KSAs that already have good item logics.
- ~ The item draft uses a preexisting item logic, but applies it in a truly novel context.
- !! The item draft does not present any significant novelty.

Alignment

Aligned to the core of the alignment reference	Aligned to the margin of the alignment reference	Aligned to some other alignment reference or level	Not aligned to any alignment reference	
✓	~	!?	!!	

- ✓ The item draft presents an idea that could elicit evidence of most important part of the alignment reference, the core that is most worth assessing given limited assessment time.
- ~ The item draft presents an idea that would elicit evidence of some less important part of the alignment reference which not worth using limited assessment time on.
- !? The item draft presents an idea that is not aligned to the alignment reference at all, but is aligned to some other alignment reference of interest in the domain model.
- !! The item draft presents an idea that is not aligned to any alignment reference of interest in the domain model.

Suitability

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

Critical Issues

- The idea does not result in a definitively correct key.
- The idea does not support a sufficient number of distractors.
- The required stimulus and task would take too much time.
- The idea requires inappropriate amounts of learning during the task by test takers for a successful response