**Item Review Procedures**

*Assessment Item Modification (AIM) for Equity Project*

**Item Analysis**

Pursuant to recommendations\* from the Council of Great City Schools (CGCS) report released in 2017,[[1]](#footnote-1) the Department of Data, Research, Evaluation, and Assessment (DREA) is piloting additional procedures for reviewing assessment items. These procedures capitalize on already available data and resources for psychometric test item analysis and will aid in detection and correction of aberrant and biased assessment items.

Procedures outlined here target test items that exhibit one (or both) of these limitations:

(A) ineffectiveness in student content mastery based on analysis of the full population of students tested;

(B) differential item functioning (DIF), wherein two given students of the same ability level but different demographics (e.g. race, gender) perform differently on the selected item.

**(A) Item Performance for General Student Population**

These procedures use techniques based on item response theory (IRT) to identify assessment items that underperform for the general student population (category A, above) based on discrimination (alpha value), difficulty (beta value), or aberrant response (c value) parameters.[[2]](#footnote-2) The thresholds for flagging items for review based on each of these criteria are as follow:

**Discrimination**: Parameter values of 0.5 or lower

**Difficulty**: Parameter values outside of +/- 2 standard deviations from the mean ability level (i.e. items that are uninformative about >95% of test takers)

**Aberrant Response**: Parameter values higher than the probability of a correct answer with an uninformed guess (e.g. >0.25 for a 4-option multiple-choice item)

**(B) Differential Item Functioning (DIF) Analysis**

In order to combat a legacy of inherent bias—particularly along lines of race, gender, socioeconomic status, and other factors of identity and privilege—among the historical instruments of standardized testing, these procedures leverage DIF analysis. Based on student response patterns, this approach seeks to identify whether a given item yields different information for two students of the same ability/content mastery level who represent different demographics (see Appendix B for more information).

Because a student’s identity is never fully represented by one factor alone, such as their race, these procedures conduct DIF analysis for the following subgroupings in all cases for which sufficient assessment data is available:

* Race
* Sex
* Gifted status
* Limited English Proficiency status
* Disability status
* Socioeconomic status.

In cases where insufficient data is available to indicate whether an item exhibits DIF, these procedures create documentation of the limitations of the underlying data.

The presence of human judgment in analyzing bias in assessment items is critical. Therefore, these procedures first identify DIF based on the statistical significance—rather than the degree or direction—of the difference between subgroups’ performance on a given item. As discussed in Appendix B**, the key task of district personnel is to review visualizations of items flagged for statistically significant DIF (i.e. unlikely to have occurred by chance) and determine whether the degree of DIF warrants further assessment.** This is because there may be many items for which DIF is statistically significant but the actual degree of difference is in fact microscopic.

**Methodology**

*Which Assessments Will Be Analyzed?*

**The analysis outlined above will be conducted for all district-level assessments for which sufficient data exists, beginning with the 2018-2019 school year**. Due to limitations of statistical analysis (particularly within item response theory methodology) certain assessments may not be feasible to analyze. These mainly include:

* Assessments with too few items
* Exclusively rubric-scored assessments (unless >5 items are included)
* Assessments with too few student responses (e.g. fewer than 500 total responses for general item characteristic analysis or fewer than ~500 per group for DIF analysis).

Assessments will be included for analysis regardless of source (e.g. purchased from a testing company, or written internally by district curriculum and instruction professionals).

*Which Students and Demographics Are Considered?*

As discussed above, DIF analysis will be conducted on bases of race, sex, giftedness, English proficiency, special needs, and socioeconomic status. Noteworthy limitations for DIF analysis among these groups:

* African American students are chosen as the reference group for DIF analysis by race due to the high priority of improving opportunities for this demographic in accordance with the current strategic plan. While inclusion of other subgroups is also a high priority, limitations of small sample sizes for non-African American, non-Caucasian student groups would make DIF analyses for these groups statistically indefensible at this time (as of 2019).
* Similarly, female and male subgroups are used for DIF analysis based on sex due to insufficient sample size among students who do not identify with binary labeling options.
* Small sample sizes of students labeled as gifted, having limited English proficiency, or having disabilities makes DIF analysis unreliable in many (but not all) cases; reports will note limitations of sample size for these groups as applicable.
* Currently, the Federal Poverty Rate (FPR) is the most accurate indicator of socioeconomic status available at the district level. The low threshold for FPR qualification, however, makes it a crude indicator of economic distress. I.e., a family of four earning, for example, $15,000/year above the FPR may still fall well within a better qualitative definition of economic distress. The DREA department is currently investigating more accurate methods for estimating socioeconomic status.

Analysis will be conducted for student responses at all school buildings and for all grade levels for which applicable assessments have been administered.

**Output and Documentation**

This analysis leverages item-level student response data and student demographic information pulled from the Unify (Performance Matters) system. Student ID numbers are used only to match individual student response patterns with the associated demographic information, then individually identifiable information (name, etc.) is deleted from the data set to protect student privacy.

The resulting assessment response information is analyzed using a program written in the R statistical analysis package; underlying code for the program is available upon request or can be viewed on the project’s GitHub repository. When run for a given assessment, the program conducts each category of analysis listed in the preceding section and creates a directory with a written report in the form of a MS Word document outlining all items flagged for any reason as well as the specific metrics that raised each flag and visuals of DIF as applicable.

Within the same output directory, the program also generates:

1. A spreadsheet of item response characteristics and DIF significance for *all* items, regardless of whether they were flagged
2. PDF files containing visualizations of item characteristic curves (ICCs) for all items exhibiting DIF (a separate file for each demographic category)
3. A CSV file of the analyzed item response data (post-processing)
4. a separate CSV of student response scores for items that could not be analyzed (e.g. for a 25-item test with 23 multiple choice items and 2 rubric-scored items, the 2 rubric-scored items would not generate sufficient data for defensible item response analysis and would therefore be included in this file).

These non-report files (A-D, above) are preserved to ensure transparency around district processes for item analysis. Should the need arise, their existence enables review and verification of the methodology and analysis used herein.

**Item Review Procedures**

A fair and effective item review process must be inclusive, rigorous, and transparent. The procedures below ensure each flagged item is reviewed thoroughly while respecting the value of the time required of experts—including Curriculum and Assessment staff, teachers, and other district personnel—and key stakeholders—including students, parents, and community members.

This process aims to submit items that are both problematic and actionable to the final review committee, encourage transparency throughout the process, and ensure that problems and actions taken are well documented.

There are three phases of the review process: (1) internal (DREA) vetting, (2) inter-departmental (e.g. Curriculum Department) review, and (3) community panel review.

1. **Internal Vetting**

*Purpose*

DREA staff will review each item to screen for low-level flagging issues. The following problems can be identified immediately, and may be remediable easily and internally, without requiring the time of other staff or external partners:

* Simple system errors—e.g. incorrect answer marked as correct on the key
* DIF flags with negligible subgroup differences

*Parties Involved*

\_\_\_\_\_, \_\_\_\_\_\_, and \_\_\_\_ (*position titles*) will conduct first round internal review of assessment items. This group will review the assessment analysis report for the issues above and eliminate items with minor errors that can be easily fixed as well as items not warranting further consideration.

For all items that raised a flag, regardless of severity, the internal review committee will document the following information (using a shared MS Access database for consistency in documentation formatting) before passing the report to the next phase of review:

* Assessment title/ID
* Item flagged
* Reason for flag
* More detailed problem identification (e.g. misleading distractor, ambiguous question phrasing, test/answer encoding error, etc.)
* Action(s) recommended
* Action(s) taken (if already done)
* Date action(s) taken

1. **Inter-departmental Review**

*Purpose*

Interdepartmental review has two purposes: (1) submit items exhibiting general problems (i.e. discrimination, difficulty, or aberrant response) to the proper experts—e.g. Curriculum/Assessments teams—for content revision, and (2) identify which items among those already flagged for DIF are actionable/eligible for community panel review.

*Departments Involved*

While the DREA Department will assist in coordination of review for step 2, the primary parties involved should be representatives from the Curriculum/Assessments team(s).

*Process*

The panel will go several steps further than the initial internal vetting committee (step 1) by first reviewing the actual content of all submitted items. Second, items flagged for general problems should be revised by the committee itself (particularly by qualified Curriculum and Assessment staff). Third, the committee should mark items with significant and substantial DIF for further review by a community panel, provided that the items in question appear actionable and are free of obvious simple errors. Items with obvious errors (e.g. small typos. etc.) should be revised by the inter-departmental review committee.

*Documentation*

After determining which items warrant submission to the final review panel and which should be revised by Curriculum and Assessment, committee members should document the decisions using the same database as the internal vetting committee in order to ensure sustainability and consistency in documentation. If the committee chooses to revise any questions, revised versions and records of assessment IDs/titles and question numbers should be sent to the proper department to be re-uploaded to the assessment management system. Revisions should also be documented in the item review tracking database.

1. **Final Review**

*Purpose*

External, community review of assessment items flagged for bias is the most important element of the item review process. Through external review, the district not only empowers parties—such as students, parents, community members, and district partners—traditionally left out of assessment decisions to become critical voices in the process, but also improves the quality of item review through critical perspectives that may capture difficult-to-spot sources of bias within item content.

*Parties Involved*

The community review panel should comprise stakeholders from across the spectrum of engagement with district assessments. Most importantly, multiple students from each DIF subgroup (e.g. both economically disadvantaged (ED) and non-ED students, if DIF based on ED status is identified) should be included in the process—either for pre-committee item review or on the committee itself. Additional members of the final review committee should include parents, teachers, district representatives from Curriculum and Assessment, school-based partners (e.g. mentoring program staff), and university partners.

*Process*

Once a significant collection of assessment items (e.g. a quarter’s- or semester’s-worth) have been analyzed and flagged for final review, the items shall be submitted to the community panel. An appointed facilitator, acting as a neutral party for the review process, will help the committee seek consensus around item content revisions as appropriate.

To avoid “anchoring” bias, each party should review the body of flagged items independently to determine their first impressions of the source of bias prior to an in-person or virtual meeting during which consensus is reached. Unlike those involved in the first two phases, the community panel should not be informed of which subgrouping was the basis of any given item’s DIF detection. This method aims to circumvent the fallacy of reviewers conducting analyses based on preconceived notions that may be biased themselves.

*Documentation*

The facilitator should take note of the group’s consensus on each item and record group decisions using the same spreadsheet as in the first two phases of review. The committee’s suggested item content revisions will be sent through the reverse-trajectory of the item review process for confirmation: first for approval by the Curriculum and Assessment Offices, then to DREA to be re-uploaded to the assessment management system.

***Appendix A***

**Reading Item Characteristic Curves (ICC’s)**

Comfort with understanding and interpreting ICCs is crucial for effective use of the procedures set out in other parts of this document. This section explains the ICC elements that users will need to comprehend for the purposes of district assessment item analysis.

**What is an ICC?**

Simply put, an ICC represents a single assessment item in terms of student ability level vs. probability of a correct answer.

*Y-axis*

The y-axis measures the probability of a student answering the question correctly; the bottom of the y-axis corresponds to a 0% chance of answering correctly, while the top corresponds to a 100% chance.

*X-axis*

*The* x-axis represents the spectrum of student ability levels among the population of students who completed the assessment. This can also be interpreted as content knowledge for the unit/class/subject being tested. At the center of the x-axis, the value of 0.0 represents the average student ability level (which is not to be confused with having zero ability), and each unit up or down is a change of 1 standard deviation. An x-value of 3.0, therefore, represents student ability level that is 3 standard deviations above the mean, while an x-value of -0.5 represents a student who has slightly lower than average mastery of the tested content. Note: “ability level” here refers only to ability **only** in the specific content represented by the test and **only** relative to the other students who completed the same test. The x-axis of an ICC does not (and does not claim to) serve as an estimation of general intelligence or student ability in other domains.

*Interpreting Coordinates*

Each coordinate on an ICC represents the intersection of a given student ability level and probability that a student with that ability level will answer correctly. If an ICC crosses a point at (3.0, 1.0), for example, this means a student whose ability level is 3.0, or 3 standard deviations above average (i.e. very, very strong content mastery) has a 100% chance of answering the question correctly. If the same ICC crossed the point (-1.5,0.5), that would indicate that a student somewhat below the average ability level (i.e. moderately poor content mastery) has a 50% chance of answering the question correctly.

*Reading the Graph*

There are three main elements of an ICC a user must know in order to understand the procedures in this document: the discrimination parameter, the difficulty parameter, and the aberrant response parameter.

**Discrimination**

*What is it?*

The discrimination parameter (also often referred to as the “alpha value”) describes how effectively an item identifies the ability level of a given student. In other words, is this question good at telling us if a student has average-level mastery of the content? Exceptional mastery? Needs remediation?

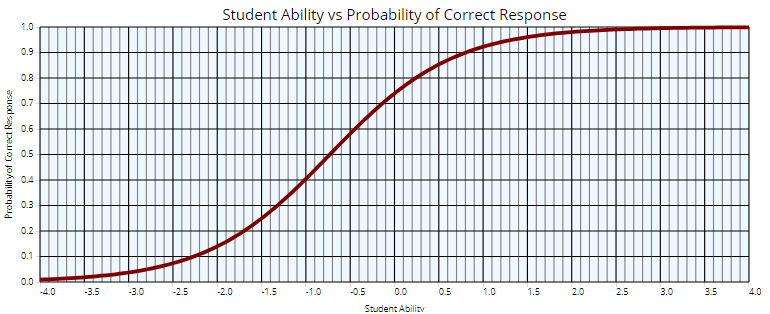
*Isn’t Discrimination Bad?*

In this context, discrimination has nothing to do with bias or prejudice; it refers only to how well an assessment item discriminates between different student ability levels.

*What Part of the Graph is it?*

The discrimination parameter for an item is represented on an ICC chart by the slope of the curve, particularly between the two “elbows” (the bottom and top bends) of the S-shape. In the example below, the yellow highlighted portion between the blue lines roughly represents the section of the curve considered in the discrimination parameter.

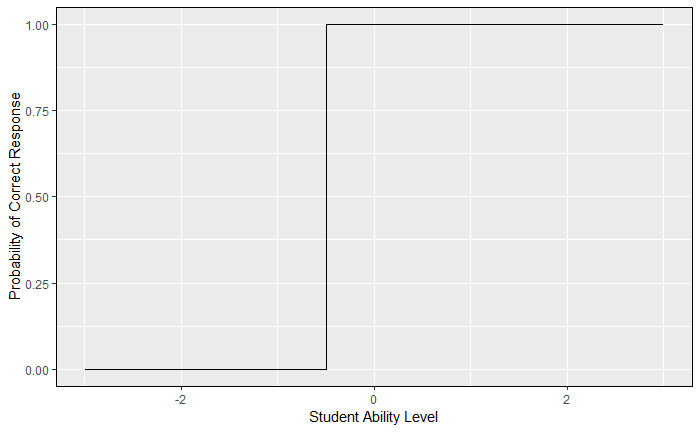
*Figure 1: Item Characteristic Curve (Strong)*



*What Does a Good/Bad Discrimination Parameter Look Like?*

Theoretically, the ICC for an item with the perfect discrimination parameter would appear as a straight vertical line, as in *Figure 2*, rather than an S-curve.

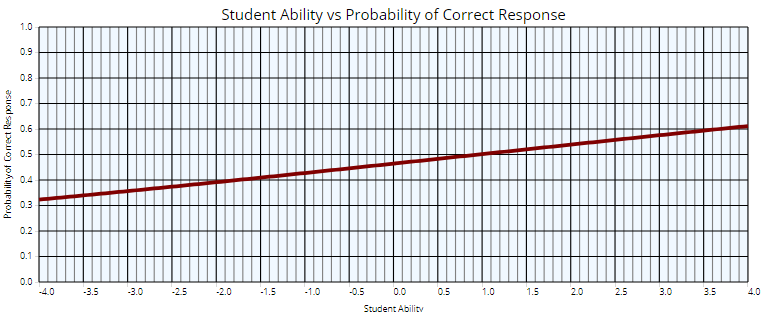
*Figure 2: Hypothetical ICC with “Perfect” Discrimination Parameter*



This would mean 100% of students below the ability level of the vertical line’s x-value would answer incorrectly, and 100% of students above that value would answer correctly. We could be completely certain that students answering incorrectly fall below the question’s x-value and may need remediation.

Questions in the real world will not be “perfect discriminators,” and the slope depicted *Figure 1*, for instance, represents a decent question in terms of its discrimination parameter. In contrast, you can easily see why **the item depicted in *Figure 3,* below, is not a strong question**: students with virtually no content mastery (-4.0 on the x-axis) have a ~33% chance of answering correctly, and students with extraordinary content mastery (4.0 on the x-axis) have only a ~61% chance of answering correctly. In short, a correct answer is not very informative about the student’s underlying content mastery.

*Figure 3: ICC with Weak Discrimination Parameter*



*What is a Good/Bad Number for the Discrimination Parameter?*

Regarding the numerical value of the discrimination parameter, the calculus required for its computation is beyond the scope of this appendix. For the purposes of these procedures, though (and as a general rule of thumb in item response theory), a discrimination parameter greater than or equal to 1.0 is considered strong. As described elsewhere, the analysis used here sets <0.5 as the threshold indicating a poor discrimination parameter that warrants further review of the assessment item.

For the sake of building your intuition, the “weak” item depicted immediately above has a discrimination parameter of 0.15 (very bad) and the preceding “decent” item (with the highlighted S-curve and blue cutoff lines) has a discrimination parameter of approximately 1.5 (fairly good!).

**Difficulty**

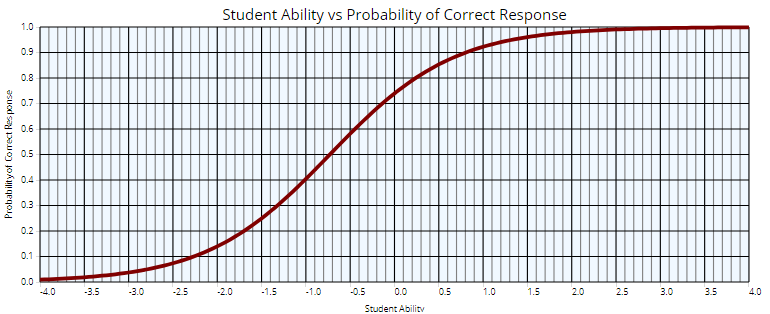
*What is it?*

The difficulty parameter (also often referred to as the “beta value”) of an assessment item technically refers to the value of student ability level (x-axis) at the point where the probability of a correct answer (y-axis) is equal to 0.5 (or 50%). In slightly simpler terms, the difficulty parameter tells us what student ability level the question is most informative about. In simplest terms, it tells us how hard the question is.

*What Does It Look Like?*

In *Figure* 4, below, the red guide lines help us see that at the point on our curve where the probability of a correct response (y-axis) is 0.5, the corresponding student ability level (x-axis) is approximately -0.8. So, a reviewer can eyeball this item as having difficulty of ~-0.8 and say that this item is best at identifying whether a given student has slightly below average content mastery. In general, a lower difficulty parameter correlates with an easier question, and vice versa.

*Figure 4: ICC Difficulty Parameter*

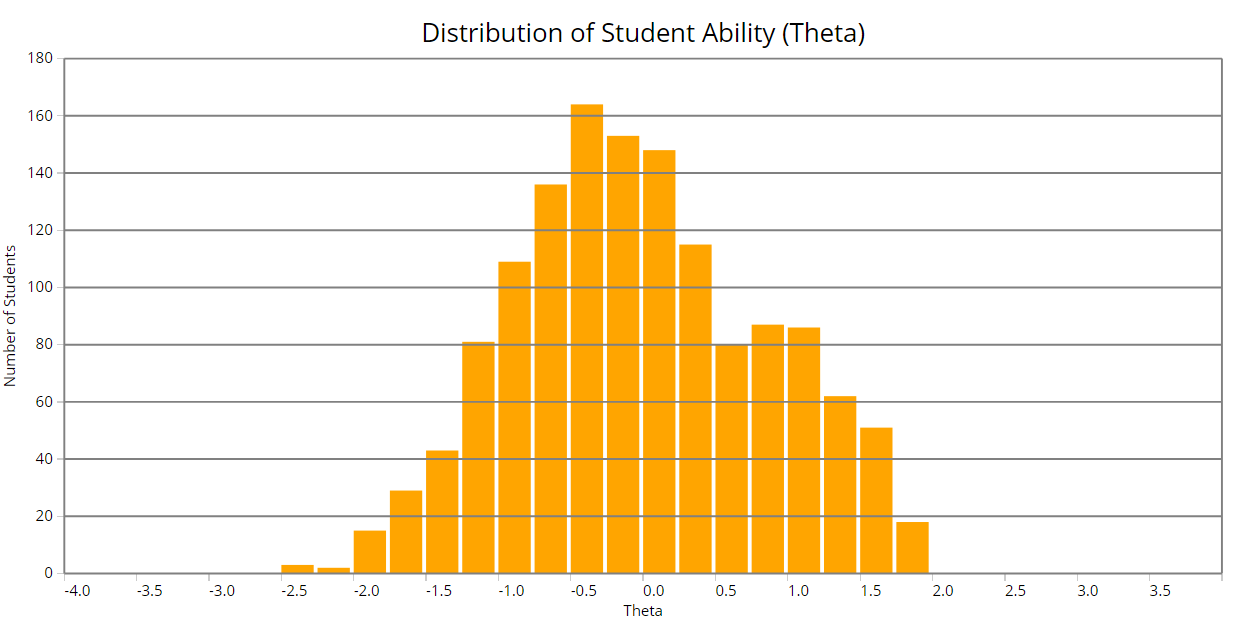


*What is a Good/Bad Difficulty Parameter?*

An effective assessment will include questions with varying difficulty parameters in order to distinguish where students fall along many points of the ability level spectrum. However, because the distribution of student ability levels tends to center around the mean (i.e. 0.0 on the x-axis), questions that measure very far above or very far below the mean will be uninformative about most students simply because there are so few, if any, who fall above or below, e.g., 3 standard deviations from the mean.

To illustrate that concept, the distribution of estimated student ability levels from the same test these example ICCs have been drawn from is pictured in *Figure 5*.

*Figure 5: Distribution of Student Ability/Content Mastery Levels*



In this case, there are in fact no students at all who fall more than 2 standard deviations above the mean or more than 2.5 below the mean. Therefore, a question on this test with a difficulty parameter of 3.5, for instance, would have been useless on this assessment, as it would have applied to zero of our students.

(More precisely, for those who are interested: such a question would only tell us that students who answered incorrectly are likely to have ability levels below 3.5—something we are already nearly certain of based on the Empirical Rule in statistical modeling (i.e. given an approximately normal distribution, which appears to apply here, 68% of observations fall within +/-1 standard deviation, 95% fall within +/- 2, and 99% fall within +/- 3). The question would not give us any new information about the student population.)

*What Does That Mean for Me as a Reviewer?*

As described elsewhere, the analysis associated with these procedures will flag items with difficulty parameters more than 2 standard deviations above or below the mean student ability level (i.e. items that inform us about fewer than ~5% of all tested students). For our purposes, an assessment having one or two items with difficulty parameters between +/- 2-3 are not cause for concern; however…

(A) items beyond 3 standard deviations from the mean should be revised, as they likely inform us about almost no students at all, and

(B) if an assessment has more than 2 items between +/- 2-3 standard deviations, reviewers should either revise all but 1-2 of those items, or choose to revise all of those items. Again, this is because >2 of them on the same assessment gives that assessment a disproportionately large number of items that are informative about a disproportionately small number of students.

**Aberrant Response**

*What Is It?*

The aberrant response parameter (also sometimes referred to as the “c value”) represents the likelihood that a student with no content mastery whatsoever will answer the item correctly. This parameter is often misleadingly called the “guessing parameter;” however, it is not a representation of whether students did or did not guess the answer to the question. “Guessing” implies a student with *any* level of content mastery was *randomly* *selecting* a response; what the aberrant response parameter actually represents is the probability that a student with a *specific* content mastery level (i.e. none) will answer correctly when they are indeed attempting to get it right (i.e. *not randomly selecting*).

*Why Does It Matter?*

A question with a high aberrant response parameter—even if it is otherwise an excellent question—is less reliable in what it tells us about student content mastery. That is, even if a question were a perfect discriminator (straight vertical line) between students above/below an ability level (x-axis) value of 1.0, if it had an aberrant response value of, e.g., 0.5, then we would be stuck interpreting a correct answer as follows: “this correct answer could indicate that the student is above a 1.0 ability level, but there is also a 50% chance that the correct answer is telling us nothing at all.”

In other words, **as our aberrant response parameter (c-value) increases, our confidence in anything we learned from the discrimination and difficulty parameters decreases.**

*What Do We Do with the “Guessing Parameter” Idea?*

With regard to the aberrant response parameter, the misnomer of “guessing” is useful in one way: it gives us a rough benchmark.

In the case of a 4-option multiple-choice item (A, B, C, D), we know intuitively that the chance of guessing the correct answer randomly is 25%. In many cases, we use that as our threshold for what an acceptable vs. unacceptable aberrant response parameter should be. I.e., in this case if a student with no content mastery whatsoever has a 40% probability of answering correctly, we know that that is 15% higher than the 25% chance from a random guess.

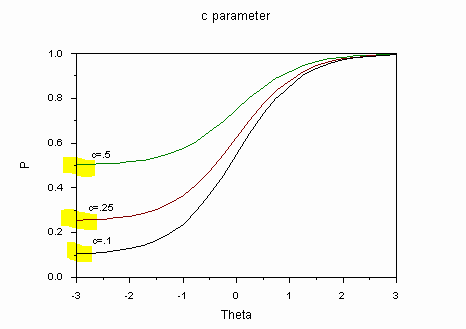
That is a bad thing. In an ideal world, higher content mastery (x-axis) should increase probability of a correct answer (y-axis), and lower content mastery should decrease that probability. Therefore, as these procedures describe elsewhere, **district procedures flag for review any items with aberrant response parameters that are higher than the probability of answering correctly with a random guess.**

*What Does Aberrant Response Look Like?*

In the ICC charts you will see, aberrant response is represented by the point at which the S-curve intersects the y-axis (basically, the y-intercept). That is, the probability of answering correctly (y-axis) for a student with the lowest possible content mastery level (x-axis).

The highlighted spots along the y-axis of *Figure* 6, below, represent the aberrant response parameters for three different hypothetical items. As you can see, the item represented by the black curve (bottom) has an aberrant response parameter value of 0.1, which is generally acceptable for a 4-option multiple choice item; the red curve’s (middle) aberrant response value is exactly the same (0.25) as the chance of answering correctly with a random guess, and would be flagged although it is only mildly concerning; the green curve’s (top) aberrant response value of 0.5 is in the unacceptable range, and that item would warrant further review.

*Figure 6: Three ICCs with Varying Aberrant Response Parameter Values*



*What Does This Mean for Me as a Reviewer?*

The cause of a high aberrant response parameter is not always obvious. However, many cases you review will be related to one (or both) of the following issues:

(A) obvious signaling in the correct answer—in these cases, there may be some characteristic of the wording or structure of the correct response that sets it apart from the other possible responses; such a question may be less a determinant of content mastery than of simple common sense.

(B) weak distractors—often, these will be easier to spot for reviewers with strong knowledge of the tested content and grade-level appropriate difficulty; in other cases, however, it will be clear that one or more distractors are silly or very easily ruled out. If an item asked students, for example, to identify the product of 11 x 9, response options between 20 and 200 may reasonably detect content mastery while an answer over 10,000 would be obviously incorrect, even to students without strong knowledge of multiplication (note, however, that obviously incorrect response options are sometimes designed to detect specific student misunderstandings—in this example, a student who chose “20” as their response may struggle to distinguish between addition and multiplication).

*Note: Figures not cited in this appendix were created by the author of this document using various libraries in the R statistical analysis program, downloads from the PPS Performance Matters (Unify) system, or open-source images from the internet.*

***Appendix B***

**Interpreting Differential Item Functioning (DIF)**

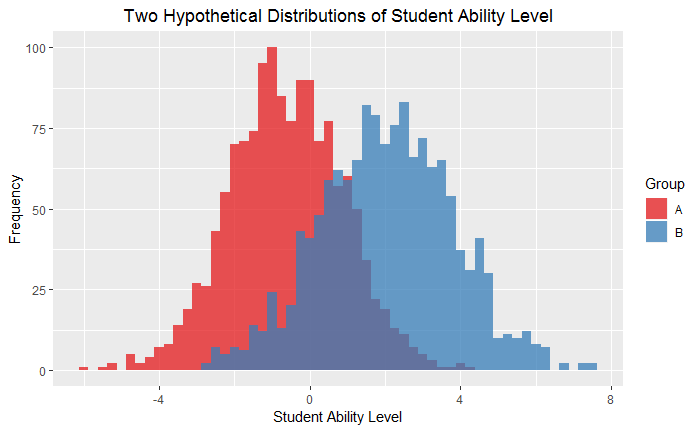
In order to contribute meaningfully to the district’s assessment item review process, reviewers should be able to understand DIF and interpret its depiction on an item characteristic curve (ICC) plot. This appendix explains the basic elements of DIF that reviewers will need to know for the purposes of district assessment item analysis. **Those unfamiliar with ICCs should review the preceding appendix, regarding ICC interpretation, before proceeding with this section**.

**What is DIF?**

DIF is a measure of how consistently an assessment item measures content mastery across different student subgroups. That is, **DIF exists when two students from different subgroups but with the exact same ability level have different likelihoods of answering the same question correctly.**

Note that **this is not the same as comparing the average scores of different groups of students on a given assessment or assessment item**. To illustrate, consider an example of the two subgroups depicted in *Figure* 1, below—for the sake of this example, assume Group A is a district’s 2nd grade students and Group B is the same district’s 3rd grade students. Group B clearly has, on average, stronger content mastery than Group A. Therefore, we expect the two groups to have different average scores on a perfectly un-biased assessment. We wouldn’t consider this assessment problematic; in fact, it is effective in differentiating between different levels of content mastery, which is the goal of most assessments.

***Figure 1: Two Hypothetical Distributions of Student Ability Level***



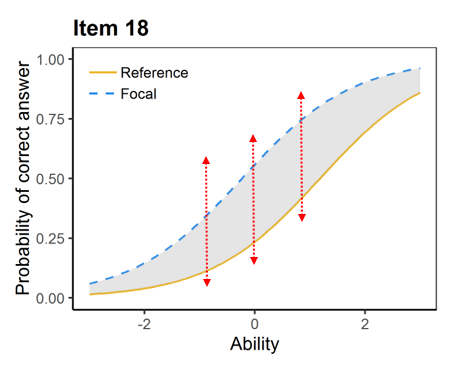
DIF is a more rigorous measure of whether an assessment item is problematic. As noted above, DIF occurs when students with the exact same level of content mastery score differently on an assessment item according to their subgroup identity (e.g. race, sex, ethnicity, socioeconomic status, etc.). In the graph above, you could think of DIF as a measure of how consistently students who fall within the overlapping section perform on a given test item, according to their membership in Group A or Group B. (This is an oversimplification of the mathematics behind DIF, but it may be useful for developing basic intuition of what DIF represents.)

**What Does DIF Look Like?**

DIF is often depicted as an ICC with two curves—each representing one of the two student subgroups under comparison. While the two subgroups could represent any conceivable categorization, for the purposes of this analysis, the lines on the charts you review will usually represent (current as of summer 2019) race, sex, gifted status, socioeconomic status, limited English proficiency status, or disability status.

In the *Figure 2* below, let the solid “Reference” group curve represent students not labeled as “gifted” and the dashed, “Focal” group curve represent students labeled as gifted. When looking at a visualization of DIF, many will find it most intuitive to compare curves along a hypothetical vertical line, as indicated by any of the superimposed arrows in *Figure 2*. This makes it easy for the reviewer to interpret, for example, as follows: “on this assessment item, students with average content mastery but *not* labeled as ‘gifted’ are over 25% less likely to answer correctly than students of the exact same ability level who *are* labeled as ‘gifted’. This item shows DIF that disadvantages students not labeled as ‘gifted.’”

***Figure 2: “Vertical” Interpretation of DIF***



Reviewers can add complexity to their interpretation by comparing the DIF at different points along the x-axis. In this case, for instance, a reviewer may notice that there is greater DIF for students who are close to average ability level; the item is slightly less biased for students far above and far below the mean. This is a common pattern for items exhibiting DIF.

**Are there Different Types of DIF?**

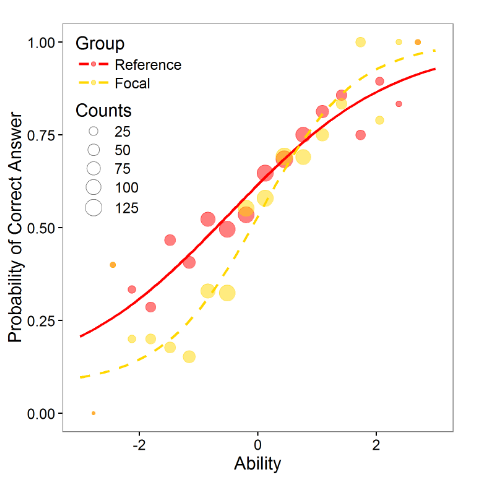
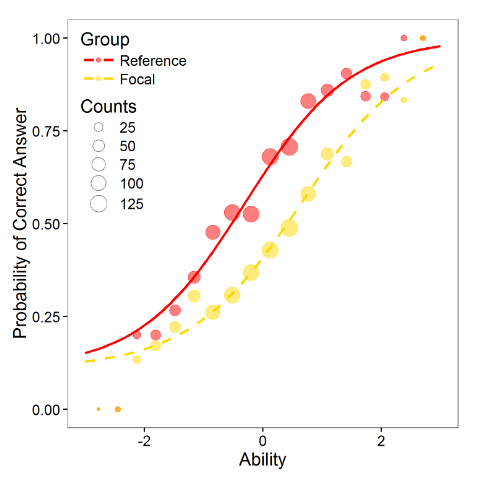
Yes. The example above depicts “uniform” DIF—by far the most common type in cases you will likely review. Items exhibiting uniform DIF favor one group over the other along all points on the ability level spectrum (x-axis). However, assessment items occasionally exhibit non-uniform DIF as well.

Non-uniform DIF is precisely what it sounds like: DIF that favors one group at certain levels of content mastery and the other group along other parts of the content mastery spectrum.

Keeping with the subgroups of the prior example (Reference = students not labeled as gifted, Focal = students labeled as gifted), *Figure 3* below shows the difference between uniform and non-uniform DIF for two different hypothetical assessment items. For the item depicted in *Figure 3A* on the left, students not labeled as gifted have a higher chance of answering correctly regardless of ability level. In the non-uniform case depicted in *Figure 3B* on the right, the item is biased in favor of students not labeled as gifted when we compare students at or below average ability level, but the item favors students labeled as gifted when we compare those with above average (roughly ~0.5 on the x-axis and above) ability level.

***Figure 3: Uniform vs. Non-Uniform DIF***

***Figure 3A: Uniform DIF***



***Figure 3B: Non-uniform DIF***

*Figure Source: Martinkova, P., et al. “About DIF.” ShinyItemAnalysis. June 2019. Retrieved from:* [*https://shiny.cs.cas.cz/ShinyItemAnalysis/*](https://shiny.cs.cas.cz/ShinyItemAnalysis/)

**How Much DIF Must an Item Exhibit in Order to Raise a Flag for Review?**

Unlike in cases of ICC parameter problems (see preceding appendix), thresholds for determining how much DIF is problematic are less appropriate to determine a priori and enforce with a computer program. As discussed elsewhere, the procedures recommended in this document acknowledge that the nature of bias in assessment items can be nebulous and is best left to be judged by the minds of experts and stakeholders.

In that spirit, **these procedures do not flag items based on the degree or direction of DIF** (i.e. how far apart the curves are at any point along the x-axis and which group is advantaged or disadvantaged).

**Instead, this analysis flags all items for which the DIF—no matter how small or large—is statistically significant.** In other words, regardless of how severe the bias is, all items wherein exhibited DIF is unlikely to be due to chance will be flagged. (For those interested, the confidence level for DIF detection is set at 90% (alpha=0.1), which maintains a credible threshold for statistical significance while erring on the side of over- rather than under-labeling. That is, if a mistake were to occur, these procedures would prefer the consequence to be consideration of a non-DIF item for review, rather than failure to detect an item that may in fact have exhibited DIF.)

**When Should Items with DIF be Reviewed?**

Considering the above, you should expect substantial ambiguity in the item review process. Setting too-specific rules for determining how much DIF is problematic would undercut the purpose of an item review panel.

For the purposes of this appendix, these are the most important points for reviewers to note:

* **Not all DIF flags will warrant further item review**

In the *Figure 4* below, the item depicted in *Figure 4A* on the left has been flagged because the DIF between two groups was statistically significant. In this case let’s presume Reference = students from families with incomes *below* the Federal Poverty Rate (FPR) and Focal = students from families with incomes *above* the FPR.

Clearly, DIF between the two groups is nearly imperceptible. A reviewer looking only at the graph on the left should be able to examine it and note that *this item does not merit further review on the grounds of DIF by socioeconomic status*.

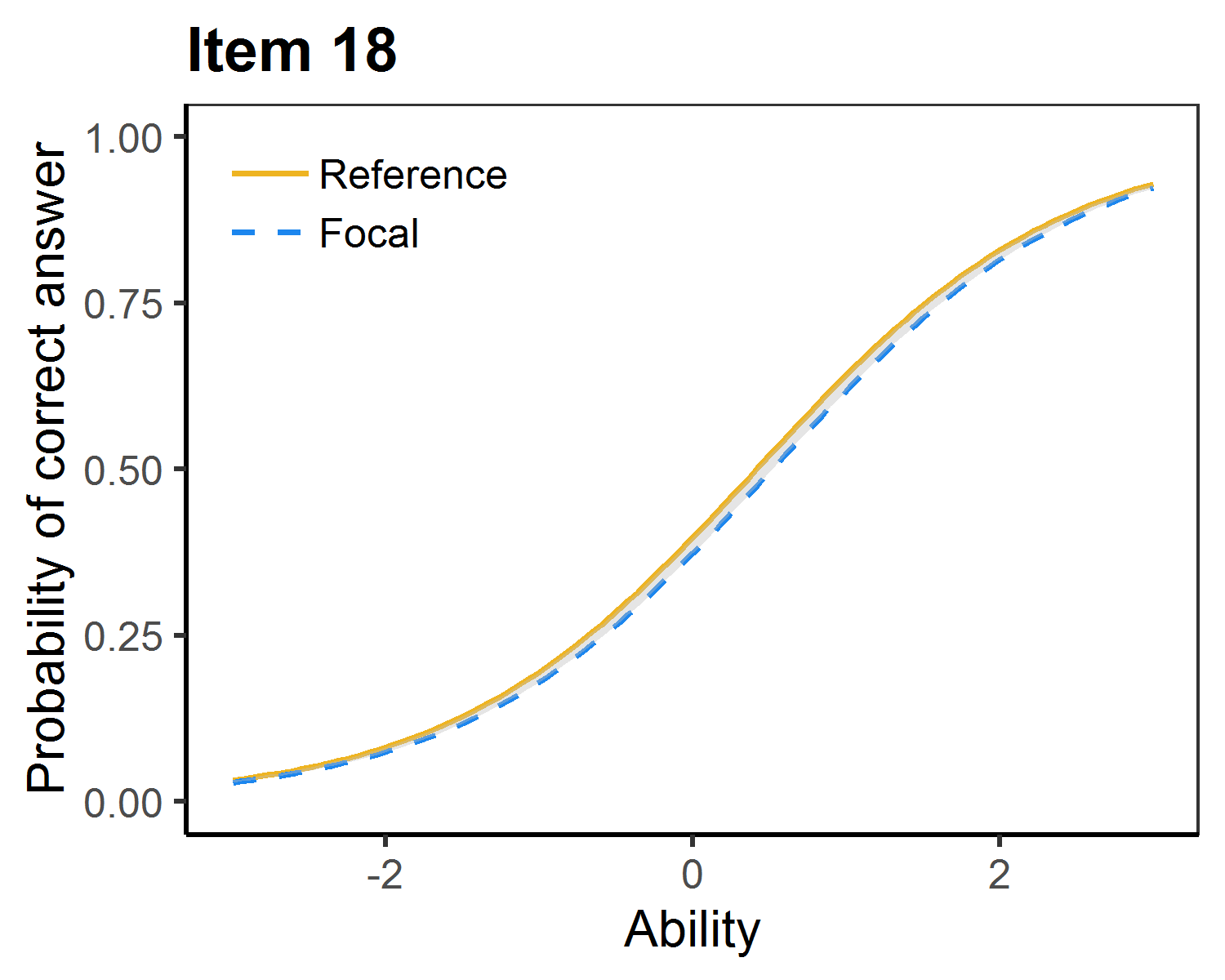
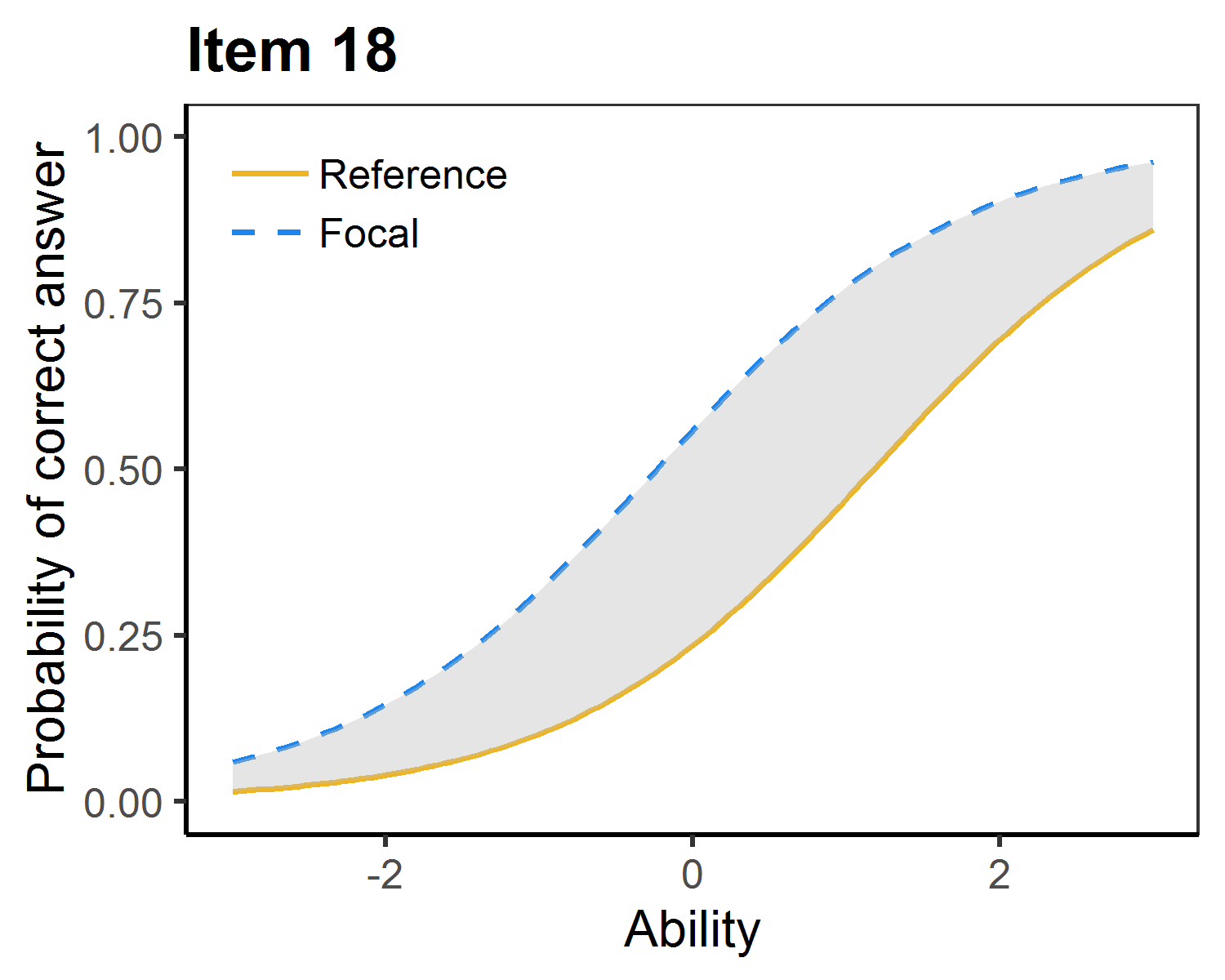
* **Items may exhibit insubstantial DIF for one subgroup and problematic DIF for other subgroups**

Continuing the prior example, examine the DIF depiction in *Figure 4B* on the right. In this case, both graphs depict the same assessment item; however, while the left graph shows DIF based on socioeconomic status, the right graph shows DIF based on ethnicity (Reference = African American students and Focal = non-African American students).

Here, while the question is “fair” when considering socioeconomic status alone, it appears to be biased against African American students when considering ethnicity. A reviewer observing both graphs together should be able to determine that *this item should be subject to further review on the grounds of DIF by ethnicity*.

***Figure 4: Comparing Scope of DIF***

***Figure 4A: Insubstantial DIF Figure 4B: Substantial DIF***

* **Items exhibiting substantial DIF for any group should be marked for further review**

An item exhibiting DIF as substantial as *Figure 4B* above should be subject to further review—this is true even if the item exhibited no ICC parameter problems and no DIF for any other group. This is simply because potential bias against any category of student is sufficient grounds for change, and assessment items cannot change for the better without equitable standards for critical review.

* **Most items will not be as simple as the preceding examples**

While some items will exhibit degrees of DIF similar to the two in *Figure 4*, most cases will fall somewhere in between. The purpose of calling on reviewers at all is to introduce human judgment from multiple perspectives into the process.

This appendix deliberately aims to avoid shaping reviewers’ values or dialog around any item (beyond building a basic understanding of the statistical reasoning behind the item flags). However, it may be useful for reviewers to be aware of the trade-off between the threshold for recommending further review, on one hand, and the time it will take to review each item’s content, on the other. I.e., once an item is recommended for further review, a committee may be convened to examine the phrasing of the question, the answer options available (if multiple choice), and any of a number of other factors; then the question’s content must be carefully revised and reintegrated into the district’s bank of assessment items.

This process is critical to ensure the inclusion of expert and community perspectives and is naturally time-intensive. Reviewers can think of the trade-off in terms of (A) lower threshold for

review 🡨🡪 broader assessment change more slowly vs. (B) higher threshold🡨🡪 narrower assessment change more quickly.

*Note: Figures not cited in this appendix were created by the author of this document using various libraries in the R statistical analysis program.*

1. Specifically, recommendations #109 (leveraging psychometric analysis of test items) and #112 (DREA assistance in developing valid, reliable, and aligned test items) [↑](#footnote-ref-1)
2. Readers unfamiliar with IRT should review Appendix A (Reading Item Characteristic Curves) and Appendix B (Interpreting DIF) to familiarize themselves with key item analysis concepts before continuing. [↑](#footnote-ref-2)