Development and Evaluation of a Composite Item Fit Statistic for Diagnostic Classification Models

Jeffrey C. Hoover, Jennifer L. Kobrin, W. Jake Thompson, and Wenhao Wang

Accessible Teaching, Learning, and Assessment Systems (ATLAS)

University of Kansas

Author Note

Paper presented at the 2022 annual meeting of the National Council on Measurement in Education, San Diego, CA. All authors contributed equally to this study and are listed alphabetically. Correspondence concerning this paper should be addressed to Jennifer Kobrin (Jennifer.kobrin@ku.edu). The authors would like to acknowledge the test development professionals who participated in this research and Amy Clark for her feedback on an earlier draft of this paper.

Development and Evaluation of a Composite Item Fit Statistic for Diagnostic Classification Models

Evaluation of item-model fit is an important part of the test development (TD) process (American Educational Research Association [AERA] et al., 2014). Item fit statistics quantify the difference between the observed performance on an item and the performance that would be expected given the estimated psychometric model (Sinharay & Almond, 2007; Sorrel et al., 2017). Poor item fit indicates that the item-level model does not conform to observed responses. A poorly fitted item-level model and/or an accumulation of poorly fitting items is less likely to produce assessment scores that can be used to make intended inferences and interpretations (e.g., Chen et al., 2013).

After item fit statistics have been calculated, TD professionals play a crucial role in the evaluation of items. For example, after an item is field tested, it is common for TD professionals to review the item statistics and make decisions about whether to promote an item for use on an operational assessment. Crucially, TD professionals use not only the item statistics, but also their own expertise in the subject and professional judgement to inform this decision-making process.

From a psychometric perspective, it is beneficial to provide as many statistics as possible to TD professionals to help inform item promotion decisions. However, providing too many measures of item performance may be overwhelming. It can be difficult to simultaneously evaluate multiple measures of item-fit, particularly for individuals who may not have extensive psychometric training to assist in the interpretation of the statistics. Therefore, it may be beneficial to synthesize different measures of model fit into a single composite statistic that

provides concise information about an item's performance. In this way TD professionals can more easily combine a rich set of empirical data with their own subject matter expertise to inform decisions about item promotion.

In this paper, we describe the development process for one such composite item fit statistic and compare the statistic to independent ratings of item quality from TD professionals. We focus our discussion on items that were made available for this study and were developed for use in a large-scale assessment system that uses a diagnostic classification model (DCM; Rupp et al., 2010; Bradshaw, 2016) for reporting student results.

Item Fit for DCMs

DCMs are confirmatory latent class models, where each class represents a profile of mastered and non-mastered attributes. Although the attributes can be polytomous, most applications of DCMs assume binary latent traits (e.g., master/non-master, proficient/not proficient, etc.). A Q-matrix (Tatsuoka, 1983) is used to map which attributes are measured by each item on the assessment. Because the focus of this paper is the development of a composite item-fit statistic, we limit our discussion to DCMs with binary attributes and a simple Q-matrix design, where each item measures only one attribute.

Previous work in DCMs has examined various methods for evaluating item-model fit (e.g., Sorrel et al., 2017). However, much of this existing research has utilized independent statistical tests to evaluate item-model fit. Such approaches are limited because they often focus on a single aspect of item-model fit. For example, Sorrel et al. (2017) applied the $S - X^2$ statistic (Orlando & Thissen, 2000) to evaluate absolute item-model fit using observed and expected item responses. While the observed and expected item responses are undoubtedly

informative in terms of item-model fit, this is just one aspect of item-model fit. Other aspects may include item difficulty, item difficulty conditioned on attribute mastery status, or the consistency between item-model fit statistics calculated from the observed data and from data simulated using the estimated model parameters. To account for these additional aspects, other item statistics such as item *p*-values and item *p*-values conditioned on attribute mastery status could be calculated, and the incorporation of these additional item-level statistics into a composite statistic may improve the evaluation of item-model fit.

Component Item Fit Statistics

In this study, we calculated five item statistics that were included in the development of the composite fit statistic. These are:

- Item *p*-value: the proportion of students who answered the item correctly.
- Standardized difference: the difference between an item's *p*-value and the average *p*-value for all items measuring the attribute. This statistic assumes that all items measuring an attribute should be approximately equally difficult (i.e., fungible).
- Expected *p*-value: The *p*-value expected by the estimated DCM, based on model parameters.
- Conditional probability for masters: The *p*-value expected by the estimated DCM for students who are predicted to be masters of the assessed attribute, based on model parameters.
- Conditional probability for non-masters: The *p*-value expected by the estimated DCM for students who are predicted to be masters of the assessed attribute, based on model parameters.

The first two statistics are based only on observed data. For this exploratory study, *p*-values were flagged if the value was below .35 or greater than .95. Most items available for this study had three answer options; thus, a *p*-value less than .35 indicates an item where the correct answer was chosen approximately less frequently than would occur with random guessing. Conversely, a *p*-value greater than .95 indicates an item that nearly all students responded to correctly, suggesting other information in the item may be cuing the answer. The standardized difference scores are on a *z*-score scale, and therefore were flagged when values were below -1.96 or greater than 1.96.

The last three item statistics (expected *p*-value, conditional probability for masters, and conditional probability for non-masters) are based on model parameters and are calculated from posterior predictive model checks, as described by Sinharay and Almond (2007). At a high level, the DCM is estimated using Markov Chain Monte Carlo (MCMC; Geyer, 2011). A new simulated data set is then generated from each retained iteration of the Markov chain. For each simulated data set, we calculate summary statistics of interest (e.g., the item statistics). This creates a posterior distribution of the model expected item statistics. Finally, we compare the statistics from the observed data to these posterior distributions, usually through a compatibility interval (e.g., the middle 95% of the posterior distribution). If the observed statistic falls outside of the compatibility interval, the item is flagged. Two compatibility intervals were used in this study. The overall expected *p*-value used a 95% compatibility interval, whereas the conditional *p*-values used an 80% compatibility interval. These flagging criteria are consistent with thresholds that have been used for evaluating the item-level fit of DCMs in previous work (e.g., Thompson, 2019).

These five item statistics evaluate multiple aspects of item-model fit. While the item *p*-values broadly evaluate item difficulty such that items can be flagged if they are overly difficult or easy, the conditional *p*-values allow for evaluating item difficulty given attribute mastery status, which may provide additional information. For example, it is possible that an item might not be flagged based on the item *p*-value if the vast majority of students have mastered the assessed attribute and the conditional probability for masters is appropriate, yet the observed conditional probability of non-masters might be excessively low and cause the item to be flagged. In the same way, the expected *p*-value assesses whether the observed *p*-value is typical, compared to what would be expected from simulating data based on the estimated model parameters. It is possible that the observed item *p*-value may be within the allowable range (i.e., .35 to .95), yet the *p*-value falls outside of the 95% compatibility interval, leading the item to be flagged. The standardized difference also provides unique information pertaining to item-model fit. The standardized difference assesses the comparability of the item *p*-values and the attribute's average item *p*-value. It is possible that the *p*-value for each item assessing an attribute is within the allowable range, but there may be differences between the items in terms of difficulty, which may have implications for model assumptions (e.g., fungibility) and/or assessment fairness if some students receive more difficult items than other students. By incorporating information from each of these five statistics, it is possible to thoroughly evaluate multiple, distinct aspects of item-model fit.

Development of the Composite Item Fit Statistic

Based on the five item statistics described earlier, we developed a composite statistic to concisely synthesize patterns of flags across the individual statistics. In total, the composite

item fit statistic has 12 categories with five severity levels based on the pattern of flags for each item (see Appendix A). The 12 categories of the composite item fit statistic are mutually exclusive. Items, based on their component item fit statistics, can only be categorized into one of the 12 categories. The categories are ordered such that lower categories indicate good fit and higher categories indicate poor fit. Accordingly, each category was also assigned a severity level. In general, a higher severity level means that the item violates the assumptions of the psychometric model to a greater degree than other items. Thus, items with high severity are the most in need of further review. In contrast, a lower severity level indicates little or no misfit, indicating the item fits the psychometric model well and likely needs less review. The individual item statistics and the composite item fit statistic were calculated for each item made available for this study. In total, this included 178 English language arts (ELA) items, 160 mathematics items, and 60 science items. Items represented a range of content across grade levels and item complexity.

To evaluate the consistency of the composite item fit statistic with other ratings of item quality, we investigated the extent to which the composite is consistent with TD professionals' independent item ratings and reviews. Our research questions were:

1. To what extent does the composite item fit statistic correspond to TD item ratings?
2. What factors do TD professionals identify in their item reviews, and to what extent are these captured in the composite item fit statistic?

Methods

Item Rating

To examine how the composite item fit statistic compares to ratings of item quality by subject matter experts, the 398 items for which the composite statistic was computed were sent to TD professionals for independent ratings. That is, the ratings were blind in that the TD professionals did not have access to the item statistics during their review. Their ratings of item quality were based only on their subject matter expertise, without item statistics to potentially bias their independent opinion, or cue certain aspects of the item to examine more closely.

We developed rating instructions and forms for TD professionals to rate the items (see Appendix B). Next, we conducted cognitive labs with two staff and used the findings to revise the rating materials. Then, TD team members were trained in the item rating process and had an opportunity to practice rating items. Following the training we collected item ratings from four TD professionals in ELA, two in mathematics, and four in science. The TD professionals worked in pairs to rate the items and were instructed to discuss the ratings and come to consensus. On the rating forms, the TD raters provided overall ratings of item quality on a five-point scale, corresponding to the five composite severity levels (0-4) and written comments about each item. Subsequently, TD professionals rated the items on a four-point scale (0-3) according to seven factors aligned closely to the flagging criteria used to develop the composite.

After collecting the item ratings, we reviewed the data for logical consistency and flagged a few ratings that needed to be revised. Specifically, in mathematics, five items were rated as both “easy for both master and non-master” and “difficult for both master and non-

master" and two attributes had more than 80% of items rated as easier than the other items. In science, one attribute had more than 80% of items rated as easier than the other items. We asked TD staff to go back to review and revise these ratings before conducting our analyses.

Analyses

We examined the correspondence between the composite item fit statistic and the TD ratings with a Chi Square Test of Association and the percentage of items with exact or adjacent agreement. We also examined the distribution of ratings across and within each composite category. To evaluate the factors TD professionals identified in their reviews and the extent to which those factors are captured in the composite statistic, we analyzed TD's open-ended comments in mathematics on each item using thematic content analysis.[1] We developed a coding protocol capturing the content of the comments. Two authors coded each comment and discussed coding to come to full consensus. We examined the frequency of codes applied to each composite category to determine if there are themes in TD's comments associated with the composite categories and severity levels.

Results

Table 1 shows the number of items from each subject that were placed into each category based on the pattern of flags in the component item statistics.[2] These classifications represented the "true" categories and severity levels that were compared to the independent ratings from TD professionals. Across all subjects, most items fell into the "outperformed expectations" category (*n* = 148, 37% of all items). These are items where one or more

[1] Due to the large number of items missing comments in ELA and science, we did not conduct qualitative analyses in these subjects.

[2] For a description of the composite item statistic categories, see Appendix A.

conditional *p*-values were outside of the compatibility interval, but in an advantageous direction (e.g., masters performed even better than expected). Additionally, very few items (*n* = 42, 10% of all items) were in the highest severity categories.

Table 1

Number of Items in Each Composite Category

Category	Severity	ELA Items	Mathematics Items	Science Items	Total Items
No flag	0	23	26	0	49
Observed *p*-value	1	0	2	0	2
Expected *p*-value	1	11	6	1	18
Overlapping compatibility intervals	1	0	0	0	0
Outperform expectation	1	65	54	29	148
Difficult for master	2	7	7	3	17
Easy for non-master	2	8	5	3	16
Overlapping p-value	3	1	2	1	4
Easy for both master and non-master	3	17	28	15	60
Difficult for both master and non-master	3	17	17	8	42
Reversal	4	4	4	0	8
Non-fungible	4	25	9	0	34

Figure 1 shows the comparison of TD's overall ratings with the composite severity levels. In ELA, the TD raters tended to give low overall ratings, indicating good item quality. In mathematics, the raters tended to give higher overall ratings, indicating poor item quality, but with greater variability in the ratings. In ELA and science, TD staff gave low (i.e., good) ratings to numerous items with the highest severity levels; and in mathematics, TD staff gave high (i.e., poor) ratings to numerous items with low severity levels. The Chi Square Tests of Association between overall rating and severity levels in all subjects were not statistically significant ($p > .05$), indicating that TD ratings and composite severity levels are independent.

Figure 1

Distribution of TD Ratings by Composite Severity Level

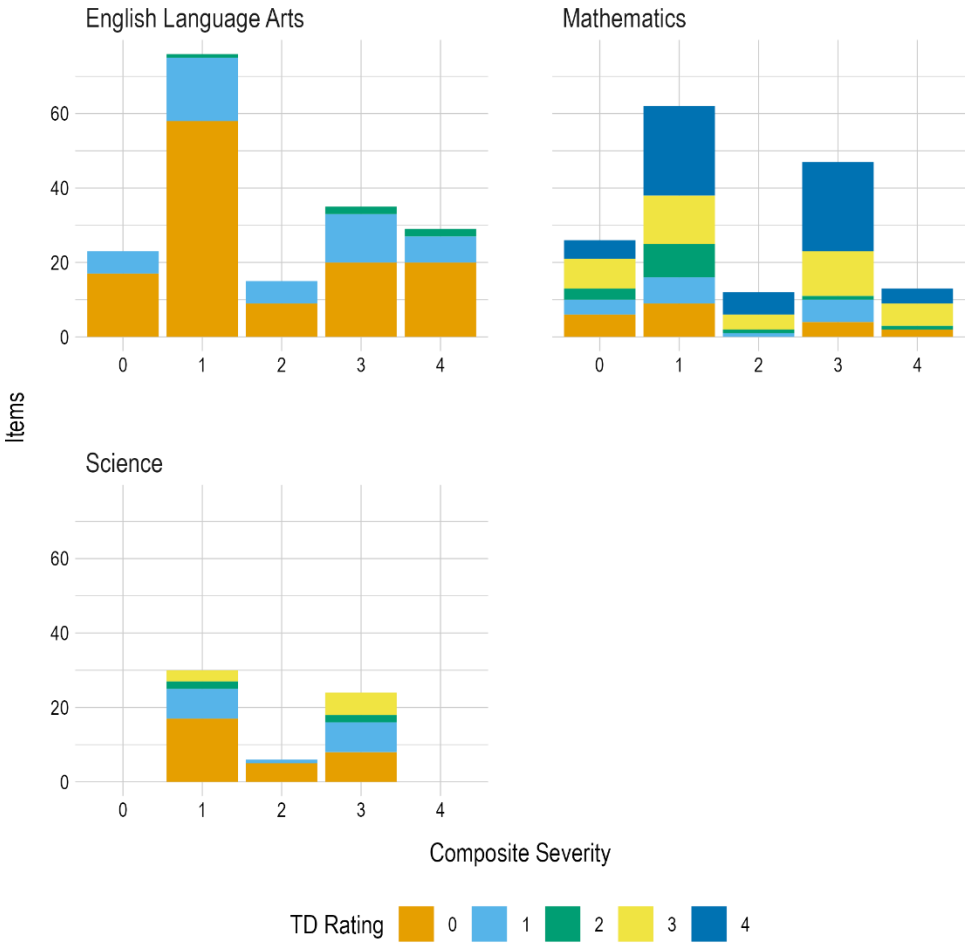

Table 2 shows the percentage of items with exact or adjacent agreement between TD ratings and composite severity levels. In all subjects, about 20% of items showed exact agreement and 40% showed adjacent agreement (plus/minus one rating level). Additionally, all three subjects showed a weak positive association between the composite severity and TD's overall rating, as measured by the polychoric correlation and Cramer's V. These results suggest moderate consistency between the ratings and severity levels, but there is a sizable number of

items with discrepancies as noted earlier, which is reflected in the non-significance of the overall Chi Square Test of Association.

Table 2

Percentage of Items with Agreement Between TD Overall Ratings and Composite Severity Levels

Subject	Number of Items	Exact Agreement (%)	Adjacent (± 1) Level Agreement (%)	Polychoric Correlation	Cramer's V
ELA	178	19	41	.167	.158
Mathematics	160	19	36	.210	.172
Science	60	23	37	.312	.241

In addition to providing overall ratings, the TD raters were also asked to rate each item on seven statements using a 4-point scale ranging from definitely not/not at all to definitely yes/extremely (see the complete statements in Appendix B). Figure 2 shows the average rating on each statement for each of the 12 categories of the composite statistic. In the figure, no bar indicates an average rating of 0 for that statement. Moving outward, the concentric circles represent an average rating of 1, 2, and 3, the maximum rating. Thus, the larger the bar, the more often items in that category received a high rating on each statement.

Overall, the ratings on these statements generally agree with what would be expected. For example, items in the "easy for both master and non-master" category had average ratings between 1 and 2 on all the "easy" statements, and near 0 ratings on the other "difficult for both master and non-master" statements. The reverse was true for items in the "difficult for both master and non-master" category. Additionally, for items in the "outperform expectation" category, almost all statements had near 0 ratings, which is expected for items that are performing better than would be expected from the model.

Additionally, the “discrimination” rating is elevated within almost all categories of items. In the ratings, a value of 0 indicated that TD thought the item could not discriminate well, and 3 indicated a good ability to discriminate. For this analysis, the coding was reversed so that a higher rating indicated a non-desired outcome, just like the other ratings. That all categories of items have a high discrimination score indicates that TD thought most items were unable to discriminate between masters and non-masters. Because almost all items received high ratings on this statement, additional training may be needed to help TD teams evaluate which items may exhibit this problem.

In total, the results from the ratings provide somewhat conflicting results. On the one hand, ratings from the seven specific statements about each item are generally consistent with the 12 categories of items determined by the composite statistic we identified. On the other hand, there is a fair amount of disagreement in the overall ratings and composite statistic severity levels. This indicates that the TD team may have different criteria for what makes an item problematic than captured by the flagging criteria for the item fit composite. Using the results in Figure 1, it is likely that although the ELA and science teams were able to identify some specific issues in the items (e.g., the item was difficult for both masters and non-masters or easy for both masters and non-masters), they still felt that the item was fine overall. Conversely, the mathematics team identified many items that performed well statistically as having some issues making them not “perfect”. Thus, it seems likely that additional criteria are being used by TD, beyond item performance and these criteria may very across content teams. These additional criteria were explored in a qualitative analysis of TD’s item review comments.

Figure 2

Average TD Rating for Each Composite Category

Qualitative Results

Table 3 shows the overall code frequency for TD's item comments in mathematics. TD staff did not identify any factors impacting performance for approximately 17% of the items (n=27). Conversely, one third of the comments (n=53) reflected multiple factors that may have impacted performance on the item. For the majority of items (n=138, 86.3%), TD staff made comments about item content, including item appearance, wording/vocabulary, and/or issues with objects or materials.

Table 3

Item Comments Overall Code Frequency in Mathematics

Code	N	%*
Item Content (includes item appearance, wording/vocabulary, issues with objects or materials and other factors)	138	86.3
Item Difficulty	66	41.2
Response Options/Distractors	70	43.9
Multiple Factors Impacting Item Performance	53	33.1
No Factors Impacting Item Performance	27	16.9
Comment mentions "masters" or "non-masters"	16	10.0

Note. *More than one code could be applied to each item so the percentages do not add to 100%.

Table 4 shows the frequency of "no factors" and "multiple factors" codes by composite severity level and TD overall ratings in mathematics. If TD ratings and the composite were perfectly aligned, we might expect the number/percent of items with "no factors" impacting performance to decrease and the number/percent of items with multiple factors to increase as severity level increases. The results show that of the 27 items coded as having "no factors", the majority had severity levels of 0 and 1. However, seven (26%) of the items coded as having "no

factors" had a severity level of 3. The 53 items coded as having multiple factors were distributed across all five severity levels.

There is a closer correspondence between TD overall ratings and their item comments, as most of the items described as having "no factors" had overall ratings of 0 and 1, and most of the items coded for multiple factors had overall ratings of 3 or 4. Yet, there were still a few items coded for multiple factors that TD rated 0-2. There was a small positive correlation between the number of factors and the overall rating from TD ($\rho = .27, p < .01$), which suggests the number of factors identified by TD roughly corresponded to the overall rating provided by TD.

Table 4

Items Coded as Having "No Factors" and "Multiple Factors" by Composite Severity Level and TD Overall Rating

Rating	No Factors	Multiple Factors
Composite Statistic		
0	7	8
1	13	14
2	0	5
3	7	19
4	0	7
TD Overall Rating		
0	10	1
1	8	2
2	4	4
3	5	14
4	0	32

We examined the codes for the items where the composite statistic had an overall severity of 0 (i.e., no statistical flags) in which the TD raters found "multiple factors". On these items, TD raters identified included factors including item appearance, difficulty (cognitive

load), throw-away distractors, and cuing. Table 5 shows the frequency of TD comment codes by composite category. If TD ratings and the composite were perfectly aligned, we might expect TD comments reflecting "no factors" items to align with category 1 (no flag) and category 5 (outperform expectations); comments referring to "masters" or "non-masters" to align with categories 6 and 7; and comments referring to item difficulty to align with categories 2, 3, 4, 9 and 10. There was generally good correspondence between TD's identification of items with "no factors" influencing performance and the expected categories, but less correspondence in the other categories.

Table 5

Code Frequency by Composite Category

Composite Category	TD Comment Code Applications in Mathematics					
	No factors	Multiple factors	"Master" codes	Content codes	Item difficulty codes	Response option codes
1. No flag	7*	8	4	19	16	9
2. Observed p-value	0	0	0	0	0*	0
3. Expected p-value	1	1	2	4	3*	3
4. Overlapping compatibility intervals	0	0	0*	0	0*	0
5. Outperform expectation	12*	13	2	39	18	24
6. Difficult for master	0	5	2*	9	1	7
7. Easy for non-master	0	0	0*	5	2	3
8. Overlapping p-value	0	1	0*	3	0*	1
9. Easy for both master and non-master	5	10	1	26	12*	8
10. Difficult for both master and non-master	2	8	1	23	8*	4
11. Reversal	0	3	2*	6	2*	2
12. Non-fungible	0	4	2	4	4	9

* These cells indicate the types of comments expected if TD review and the composite categories were perfectly aligned.

Conclusions, Implications and Next Steps

Traditional item review practices generally involve TD staff or external reviewers using item statistics to help guide reviews. However, as these professionals often lack psychometric training, especially training on DCM, they may have difficulty interpreting and synthesizing model-based statistics, especially when multiple item statistics are available. While several studies examined relationships between expert judgment and item difficulty (e.g., Bejar, 1983; Bramley & Wilson, 2016; Kibble & Johnson, 2011), no research has compared TD's independent item reviews to model-based item flagging based on a set of item statistics. Investigating this relationship can shine light on the most salient factors in TD's item review, as well as provide evidence that the composite is functioning as expected.

In this study, TD professionals independently rated items based on their subject matter expertise, and these ratings were compared to a composite fit statistic that incorporate many different aspects of item-level model fit. The findings suggest that while there is some level of agreement between TD ratings and the composite item fit statistic, TD ratings are based on different factors or constructs than those captured by the composite. Thus, the composite may provide useful supplemental information for TD to consider in the item review process. Additionally, the findings show variability across TD content teams in the criteria they use to determine whether an item is good or poor, with some teams more stringent and others more lenient in their overall ratings.

Findings from this work have implications for any assessment program that seeks to integrate TD subject matter expertise with empirical psychometric data during item reviews. From a psychometric perspective, it is beneficial to include as many indicators of item performance as possible. In this way, we can capture many aspects of item-level model fit,

ensuring that items used on an operational assessment perform as expected. However, too many indicators can be overwhelming and difficult to synthesize. In this paper, we demonstrated how multiple indicators of item performance can be combined into a single composite metric that can be used for item review. The composite metric allows TD reviewers to easily identify which items are seen, by the psychometric model, as problematic. The individual component statistics can then be used to dive deeper into specific issues of item performance when needed. Thus, the composite metric approach effectively balances the desire for a rich set of psychometric information with the practical needs of individuals who must process that information.

Future work will continue to work with TD professionals to refine the composite statistics. For the items evaluated in this study, the composite and component item statistics were provided to the TD professionals after the study was complete. The TD professionals were then able to compare their ratings to the empirical data to evaluate where there were differences and formulate subject matter rationale for why differences may exist. Based on the findings of this study and the TD review of the actual statistics, it may be appropriate to refine the composite statistic. Specifically, given some of the discrepancies observed between the composite statistic and TD ratings, it is possible that improvements to the composite statistic may be possible. Additional component statistics may help capture an even wider range of item performance. Alternatively, refining the existing categories and severity levels may also be appropriate. Ultimately we aim to have the composite statistic incorporated into the operational item review workflows to ensure that item promotion decisions are made with fullest set of information possible.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303–310. https://doi.org/10.1177/014662168300700306

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (1st ed., pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Bramley, T., & Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. *Research Matters*, *21*, 48–54. https://www.cambridgeassessment.org.uk/Images/374813-maintaining-test-standards-by-expert-judgement-of-item-difficulty.pdf

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Dynamic Learning Maps Consortium. (2016). *2014–2015 Technical Manual—Integrated Model.* University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf

Geyer, C. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC. https://doi.org/10.1201/b10905-2

Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations. *Advances in Physiology Education*, *35*(4), 396–401. https://doi.org/10.1152/advan.00062.2011

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1)*,* 50-64. https://doi.org/10.1177/01466216000241003

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. *New York: Guilford*.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631. https://doi.org/10.1177/0146621617707510

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report No. 19-01). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://doi.org/10.35542/osf.io/jzqs8

Appendix A: Composite Item Fit Statistic Categories and Severity Levels

Category	Severity	Description
1. No flag	0	The item fits the model well. The item is not flagged by any of five model fit statistics.
2. Observed *p*-value flag	1	The item adequately fits the model; the item is flagged only by observed *p*-value, i.e., the observed *p*-value is either too low or too high.
3. Expected *p*-value flag	1	The item is only flagged by the compatibility interval of the expected *p*-value
4. Overlapping compatibility intervals	1	The item adequately fits the model, but the compatibility intervals for masters and non-masters overlaps. Items could have an observed *p*-value flag or not.
5. Outperform expectation	1	The item has an observed conditional *p*-value for masters higher than expected, an observed conditional *p*-value for non-masters lower than expected, or both. Items in this category could have an expected *p*-value flag or not. Items could have an observed/expected *p*-value flag or not.
6. Difficult for master	2	The item has an observed conditional *p*-value for masters lower than expected. Items can have an observed/expected *p*-value flag or not. Items can have an observed conditional *p*-value for non-masters lower or within the expected range.
7. Easy for non-master	2	The item has an observed conditional *p*-value for non-masters higher than expected. Items can have an observed/expected *p*-value flag or not. Items can have an observed conditional *p*-value for masters higher or within the expected range.
8. Overlapping p-value	3	The item has an observed conditional *p*-value for masters lower than expected and the observed conditional *p*-value for non-masters is higher than expected. However, the observed conditional *p*-value for masters is still higher than the observed conditional *p*-value for non-masters. Items in this category could have an observed/expected *p*-value flag or not.
9. Easy for both master and non-master	3	The item has all three expected model fit statistics higher than the expected range.
10. Difficult for both master and non-master	3	The item has all three expected model fit statistics lower than the expected range.
11. Reversal	4	The item has a conditional probability of a non-master providing a correct response higher than that of a master providing a correct response. In other words, the item is easier for non-masters than masters.
12. Non-fungible	4	The item is flagged based on standardized difference statistic as violating the fungibility assumption.

Appendix B: Item Review Rating Instructions

The following table includes a list of statements about item quality. For rating_1 through rating_7, indicate the extent to which the statement is true using the following scale: 0 = definitely not or not at all, 1 = slightly, 2 = moderately, and 3 = definitely yes or extremely. *Reminder*: Review all stimulus materials for the testlets before doing your ratings, including EECMs, alt text, and all items in the assigned testlets. You may find it helpful to rank the items across the testlets or put them into tiers based on difficulty before doing your ratings.

Column Name	***Statements***	***Response Options***
overall_rating	Provide an overall rating for the item on a 0 to 4 scale (0 means this item does not have any issues and 4 means this item has one or more severe issues and would be removed from the test). In other words, a 0 indicates that the item is the best we could write for the attribute and a 4 indicates that the item should be removed from the test. This rating should consider the item in isolation (i.e., without considering other items for the attribute). For example, just because an item is worse than other items doesn't necessarily mean it's bad. Conversely, an item that is better than all the others isn't necessarily good, it might just be the least bad.	0,1,2,3,4
comments	Please summarize your thoughts about this item. If you detect any issues/problems in this item, please describe those.	[Text]
rating_1	This item is easy and most students will be able to answer it correctly (for example, the stem has a clue making the item very guessable).	0,1,2,3
rating_2	This item is easier than other items measuring the same attribute. While the item is easier than other items from the same attribute, it is not necessarily easy. For example, if most of the items for an attribute are very difficult, then an item that is only moderately difficult would be much easier than the other items, even if the moderately difficult item isn't easy in isolation. In other words, once you look at all items in the spreadsheet for an attribute, how does each item compare to the overall group? Note that not all items should get ratings of 2 or 3. That is, every item can't be easier than the overall group.	0,1,2,3
rating_3	This item is easier than other items for students who have not mastered the attribute (i.e., more non-masters will be able to answer it correctly, compared to other items from the same attribute). A non-master is a student demonstrating the skills less than 50% of the time. While the item is easier for non-masters	0,1,2,3

Column Name	*Statements*	*Response Options*
	than other items from the same attribute, it is not necessarily easy for non-masters in isolation. In other words, if you think about just the subgroup of students who have not mastered the attribute, is this item easier for this subgroup than other items for this attribute? Examples: • An item with 3 answer options, where one of the distractors is clearly wrong. This doesn't help masters, because they still need to have mastered the attribute in order to pick the correct answer. But this could make the item easier for non-masters, who are now guessing between 2 options instead of 3. • The stem cues the correct answer option, so non-masters can provide a correct response without mastering the attribute, which may not be true for other items measuring the same attribute.	
rating_4	This item is difficult and few students will be able to answer it correctly (for example, there is an error in the answer key or there are two correct answers).	0,1,2,3
rating_5	This item is more difficult than other items measuring the same attribute. While the item is more difficult than other items from the same attribute, it is not necessarily difficult. For example, if most of the items for an attribute are very easy, then an item that is only moderately easy would be much more difficult than the other items, even if the moderately easy item isn't difficult in isolation. In other words, once you look at all items in the spreadsheet for an attribute, how does each item compare to the overall group. Note that not all items should get ratings of 2 or 3. That is, every item can't be easier than the overall group.	0,1,2,3
rating_6	This item is more difficult than other items for students who have mastered the attribute (i.e., fewer masters will be able to answer it correctly, compared to other items from the same attribute). A master is a student demonstrating the skills at least 50% of the time. While the item is more difficult for masters than other items from the same attribute, it is not necessarily difficult for masters in isolation.	0,1,2,3

Column Name	*Statements*	*Response Options*
	In other words, if you think about just the subgroup of students who have mastered the attribute, is this item more difficult for this subgroup than other items for this attribute? Examples: • An item has a very tricky/attractive distractor. This doesn't affect non-masters, because they haven't mastered the attribute and therefore may not realize the distractor is tricky. But this would make the item more difficult for masters, who might select the attractive distractor and therefore answer incorrectly at a higher rate than on other items with less attractive distractors. • There is one best answer option (the key), but other options are arguably correct.	
rating_7	This item can discriminate well between students who have and have not mastered the attribute. That is, masters have a much better chance of providing a correct response to the item than non-masters. An item that cannot discriminate well is one that masters and non-masters have the same chance of answering correctly. Examples of well-discriminating items: • An item has a key that is not guessable and distractors that are well aligned to misconceptions for the attribute, such that non-masters almost always answer incorrect, and masters almost always answer correctly. Examples of non-discriminating items: • An item has no correct answer, so both masters and non-masters are randomly choosing between the available options. • An item with 3 answer options has one option that is clearly wrong and one option that is a very tricky distractor. Non-masters can eliminate the incorrect option and guess between the other 2. Masters can also eliminate the obviously incorrect option, but select the tricky distractor a high percentage of the time. This could result in both groups having around a 50% chance of providing a correct response, or masters might be slightly above 50%, depending on how tricky the distractor is.	0,1,2,3