Item Response Demands, Predicting Item Difficulty, and Validity of Inferences from Test Scores

Steve Ferrara, Measured Progress; Jeffrey T. Steedle, ACT; Roger S. Frantz, Questar[1]

In current state accountability testing practice, we develop items to meet (a) specifications such as coverage of target content and process standards; and (b) review criteria such as alignment with content standards, Depth of Knowledge, language simplicity, freedom from bias and sensitive topics, and acceptable ranges of item statistics. These design specifications and review criteria do not fully account for content, cognitive, and linguistic response demands that items place on examinees as they process, understand, and respond to achievement test items. It is likely the case that item writers address these unspecified response demands in unique, non-standardized, and intuitive ways rather than in standardized, conscious ways.

This approach has worked well to enable valid interpretations and uses of educational test scores. However, misalignment between item content, cognitive, and linguistic demands and the knowledge and skill demands specified in achievement level descriptors undermines the inferences we make about what students know and can do, based on their achievement level and corresponding achievement level descriptors. Frameworks for assessment engineering (Luecht, 2013) and other principled approaches to assessment design and development and calls for full alignment to enable engineered cut scores (Ferrara, 2017; Lewis & Cook, 2018) aim to avoid misalignment and faulty inferences and enable automated item generation.

To improve understanding of what makes items easy or difficult, we summarize findings from prior item difficulty modeling studies, and we identify item response demand features that predict item difficulty in three state or national assessment programs. We then demonstrate the application of that knowledge by illustrating how to use those features to target achievement levels so that item response demands and achievement level descriptors are aligned. Results from this study provide practical

[1] The authors thank Amy Kinsman, formerly of Pearson, and Amy Reilly of Pearson for their contributions to studies 1, 2, and 3; the managers of three state assessment programs who provided data and items for these studies; and Kristin Morrison of ACT who pointed us to many of the background studies reviewed here.

guidance for item developers that have the potential to improve the accuracy of inferences drawn about student achievement from achievement levels.

What are Item Response Demands? Why Do We Care about Them?

Item response demands are the content area, cognitive, and linguistic knowledge and skills in test items that examinees must recognize, understand, and process when they respond to test items. These demands are operationalized as item features, such as prompts in item stems on how to respond or select from among response options. Item response demands may represent proxies for the cognitive processing that examinees activate during testing, as indicated in cognitive laboratory studies (e.g., Ferrara et al., 2004) and through the item difficulty studies we summarize in this paper.

Understanding item response demands—and their predictive relationship with item difficulty—provides several practical benefits. It enables us to (a) assemble test forms in which content, cognitive, and linguistic item response demands are consistent with the knowledge and skills in achievement level descriptors, thus clarifying interpretations of test scores; (b) develop items for specified ranges of a test's score scale where item availability is sparse; and (c) train item writers to develop items targeted to ranges of a test's score scale with more accuracy than is achieved currently.

For almost 40 years, measurement researchers and test developers have conducted item difficulty modeling research. In these studies, researchers and content experts identify hypotheses about item response demand features that they expect to determine and predict item difficulty. They code the items for these features and treat them as independent variables in statistical analyses (primarily regression approaches) to predict item difficulties as a means of empirically validating item response demand features that accurately predict item difficulty and eliminating those that do not. As we will see in the literature review below, prediction accuracy has varied widely, with R-squares as low as .10 to as high as .90 across studies that examined a variety of measures such as graduate admissions tests, international assessments, and state accountability tests.

Previous Item Difficulty Modeling Research

Other researchers have undertaken studies related to item response demands that predict item difficulty (and other item statistics and parameters). They have investigated item types from a range of assessments that target different domains and examinee age groups; used three different statistical prediction methods; selected a range of item design, content, cognitive, and linguistic response demands to predict item difficulty; found a wide array of significant item response demand predictors; and achieved a wide range of explained variance. Huff (2003) summarized results from four item difficulty modeling studies, two which overlap with the studies we review below. These studies focused on reading and listening comprehension, used OLS and classification tree based regression, and found R-squares of .35, .35, .51, .58, and .87. Table 1 below summarizes the studies we have reviewed.

As Table 1 indicates, these 24 studies focus primarily on (a) reading, literacy, and verbal reasoning; and (b) mathematics and quantitative reasoning. Two studies examine science items, one an insurance certification test. The studies include the Program for International Student Assessment (PISA; 4 studies), Graduate Record Examination (4 studies), state assessment programs (4 studies), adult literacy surveys (3 studies), and one study each on a variety of other assessments. Researchers have relied primarily on ordinary least squares regression (15 studies), but also have used classification and regression tree analysis (CART) and the log-linear test model. These studies have achieved R-squares in reading, literacy, and verbal reasoning of .11 to .94 in predicting p-values and .17 to .89 in predicting IRT b-values, threshold values, or response probabilities (i.e., RP-values). In mathematics and quantitative reasoning tests, observed R-squares for predicting p-values range from .03 to .62 and .36 to .90 for IRT difficulty indicators. In the two science studies, R-squares were .11, .13, and .23; in the certification test, the training sample R-square was .38 and .00 in the cross-validation sample.

Insert Table 1 about here

These studies tend to rely on specifically selected predictors of item difficulties, probably because they are closely aligned with item designs, stimuli, and content, cognitive, and linguistic response demands of particular interest to the researchers or relevance to the test. Thirteen of the 24 studies report R-squares greater than .50, and another nine with R-squares greater than .20.

Table 2 summarizes R-squares from the studies in the literature review. The response demands are organized in five categories: Item Design Demands, Stimulus Demands, and the Content, Cognitive, and Linguistic Demands that we focus on in our studies. We added the Item Design and Stimulus Demands categories because they were the focus of so many of these earlier studies. It is noteworthy that several of the studies in the Item Design demands category focus on item stems and distractors.

Linguistic response demands also were focal points in many of these studies, as was the readability of reading passages.

Insert Table 2 about here

Content area response demands (e.g., main idea in reading, comparisons in mathematics) appear in only five studies, cognitive response demands in nine studies. These two areas represent an opportunity to contribute new findings to the item difficulty modeling literature.

The wide array of response demand variables and wide range of R-squares in these studies suggest that research in item response demands is identifying statistically significant explanatory variables for predicting item difficulty. Several of the studies explain less than 50% of the variance in item difficulties and sometimes as little as 20%. These studies suggest both that this empirical literature is promising and that there is plenty of opportunity for improvements in theoretical development and empirical results.

Method

Data Sources

We report results from three studies. For each study, a state or national assessment program provided items and stimuli (in PDF format) and accompanying metadata and item statistics. The first study analyzed data from high school achievement tests in four content areas: language arts, mathematics, science, and social studies. In the second study, CART models were fit to data from state assessments in science and social studies administered in elementary school and middle school. Study 3 examined predictive models for a national achievement test program spanning grades 3 through 11 in English language arts and mathematics. The identity and specific details of those assessment programs must be withheld to maintain anonymity.

Item Response Demand Frameworks and Coding Procedures

In selecting response demands for studies 1 and 2, we applied three selection criteria: (a) relevance to the item formats and content standards targeted in each test; (b) item features that represent response demands that can be manipulated to target item difficulty and align items with targeted ALDs; and (c) construct relevance. We chose item design, content, cognitive, and linguistic item response demands with empirical support from previous studies (e.g., Ferrara et al., 2011) and five additional hypothesized demands:

- **Item design**: Item Type, Maximum Points
- **Content**: Standard/Objective, Indicator
- **Cognitive:** Depth of Knowledge, Reading Load, Question Type, Relational Complexity (new), Visualization/Graphic (new)
- **Linguistic:** Prepositional Phrases, Grammatical Density (new), Tier 2 Vocabulary (new), Vocabulary Density (new)

Definitions of these demand categories appear in Appendix A. We recruited raters from pools of professional item writers and constructed response scorers. Training content and activities included definitions of each demand coding category, demonstrations of coding items, procedures for coding items independently, procedures for resolving discrepancies between independent coding decisions, and ***practice in coding item response demands. The study 2 rater agreement rates prior to consensus discussions, which reflect agreement rates in study 1, appear in Appendix B.

Assessment program leaders selected the response demands that we included in study 3. These response demands are generated primarily from the Common Core State Standards, which were the knowledge and skill targets for this assessment program. The response demands analyses were part of a larger study that included a focus on validating weighted composite item cognitive complexity measures and surveys and focus groups with item writers.

- **ELA/Literacy**: Text Complexity, Command of Textual Evidence
- **Mathematics**: Mathematical Content, Mathematical Practices, Stimulus Material
- **Both content areas**: Response Mode, Processing Demands

Definitions of these demand categories also appear in Appendix A. Item writers from this program coded the items included in this study. Training content and activities included definitions of each demand coding category, demonstrations of coding items, procedures for coding items independently, and procedures for resolving ambiguities about coding decisions.

Classification and Regression Tree (CART) Analysis

For the series of studies reported here, we applied classification and regression tree analysis (CART), which is a multivariate statistical modeling approach used to generate binary decision trees (Breiman, Friedman, Olshen, & Stone, 1984). That is, rather than making predictions based on regression coefficients, CART “grows” a decision tree that makes predictions based on independent variables’ values. For example, a decision tree might first distinguish between item types (e.g., multiple choice versus constructed response), and each of those groups might be further subdivided based on other variables (e.g., alignment to a particular content standard, depth of knowledge, linguistic demands, etc.). The terminal nodes of the tree include some number of items with several shared features and similar item difficulty. The mean of their item difficulties is the predicted value for a new item with the same feature set.

CART is especially useful with large numbers of predictor variables because it automatically performs variable selection and identifies important interactions between predictors. It also provides measures of the relative importance of variables as predictors, even when predictors are highly correlated. This is achieved by examining how well each predictor would work as a surrogate for the actual variable chosen to split a given node in two. The importance statistics typically are scaled to have a maximum of 100. Other advantages include that it is nonparametric—that is, it makes no distributional assumptions and it does not require pre-specifying a statistical model—and it easily handles noisy data, outliers, and missing data. In this study, we fit conditional trees (Hothorn, Hornik, & Zeileis, 2012) to the data to correct for bias in variable selection due to categorical variables with many values. In addition, we applied the random forest approach (Breiman, 2001), which is a bootstrap technique wherein many trees (1,000 in this study) are grown using random samples of items and random samples of predictor variables.

In the Results section, we report the conditional random forest R-squares because the random forest approach provides unbiased evaluation of predictive accuracy (in contrast to many prior IDM studies that did not use cross validation). Cross validation is built into the conditional random forest R-squared values because they reflect the accuracy of the “out-of-bag” (OOB) predictions. With random forests, we fit 1,000 different trees with 1,000 different random samples of predictors and items. The OOB items are those not used to fit a given tree (like holding out a cross-validation sample), so we use them to evaluate the model in an unbiased way. Also, those R-squared values come from the same analyses that generated the importance statistics, supporting many of our conclusions.

Results

Importance statistics are scaled to have a maximum of 100, so there is always a predictor with importance of 100, even if it is a poor predictor as indicated by low R-square. To ensure that results have meaningful interpretations, we report regression tree results from predicting item p-values for grades and content areas with R-squares greater than or equal to 0.10, and CART importance statistics greater than or equal to 20. We have chosen these criteria because, as is evident in Table 1 and in our results, item difficulty modeling studies for state achievement tests often produce low R-squares. Tables 3 and 4 display only those importance statistics from our analyses that meet these criteria.

Importance Statistics for Studies 1 and 2

The upper panel of Table 3 displays R-squares for all response demands, including Item Type and Maximum Points, and their importance statistics. As is evident in Table 3, the content response demands Item Type and Maximum Points per item are the most important predictors of item difficulty (see the Item Design Demands panel); they overwhelm the relative importance of the other predictors. This result reflects the finding that dichotomous and selected response items (including some TEIs) generally are easier than polytomous and constructed response items. The lower panel of Table 3 provides importance statistics and R-square values after excluding Item Type and Maximum Points from the analyses. These exclusions enable us to examine the contribution of other predictors to R-squares. Moreover, this enables a clearer look at the relative importance of content, cognitive, and linguistic demands because their importance statistics are higher than they would have been with Item Type and Maximum Points included. In subsequent discussion, we address only results from the lower panel of Table 3.

Insert Table 3 (studies 1 and 2 summary) about here

In study 1, we used item metadata provided by the testing program and coded items for Cognitive Demands and Number of Prepositional Phrases as the Linguistic Demand. In the lower panel of Table 3, we can interpret results from the language arts and social studies tests, where R-squares were

.44 and .18, respectively. Unlike mathematics and science, these subject area tests include polytomous, constructed response items. Thus, results may reflect the influence of Item Type and Maximum Points through their correlation with other predictors. The Content Demands, Standard/Objective, and Indicator content response demands are the most important predictors of item difficulty in these two content areas. Question Type also is a relatively important predictor in social studies, consistent with results in Ferrara et al. (2011).

In study 2, we used item metadata provided by the testing program and coded items for cognitive demands. We used automated grammar parsing and vocabulary analysis to code items to indicate the density of tier 2/3 vocabulary, dependent clauses, prepositional phrases, complex noun phrases, complex verb phrases, and passive voice. In the study 2 section of the table, we can interpret results from the elementary grade tests, grade 4 social studies and grade 5 science tests, where the R-square values were .19 and .13, respectively. Here, in contrast to the study 1 results, Cognitive Demands played a more important role in predicting item difficulty. Question Type is the most important demand in grade 5 science while Relational Complexity is the most important in grade 4 social studies. Question Type and Depth of Knowledge also are relatively important predictors in grade 4 social studies, consistent with Ferrara et al. (2011), as is Grammatical Density. Vocabulary Density (Tier 2 and Tier 3 words per sentence), Dependent Clauses per sentence, Objective, Depth of Knowledge, Visualization/Graphic, Grammatical Density, and Tier 2 Vocabulary also are relatively important in grade 5 science.

Importance Statistics for Study 3

Table 4 displays importance statistics for study 3. As we described above, this study includes response demands that are unique to the focuses and goals of this assessment program. The design of this assessment program focuses on reading selections and innovative item designs not seen in studies 1 and 2. The subsections and response demands in Table 3 reflect the uniqueness of this assessment program. As is evident in Table 3, response demands that are important predictors of item difficulty vary considerably across grade levels/content areas in both reading and mathematics.

English Language Arts. Total R-squares range across grades from .10 to .37 (excluding grades 3 and 10, which are below the .10 criterion). They are highest in grades 6 and 7. Some observations about the importance statistics:

- Several Item Design demands (e.g., Number of Score Categories, Item Type) are among the most important response demands in grades 5–7. These results resemble the dominant role of Item Type and Maximum Points demands in studies 1 and 2.
- Several Reading Selection Demands reflect the importance of Text Complexity as a predictor of item difficulty, especially in grades 6–8 and 11. Text complexity is an emphasis in the Common Core Standards and the design of this assessment.
- Several Content Demands are moderately to highly important (except in grades 9 and 10), resembling results in studies 1 and 2 where Standard/Objective and Indicator demands were the most important predictors in some content areas.
- The Cognitive Demand, Command of Textual Evidence, is relatively important only in grades 6 and 11.
- Processing Demands, an amalgam of Linguistic Demands and Reading Load, are relatively important in grades 4, 6, 7, and 9.

Insert Table 4 about here

Mathematics. Total R-squares range across grades from .33 to .50, with a mean of .37. We can interpret results in all of grades 3–8 plus the three high school mathematics content area tests. Some observations about the importance statistics:

- Item Design Demands, particularly TEI Type in grades 5 through high school and Task Model 1 in grades 3, 6, and 8 are by far the most important predictors of item difficulty.
- Stimulus Demands are not important predictors in any grade or content areas test.
- Several Content demands play an important role, especially in grades 3 and 4. Evidence Statement 1 is an important predictor in grades 3–8 and in the two Algebra tests.
- Cognitive and Linguistic Demands are not important predictors in any grades or content areas.

Overall, for both the English Language Arts and Mathematics tests, the importance of Item, Reading Selection, and Content Demands overwhelm the Cognitive and Linguistic Demands that were the focus of studies 1 and 2. These important predictors are primarily item metadata indicators coded by item writers. This may occur because of the emphasis in this assessment program on text complexity, the Common Core Standards, and innovative item designs. The hypothesized response demands selected specifically for this program—Text Complexity, Command of Textual Evidence in English Language Arts; Stimulus Material in Mathematics; and Response Mode and Processing Demands in both content areas—play no important role in explaining item difficulty. The hypothesized demands, Mathematical Content and Mathematical Processes, play a small role in grades 3 and 4.

Using Item Response Demands to Guide Item Writing

A primary reason for conducting item difficulty modeling research is to find item features that can be specified to achieve accurate item difficulty targeting. The next empirical question is *In what ways can we use these findings to revise items to re-target them?* We use items and a reading passage to illustrate a way to respond to that question. The items below are developed exclusively for this illustration. The passage excerpt comes from an award winning 2001 young adult novel, *A Single Shard* (see https://en.wikipedia.org/wiki/A_Single_Shard).

The four multiple choice items below focus on paragraphs 10 and 11 of the excerpt. The items are designed to ask essentially the same thing, namely, *What does the reader learn about the character, Tree-ear, from these two paragraphs?* The items are differentiated by their linguistic demands and, correspondingly, their level of abstractness. Item 1 is the starting point; items 2—4 are derived from item 1.

1. Which word **best** describes Tree-ear's behavior in paragraphs 10 and 11?

   A bored

   B curious *

   C surprised

   D uncertain

2. Which words **best** describe Tree-ear's behavior in paragraphs 10 and 11?

   A bored but patient

   B curious and observant *

C surprised but quiet

D confused and uncertain

3. Which sentence **best** states what the reader learns about Tree-ear from paragraphs 10 and 11?

   A He is bored but patient while watching Min.

   B He is curious and observant of Min's actions. *

   C He is surprised but quiet while watching Min.

   D He is confused and uncertain about Min's actions.

4. What do paragraphs 10 and 11 **most** reveal about Tree-ear's character?

   A He is patient in new situations.

   B He is observant of details. *

   C He has a quiet nature.

   D He lacks confidence.

The question stems in the first two items are relatively concrete, asking for the word or words that describe Tree-ear's behavior in the two paragraphs. The words used are basic, at grade-level 4 or below, according to *EDL Core Vocabularies* (Taylor, Nieroroda, & Birsner, 1997). The second item is the more difficult of the two because the options require more processing: each includes three words and, thus, three ideas.[2] Most of the words in the options of both items are basic, with "observant" in the second item being the highest level word.

[2] If each pair of adjectives in these options were joined by the conjunction *and*, each option might be coded as containing only two ideas, represented by the adjectives. The use of the conjunction *but* in two of the options and *and* in the other two options indicates a difference in the relationship between the adjectives, and this is why we argue that each option in the second item should be coded as containing three ideas.

The question stems in the third and fourth items are more abstract, asking about Tree-ear's character. They ask that examinees deduce something about Tree-ear's character, based on observable action and his thoughts. This is an appropriate question for a grade 6 assessment, probably more appropriate at grade 6 than asking for words that describe behavior. Based on a simple count of a linguistic feature like prepositional phrases, the question stem in the third item would be coded as more linguistically complex than the stem in the fourth item, and its empirical difficulty would be expected to be higher. However, the question in the fourth item is arguably more abstract, because of the use of the verb "reveal," which actually assumes the prepositional phrase "to the reader"—that is, "What do paragraphs 10 and 11 reveal to the reader about . . ."

Although each option in the third item is more complex—in terms of number of ideas and the language used to present those ideas—than the options in the fourth item, the options in the third item are concrete descriptions, while the options in the fourth item are higher level abstractions. This points to a problem in simply counting number of ideas and certain features of linguistic complexity as part of a determination about response demand and difficulty level. The abstract nature of the options in the fourth item arguably make this item the most difficult.

Discussion and Conclusions

Identifying Study Results that May be Useful in Item Writer Training, Item Development, and Forms Assembly

The R-squares from the studies in the literature and our three studies range widely, suggesting that we have much to investigate and learn about the relationship between item features that are response demands and item difficulty. In studies 1 and 2, we were not able to interpret results for two of the four high school achievement tests and two of the four social studies and science performance tasks because their R-squares were less than our threshold, .10. For the same reason, we did not interpret study 3 English language arts results for two of nine grades. (We were able to interpret study 3 mathematics results for all six grades and three high school mathematics tests.) We argue that high importance statistics, even in analyses with low R-squares, provide weak signals that these response demands may be somewhat useful during item development to target item difficulty levels.

The 24 studies in the literature review place specific focus on item stems and distractors plus Linguistic Demands the reading passage readability. Item Type and related predictors and Maximum Points and related predictors are important predictors in almost all tests in studies 1–3. Item Type, which is often correlated with other variables (e.g., Number of Points, Depth of Knowledge, Relational Complexity) may reflect the general phenomenon in K–12 achievement tests that multiple choice and other selected response items sometimes are easier than constructed response items, especially extended response essay items. Question Type appears as an important predictor in studies 1 and 2, perhaps a reminder that the cognitive processes we strive to elicit from examinees can make items relatively easier or more difficult to answer.

Quality of the Evidence Supporting Claims about the Relationships between Item Response Demands and Item Difficulty

As we said earlier, many of these studies provide weak signals to follow. Overall, the empirical literature on item difficulty modeling is too diverse to find replications of evidence for specific response demands across studies, which we believe can compensate for predictors with weak variables and validate others that provide stronger signals. These studies are spread thinly over a range of tests that target (a) sharply focused, well defined constructs (e.g., quantitative reasoning) that contain items that focus on narrow facets of constructs (e.g., rate problems in mathematics) and that have been crafted and refined over decades of research; versus (b) broadly defined constructs that contain a range of item types that are developed under challenging time constraints and have not received the research scrutiny that they need. They also represent a diverse array of examinee grade and age levels, learning areas, and intended score interpretations and uses. Of the five studies with R-squares greater than .75, four focus on reading comprehension and literacy (i.e., Calfee et al., 1981; Kirsch, 2001; Kirsch & Mosenthal, 1990; Sheehan & Ginther, 2001), and one on quantitative reasoning (Enright et al., 2002).

Construct Relevance

Item construct validity (e.g., Ferrara, Duncan, Perie, Freed, McGivern, & Chilikuri, 2003) or construct relevance requires consideration in evaluating item response demands and attempting to use item difficulty modeling results to item development. Some item response demands may represent construct irrelevant sources of variance in item difficulties; for example, item formats that may not be construct relevant (e.g., multiple choice) and others that may be (e.g., constructed responses); Maximum Points, which may be an item design artifact rather than an indicator of construct relevant knowledge and skills; and linguistic complexity in items that is an impediment to some examinees (e.g., students with disabilities, English language learners) or even all examinees.

Application of Response Demands Frameworks to Item Development and Test Forms Assembly

Earlier, we demonstrated how item features that represent response demands can be manipulated to target difference levels of difficulty—and without introducing any apparent sources of construct irrelevance. Improving difficulty targeting enables (a) improvements in item-achievement level descriptor alignments and, thus, improves score interpretations; and (b) reductions in sources of construct irrelevance such as unnecessary linguistic complexity. We need to know more about targeting item difficulty because (a) content experts do not predict item difficulty accurately (e.g., Hambleton & Jirka, 2006), (b) item writers do not hit intended difficulty targets accurately (e.g., Ferrara et al., 2011), and (c) item response demands often are misaligned with corresponding achievement level descriptors (e.g., Ferrara, 2017). We can use current and future findings to train item writers, create or revise items to hit achievement level targets as well as content targets, and create aligned test forms.

Hypotheses about Distance from Specific Cognitive Processing, Levels of Generality, and Efficiency of Coding Item Response Demands

Tables 1-4 contain a wide range of item response demands. Some focus on solutions of specific types of problems that require specific types of solutions (e.g., item-text interactions, Freedle & Kostin, 1992; rate and probability problems, Enright et al., 2002). Other response demands are at a higher level of generality (e.g., Question Type in studies 1 and 2, Table 3). Others are even further removed from cognitive processing (e.g., Text Complexity in study 3, Table 4). Response demands that are closer to a specific type of problem and solution are likely less generalizable to items that require different solutions, knowledge, and cognitive processes but may provide closer looks are the knowledge, skills, and processes required to respond to items. Other response demands may promise greater generalization and application to a wide range of items and item response demands but may provide insights that are somewhat removed from response demands of specific items, item types, and item families.

Depth of Knowledge is an example of a response demand category that is widely useful but provides very general, even confusing information about the response demands of specific items. For example, DOK level 2, Skill/Concept, includes widely different skills as classifying, organizing, estimating, making observations, collecting and displaying data, and comparing data (Webb, 2007). Subscore reporting categories are an example of further distance from inferable response demands and an attempt at a higher level of generality. Typical subscore categories include literary analysis in reading tests and mathematical reasoning and problem solving.

References

Alderson, J. C., de Jong, J., Kirsch, I., Lafontaine, D., Lumley, T., Mendelovits, J., & Searle, D. (2009, November). How can we predict the difficulty of PISA reading items? The process of describing item difficulty. Presentation at the Language Testing Forum, University of Bedfordshire, Bedfordshire, England.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth.

Cai, L., Baker, E., Choi, K., & Buschang, R. (2014, April). CRESST functional validity model: Deriving formative and summative information from Common Core assessments. Paper presented at the annual meeting of the American Educational Research Association, Philiadelphia, PA.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16(4), 486–514. doi:10.2307/747313

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J-A. (2016). Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. The Curriculum Journal. DOI: 10.1080/09585176.2016.1232201

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11(2), 175–193. doi:10.1177/014662168701100207

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. Applied Measurement in Education, 15(1), 49–74. doi:10.1207/S15324818AME1501_04

Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (2003). Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment. Paper presented in S. Ferrara (Chair), Cognitive and other influences on responding to science test items: What is and what can be, a symposium conducted at the annual meeting of the American Educational Research Association, Chicago IL.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. Educational Measurement: Issues and Practice, 30(4), 3–15. doi:10.1111/j.1745-3992.2011.00218.x

Freedle, R., & Kostin, I. (1992). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: Main ideas, inferences, and explicit statements. GRE Board Professional Report No. 87-10P and ETS Research Report 91-59.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. Applied Psychological Measurement, 30(5), 394–411. doi:10.1177/0146621606288554

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 399-420). Mahwah, NJ: Lawrence Erlbaum Associates.

Hothorn, T., Hornik, K., & Zeileis, A. (2012). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

Kirsch, I. (2001). The International Adult Literacy Survey (IALS): Understanding what was measured. (ETS RR-01-25). Princeton, NJ: Educational Testing Service.

Kirsch, I., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. Reading Research Quarterly, 25(1), 5–30. doi:10.1002/j.2330-8516.1988.tb00318.x

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: example of PISA science items. International Journal of Science Education, 39(4), 468–487. doi:10.1080/09500693.2017.1294784

Lewis, D., & Cook, R. (2018 June). The efficacy of engineered cut scores: Embedding standard setting in principled assessment design. Presentation in S. Ferrara (Organizer), Principled approaches to standard setting: Benchmarked and engineered cut scores, presentation in the annual National Conference on Student Assessment, San Diego, CA.

Luecht, R. M. (2013). Assessment Engineering task model maps, task models and templates as a new way to develop and implement test specification. *Journal of Applied Testing Technology*, *14*. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/45254

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012, April). A framework for predicting item difficulty in reading tests. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC. Retrieved from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1004&context=pisa

McLeod, J., Butterbaugh, D., Masters, J., & Schaper, E. (2015, April). Predicting item difficulty by analysis of language features. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Morrison, K. M., & Embretson, S. E. (2014). Abstract: Using cognitive complexity to measure the psychometric properties of mathematics assessment items. Multivariate Behavioral Research, 49(3), 292–293. doi:10.1080/00273171.2014.912922

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. American Educational Research Journal, 35(2), 269–307. doi:10.2307/1163425

Rowe, M., Ozuru, Y., & McNamara, D. (2006). An analysis of standardized reading ability tests: What do questions actually measure? In ICLS 2006 - International Conference of the Learning Sciences, Proceedings (Vol. 2, pp. 627-633).

Sano, M. (2016 April). Improvements in automated capturing of psych-linguistic features in reading assessment text. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. Cognition and Instruction, 14(3), 285–343. doi:10.1207/s1532690xci1403_2

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. Educational Assessment, 11(2), 105–126. doi:10.1207/s15326977ea1102_2

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. Journal of Educational Measurement, 34(4), 333–352. doi:10.1111/j.1745-3984.1997.tb00522.x

Sheehan, K. M., & Ginther, A. (2001, April). What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sheehan, K. M., & Mislevy, R. J. (1994). A tree-based analysis of items from an assessment of basic mathematics skills. (ETS RR-94-14). Princeton, NJ: Educational Testing Service.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, Jossey-Bass.

Taylor, S. E., Nieroroda, B. W., & Birsner, E. P. (1997). EDL core vocabularies in reading, mathematics, science, and social studies. Austin, TX: Steck-Vaughn.

Turner, R. (2012 April). Some drivers of test item difficulty in mathematics. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC. Retrieved March 26, 2018 from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1003&context=pisa

Webb, N. (2007). Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education, 20(1), 7–25.

Appendix A

Definitions of Response Demands Codes

Response Demand	Definition and Rating
Item Design Demands	
Definition: Features of an item that are related to the complexity of understanding, processing, and formulating a response to a test item (e.g., Sudman, Bradburn, & Schwarz, 1996, Figure 1)	
Item Type	Selected response, constructed response, other (studies 1 and 2)
Maximum Points	0, 1 for selected response items; 0, 1, 2 etc. for constructed response items (studies 1 and 2)
Response Mode	How an examinee is required to respond to an item; in general, selecting a response is less demanding than constructing a response: low, medium, high (study 3)
Item Stimulus Demands	
Definition: Textual, tabular, graphical and other material that provides information with an item and requires examinees to process and understand the information in order to respond	
Text Complexity	Qualitative and quantitative score assigned to each text in separate process prior to item writing and cognitive complexity scoring: Readily Accessible, Moderately Complex, Very Complex (study 3)
Command of Textual Evidence	Amount of text examinees and location and explicitness of information in one or more texts must process in order to respond correctly to an item: low medium, high (study 3)
Response mode	How an examinee is required to respond to an item; in general, selecting a response is less demanding than constructing a response and closeness of response options: low, medium, high (study 3)
Stimulus Material	Numbers of pieces of stimulus material, role of mathematic tools: low, medium, high (study 3)
Content Demands	
Definition: Content area knowledge and skills, defined in content standards or job analyses, required to understand, process, and respond to items; content area declarative knowledge	
Standard/Objective	The content standard(s) targeted by an item
Indicator	A specific, more fined grained member of a group that comprises a content standard

Response Demand	Definition and Rating
Mathematical Content	Relative to the typical mathematical knowledge expectations at the grade level, the extent to which an item or task requires the content to be accessed and applied: low, medium, high (study 3)
Mathematical Practices	What the student is asked to do with the mathematical content , relative to the four sub-components below: low, medium, high (study 3)

Cognitive Demands

Definition: General and content area specific procedural knowledge required to understand, process, and respond to items; content area declarative knowledge

Depth of Knowledge	One of four levels of item cognitive complexity: recall, skill/concept, strategic thinking, extended thinking; see Webb (2007)
Reading Load	Amount and complexity of the textual and visual information provided with an item that an examinee must process and understand in order to respond to an item (Ferrara et al., 2011)
Question Type	Cognitive process/skill required of examinees to respond to an item, using content area knowledge and skills and often but not always posed as a question (Ferrara et al., 2011)
Relational Complexity	The number of (a) concepts that examinees must hold in mind, (b) facts that examinees must hold in mind, or (c) cognitive processes that examinees must undertake in order to process and respond to an item and their relationships to one another
Visualization/Graphic	Items that require or enable examinees to use internal visualization and/or external graphical representations (e.g., sketches, charts, graphs) to process information given in or otherwise relevant to the item

Linguistic Demands

Definition: Grammatical, syntactical, and lexical elements of test items that must be processed in order to process, understand, and respond.

Number of Prepositional Phrases	Number of prepositions counted in item stems, response options, stimuli, and other directions (studies 1 and 2)
Grammatical Density	(Dependent clauses + prepositional phrases + verb phrases) per sentence
Tier 2 Vocabulary	Tier 2 words per sentence; these are general academic words; see Appendix A of the Common Core State Standards at http://www.corestandards.org/assets/Appendix_A.pdf
Vocabulary Density	(Tier 2 + Tier 3) words per sentence; Tier 3 words are domain specific words; see Appendix A of the Common Core State Standards at http://www.corestandards.org/assets/Appendix_A.pdf

Response Demand	Definition and Rating
Processing Demands	Linguistic demands (i.e., vocabulary, grammatical complexity) and reading load (from above) in item stems, item directions, and response options: low, medium, high (study 3)

Appendix B

Rater Agreement Rates for Study 2

Grade 4 Social Studies (137 items)

Reading Load (85.4%), Visualization/Graphic (88.3%), Relational Complexity (59.1%), Primary Question Type (65.0%)

Grade 7 Social Studies (146 items)

Reading Load (83.6%), Visualization/Graphic (83.6%), Relational Complexity (69.9%), Primary Question Type (71.2%)

Grade 5 Science (202 items)

Reading Load (54.0%), Visualization/Graphic (73.3%), Relational Complexity (33.2%), Primary Question Type (35.1%)

Grade 8 Science (226 items)

Reading Load (63.7%), Visualization/Graphic (80.1%), Relational Complexity (25.7%), Primary Question Type (35.0%)

Note: Relational Complexity and Question Type may seem low; however, there are many possible values that an item coder could choose from, as opposed to the 2-3 categories available for other response demands. All final item response demands are resolved in consensus meetings.

Table 1. Item Difficulty Modeling Studies: Variables, Methods, Significant/Important Predictors of Item Difficulty, and Percentages of Explained Variance

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
Drum, Calfee, & Cook (1981)	p-values for multiple-choice paragraph comprehension items from the California Achievement Test	Stepwise OLS regression	Variables related to word translation, word meaning, syntactic/semantic forms, and test format for passages, item stems, correct answers, and distractors	Actual information and % content words in the passage, % content-function words in the stem, ratio of uncommon to common words and % new content words in the correct answer, and plausibility and % new content words for the distractors	.55–.94 (mean = .71) with the 10 best predictors
Embretson & Wetzel (1987)	IRT *b* for multiple-choice paragraph comprehension items from the Armed Services Vocational Aptitude Battery	Linear logistic latent trait model (LLTM)	Propositional analysis variables (arguments plus modifier, predicate, connective, and total propositions) and decision process variables (e.g., confirmability, falsifiability, plausibility of options, and external knowledge requirements)	Modifier propositional density connective propositional density, percent content words, percent relevant text, falsification, confirmation, distractor word frequency, and correct answer reasoning	.17 for propositional analysis variables, .25–.28 for decision process variables, .37 for all variables
Kirsch & Mosenthal (1990)	p-values for open-ended and multiple-choice document literacy items and tasks from the 1985 Young Adult Literacy study	OLS regression	Structure and complexity of documents, task complexity, and solution processes	Number of task organizing categories, number of task specifics, text correspondence between document and task, type of information, and distractor plausibility	.89 (.87 adjusted)
Freedle & Kostin (1992)	p-values converted to equated deltas for GRE reading comprehension items	Stepwise OLS regression	Rhetorical and syntax features; vocabulary; sentence, paragraph, and text length; abstractness of text; location in text of relevant information; subject matter (i.e., humanities, social science, science); rhetorical organization; coherence; text x item interactions (e.g., location of main idea statement, inferences from a single word or multiple locations)	Main idea items: Special references (e.g., “they” for “the girls”), passage sentence length, first paragraph sentence length Inference items: Seven variables, including information location, concreteness of text, negative stems, and length of distractors Explicit statement items: Six variables, including referentials, sentence length, concreteness, rhetorical organization, frontings, and location	.20 for main idea items .49 for inference items .41 for explicit statement items

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
			Item type (i.e., main idea, inference, explicit statements), features of item stem, correct option, distractors		
Sheehan & Mislevy (1994)	3PL item parameters for multiple choice and free response Praxis I mathematics items	CART followed by OLS regression	Item surface features (e.g., equations in item stem), solution process variables (e.g., application of a formula), response type (e.g., multiple choice), and expert judgments of item difficulty	Judgments of item difficulty, making a quantitative comparison, applying a standard algorithm, interpreting a histogram, translating words to symbols, and ordering and matching	.36 for IRT *b*, .12 for IRT *a*, .85 for IRT *c* (all values adjusted, based on stepwise OLS regression)
Sebrechts, Enright, Bennett, & Martin (1996)	p-values for GRE quantitative items	OLS regression	Attributes of the problem statement and problem representation	Need to manipulate multiple variables, problem complexity, and content (money, time, and metric measurements)	.62 for money indicator, .54 for time indicator, .37 for metric indicator
Sheehan (1997)	3PL item parameters for multiple-choice SAT verbal reasoning items	CART	Reading schema (vocabulary in context, main idea and explicit statement, inference about author intent, and application or extrapolation), word usage (standard or poetic/unusual), passage type (complex or simple), and others	Reading schema; contribution of other variables not reported	.20 for reading schema indicator
Mosenthal (1998)	RP80 based on 3PL model for prose-tasks from adult literacy surveys	OLS regression	Readability and three process variables (type of information requested, type of match, and plausibility of distractors)	All significant, readability was a relatively weak predictor	.77
Kirsch (2001)	RP80 based on 3PL model for literacy items from the International Adult Literacy Survey	OLS regression	Text content, continuous text (e.g., narration or argumentation), non-continuous text (e.g., graphics, maps, or forms), and process variables (type of match, type of information requested, distractor plausibility, type of calculation, and operation specificity)	Literacy tasks: Type of match and distractor plausibility	.79-.89 for literacy processing variables

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
Sheehan & Ginther (2001)	IRT difficulty based on 3PL model for TOEFL reading comprehension items related to main ideas	CART	Predictors based on theory of memory activation (eliminating obviously incorrect responses, then activating memory related to the correct response and remaining distractors)	Location of relevant information in the passage (location effects), similarity between correct response wording and text in the passage (correspondence effects), and elaboration on the subject of the correct response or distractors in the text (elaboration effects)	.86
Enright, Morley, & Sheehan (2002)	3PL parameters for variant GRE quantitative items that were created by systematically manipulating item features	CART followed by OLS regression	Need to manipulate variables, problem complexity, and mathematical content (e.g., rate and probability problems)	All significant; certain items with a cost context tended to be easier	For rate problems, .90 for IRT *b*, .50 for IRT *a*, .41 for IRT *c*; for probability problems, .62 for IRT *b*, others non-significant (all values adjusted)
Gorin & Embretson (2006)	IRT *b* for GRE reading comprehension items	OLS regression	Passage and item features such as modifier prepositional density, predicate propositional density, content word frequency, percent of relevant text, and vocabulary level of the distractors	Vocabulary level of correct response, amount of reasoning needed to confirm the correct response, special item format (including Roman numerals), and length of passage	.34 with all predictors, .00 with text encoding variables only (all values adjusted)
Rowe, Ozuru, & McNamara (2006)	p-values for Gates-MacGinitie Reading Tests items	OLS regression	Text features: Word frequency, sentence length, adjacent sentence argument overlap, propositional density Item characteristics: Reasoning required for correct response, confirmability of correct response, number of falsifiable distractors, plus degree of inference required, abstractness of relevant information	Text features: Word frequency and sentence length for expository passages only Item characteristics: None were significant	.11

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
Shaftel, Belton-Kocher, Glasnapp, & Poggio (2006)	p-values for mathematics multiple choice items from a state assessment at grades 4, 7, and 10	OLS regression	Number of words, sentences, and clauses per item, syntactic features (complex verbs, passive voice, and pronoun usage), mathematics vocabulary, and ambiguous terms	Math vocabulary, preposition, ambiguous words, complex verbs, pronouns, and comparatives (e.g., "greater than")	.13 for grade 4, .07 for grade 7, and .40 for grade 10
Alderson, de Jong, Kirsch, Lafontaine, Lumley, Mendelovits, & Searle (2009)	IRT scale locations for PISA reading items	OLS regression	Study 1: Four aspect variables (e.g., pieces of information to be retrieved from text) and four text format variables (e.g., length and complexity) Study 2: 10 item features	Study 1: Not reported Study 2: Familiarity of information needed, structural prominence, competing information, semantic match between task and target information, information concreteness	Study 1: Not reported Study 2: .52 for five strongest aspect variables
Ferrara, Svetina, Skucha, & Davidson (2011)	Item p-values and discriminations for multiple-choice, mathematics items from a grades 3–5 state testing program	OLS regression	Four cognitive response demands and five linguistic response demands	Reading load, question type, number of ambiguous words, number of mathematics terms, number of relative pronouns	.28 for difficulty (all items), .03 for discrimination (all items), .26 for discrimination (study items)
Lumley, Routitsky, Mendelovits, & Ramalingam (2012)	Difficulty of PISA reading items	OLS regression	10 variables related to type and location of information needed to respond correctly, competing information in distractors, and semantic match between the item and passage	Concreteness of information, reference to information outside the text, familiarity of information needed to answer the question, relationship between task and required information, and competing information	.57
Turner (2012)	Difficulty of PISA mathematics items	OLS stepwise regression	0–3 ratings of required competencies: communication; devising strategies; mathematizing; representation; using symbolic, formal, and technical language and operations; and reasoning and argumentation	Reasoning and argumentation, symbols and formalism, problem solving, and communication	.71 (using best subset and stepwise regression)

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
Cai, Baker, Choi, & Buschang (2014)	IRT difficulty parameters for 4th and 8th grade state assessment program English language arts and mathematics items	Combination of IRT with item demand features as covariates	Item features (e.g., requires recall, application of an algorithm), explicitness and relevance of information in the item, content area knowledge and procedural skill, language features	A “dozen or so [item] features” (p. 7)	“About 50%” (p. 7)
Morrison & Embretson (2014)	IRT *b* for middle-school summative test mathematics items	Linear logistic test model (LLTM)	19 response demands related to translation, integration, solution planning, solution execution, and decision processing	All statistically significant	$\triangle$ = .51
McLeod, Butterbauch, Masters, & Schaper (2015)	Rasch item difficulty for an insurance certification test	CART	Content plus linguistic features related to readability, sentence structure, parts of speech, and verb tense (from natural language processing software)	Flesch readability of the stem, Flesch-Kincaid readability of the options, Average word length of the stem, average word length of the key, total syllables of the key, total phrases in the options, average word length of the options Future tense verb frequency of the options, total nouns in the key	.38 (training sample), .00 (cross-validation sample)
El Masri, Ferrara, Foltz, & Baird (2016)	Two-parameter graded response model threshold parameters for selected and short constructed response items for a UK national sample science test of 11-year-olds	Stepwise OLS regression	Curricular variables (i.e., science topic, subtopic, and concept), question type (e.g., apply, infer), depth of knowledge, nature of the stimulus (i.e., text, photo, graph, schematics representation), and language variables (five dimensions generated by *Coh-Metrix* software)	Extended constructed response item, presence of photograph(s)	.23

Study	Outcome/Predicted Variable	Method	Predictors	Significant/Important Predictors of Difficulty	*R*-square
Sano (2016)	Average examinee scale scores of correct responders as proxies for item difficulties for grade 8 1992–2013 NAEP Reading multiple choice items and Reading to Gain Information passages	OLS regression and CART	Twelve psycho-linguistic features extracted from passages and items using an automated natural language processing tool: Overlap of lemmas: stem and distractors, item and passage, location in passage and item, distance between lemmas and item Parts of speech in passages and items, noun chunks in the keyed option Occurrence of "passage" and "author" lemmas in item stems	OLS regression: 12 psych-linguistic features; 4 in the keyed option, 5 in the item stem, 3 in the passage CART: 7 psycho-linguistic features; 3 in item stems, 1 in the response options, 2 in passage-stem-key combinations, 1 in the reading passages	OLS regression: .52 CART: .83
Le Hebel, Montpied, Tiberghien, & Fontainieu (2017)	p-values for PISA scientific literacy items	OLS regression	Depth of knowledge, whether responding correctly depends on provided information, item format, and PISA competency	Depth of knowledge and item format	.11 for depth of knowledge, .13 for question format

CART = classification and regression tree analysis. OLS = ordinary least squares

Table 2. Significant Item Response Demands from the Literature Review

Item Design Demands	
Item/response type, item surface features (e.g., equation in the stem)	Sheehan & Mislevy (1994)
Process variables: literacy (type info requested, type of match); mathematics (operation specificity)	Mosenthal (1998), Kirsch (2001)
Item stems and distractors: reading load, readability, sentence length, negation	Freedle & Kostin (1992), Ferrara et al. (2011), McLeod et al. (2015)
Competing information between passage and distractors, distractor plausibility, content words in correct response and distractors, distractor word frequency/length, vocabulary overlap	Drum et al. (1981), Embretson & Wetzel (1987), Kirsch & Mosenthal (1990), Freedle & Kostin (1992), Mosenthal (1998), Kirsch (2001), Alderson et al. (2009), Lumley et al. (2012), Sano (2016)
Amount of reasoning required to confirm correct response	Gorin & Embretson (2006)
Judgment of item difficulty	Sheehan & Mislevy (1994)

Stimulus Demands	
Readability: concreteness, rhetorical organization (e.g., argument), structural complexity, cohesion elements (e.g., fronting), grammatical complexity (i.e., parts of speech), continuous vs. non-continuous text	Kirsch & Mosenthal (1990), Freedle & Kostin (1992), Mosenthal (1998), Kirsch (2001), McLeod et al. (2015), Sano (2016)
Text content	Kirsch (2001)
Type, location, familiarity, frequency, proximity of information needed (e.g., concreteness of information, information outside the text)	Kirsch & Mosenthal (1990), Freedle & Kostin (1992), Sheehan & Ginther (2001), Rowe et al. (2006), Alderson et al. (2009), Lumley et al. (2012)
Degree of correspondence between item response demands and stimulus, semantic match between item and passage	Kirsch & Mosenthal (1990), Sheehan & Ginther (2001), Alderson et al. (2009), Lumley et al. (2012), Sano (2016)
Presence of photographs	El Masri et al. (2016)
Amount of relevant text, passage length	Embretson & Wetzel (1987), Gorin & Embretson (2006), Rowe et al. (2006)

Content Response Demands	
Reading schema: main idea and explicit statement, inference about author intent, application or extrapolation	Sheehan (1997)
Declarative and procedural knowledge (i.e., or e.g., rate and probability problems) (mathematics)	Enright et al. (2002)
Translation, integration, solution planning, solution execution, decision processing (mathematics)	Morrison & Embretson (2014)
Quantitative comparison	Sheehan & Mislevy (1994)
Applying a standard algorithm (mathematics)	Sheehan & Mislevy (1994)
Interpreting a histogram	Sheehan & Mislevy (1994)
Translating words to symbols	Sheehan & Mislevy (1994)
Money, time, and metric measurements	Sebrechts et al. (1996)

Cognitive Response Demands	
Propositional density, problem complexity, depth of knowledge	Embretson & Wetzel (1987), Sebrechts et al. (1996), Enright et al. (2002), Ferrara et al. (2011), Le Hebel et al. (2017)
Need to manipulate multiple variables (mathematics)	Sebrechts et al. (1996), Enright et al. (2002)
Question/task type (e.g., recall, application of an algorithm), reasoning, problem solving, and argumentation (mathematics), symbols and formalism, communication (mathematics)	Gorin & Embretson (2006), Ferrara et al. (2011), Turner (2012), Cai et al. (2014), El Masri et al. (2016), Le Hebel et al. (2017)
Falsification, confirmation, correct answer reasoning	Embretson & Wetzel (1987)

Linguistic Response Demands	
Syntax, parts of speech	Freedle & Kostin (1992), Shaftel et al. (2006), Ferrara et al. (2011), McLeod et al. (2015), Sano (2016)
Sentence length	Freedle & Kostin (1992)

Vocabulary, vocabulary level, ambiguous words, vocabulary in context, word usage (standard, poetic/unusual), content area terms	Drum et al. (1981), Embretson & Wetzel (1987), Gorin & Embretson (2006), Shaftel et al. (2006), Ferrara et al. (2011), McLeod et al. (2015)

Table 3. Importance Statistics and R-squares for Studies 1 and 2: Empirical Evidence for Item Response Demands

	Study 1				Study 2			
	Lan-guage Arts	Mathe-matics	Science	Social Studies	Grade 4 Social Studies	Grade 7 Social Studies	Grade 5 Science	Grade 8 Science
Item Design Demands[1]								
Item Type[2]	100	100	100	100	97	--	100	--
Maximum Points[2]	NA	NA	NA	NA	100	--	42	--
Conditional Random Forest R-squares (for analyses including all variables in this table)	0.46	0.12	0.21	0.23	0.23	0.09	0.20	0.05
Content Demands								
Standard/Objective	100	--	--	--	--	--	55	--
Indicator	--	--	--	100	NA	NA	NA	NA
Cognitive Demands								
Depth of Knowledge	--	--	--	--	41	--	54	--
Reading Load	--	--	--	--	--	--	--	--
Question Type	--	--	--	43	76	--	100	--
Relational Complexity	NA	NA	NA	NA	100	--	--	--
Visualization/Graphic	NA	NA	NA	NA	--	--	23	--

	Study 1				Study 2			
	Language Arts	Mathematics	Science	Social Studies	Grade 4 Social Studies	Grade 7 Social Studies	Grade 5 Science	Grade 8 Science
Linguistic Demands								
No. of Prepositional Phrases	--	--	--	--	29	--	--	--
Grammatical Density	NA	NA	NA	NA	41	--	34	--
Tier 2 Vocabulary	NA	NA	NA	NA	--	--	33	--
Vocabulary Density	NA	NA	NA	NA	--	--	74	--
Conditional Random Forest R-squares[3] (for content, cognitive, and linguistic demands only)	0.44	0.07	0.08	0.18	0.19	0.07	0.13	0.02

Note. "--" indicates importance statistics less than our selected threshold 20, or R-squares from regression tree analyses with less than our selected threshold, 0.10. See main text for details. "NA" = not applicable for this study. All importance statistics are rounded. Comparing values is appropriate within columns only.

[1] Taken from item metadata, which is determined by the test developer; all other response demand codes were developed for these studies. [2] Importance statistics for these response demands come from regression tree analyses that include these two variables plus all other response demands in the table. See main text for details. [3] Conditional Random Forest R-squares for analyses with first two response demand categories excluded.

Sources:

Table 4a. Importance Statistics Study 3: Empirical Evidence for Important Item Response Demands, English Language Arts

	Grade									
	3	4	5	6	7	8 [1]	9	10	11	Mean
Item Design Demands										
Response Mode	--	--	27	83	--	--	--	--	--	20
Number of Score Categories	--	---	71	79	88	--	--	--	15	29
Item Type	--	--	58	80	61	--	--	--	21	26
Response Type	--	--	45	79	59	37	--	--	35	31
Interaction Type	--	--	48	55	47	32	11	--	26	26
TEI Type	--	--	100	94	49	38	26	--	82	46
Task Type	--	--	--	--	--	--	--	--	--	6
Task Model 1	--	--	--	21	--	--	--	--	--	7
Number of Points	--	--	--	--	29	--	--	--	--	9
Reading Selection Demands										
Text Complexity	--	--	--	--	--	51	--	--	--	9
1st Passage Identifier	--	41	--	100	64	82	--	--	98	61
Media Type	--	--	--	--	--	--	--	--	--	2
Set Identifier	--	35	--	75	53	62	--	--	100	55
Passage Word Count	--	31	--	68	46	56	--	--	98	48
Passage Type	--	53	--	--	--	23	--	--	--	11
Stimulus Identifier	--	21	--	38	33	48	--	--	58	33

	Grade									
	3	4	5	6	7	8 [1]	9	10	11	Mean
					Content Demands					
Evidence Statement 1	--	21	--	--	--	36	--	--	--	19
Evidence Statement 2	--	--	28	43	39	44	--	--	61	44
Evidence Statement 3	--	33	--	62	22	--	--	--	55	33
Sub-claim	--	100	65	32	47	30	--	--	88	52
					Cognitive Demands					
Command of Textual Evidence	--	--	--	40	--	--	--	--	28	13
					Linguistic Demands					
Processing Demands	--	29	--	26	100	--	100	--	0	39
Conditional Random Forest R-square	.05	.12	.17	.32	.37	.20	.14	.00	.10	.16

Note. See text for definitions of Item Design, Reading Selection, Content, Cognitive, and Linguistic response demands.

[1] The importance of Overall Cognitive Complexity in grade 8 English Language Arts, not included in this table, is 100.

Source: *z z 1. CC Final Report FINAL TO 07-27-15*; Table 3.9. *Importance Statistics for Predictors of ELA/Literacy Task P-Values*and Table 3.3. *Importance Statistics for Predictors of Mathematics Task P-Values*

Table 4b. Importance Statistics Study 3: Empirical Evidence for Important Item Response Demands, Mathematics

	Grade/Content Area									
	3	**4 [2]**	**5**	**6**	**7**	**8**	**Algebra 1**	**Algebra 2**	**Geom.**	**Mean**
					Item Design Demands					
Response Mode	--	--	--	--	--	--	--	--	--	--
Number of Score Categories	--	51	--	--	--	--	--	--	--	--
Item Type	26	72	--	--	--	--	--	--	--	--
Response Type	--	20	36	40	--	--	22	23	38	23
Interaction Type	--	--	38	32	--	--	--	21	31	20
TEI Type	46	71	100	100	100	100	100	100	100	93
Task Type										
Task Model 1	84	45	36	58	23	53	--	21	--	31
Number of Points	--	--	--	--	--	--	--	--	--	--
					Stimulus Demands					
Stimulus Material	--	--	--	--	--	--	--	--	--	--
Companion Materials	--	--	--	--	--	--	--	--	--	--
Stimulus Identifier	--	--	--	--	--	--	--	--	--	--
					Content Demands					
Mathematical Content	28	40	--	--	--	--	--	--	--	--

	Grade/Content Area									
	3	4 [2]	5	6	7	8	Algebra 1	Algebra 2	Geom.	Mean
Mathematical Practices	--	36	--	--	--	--	--	--	--	--
Evidence Statement 1	100	61	40	69	30	58	18	26	--	37
Sub-claim	--	48	--	--	--	--	--	--	--	--
CCSS Identifier 1	54	64	--	31	--	33	--	--	--	22
CCSS Identifier 2	--	--	--	--	--	--	--	--	--	--
Cognitive Demands										
Calculator Code	--	--	--	--	--	--	--	--	--	--
Linguistic Demands										
Processing Demands	--	--	--	--	--	--	--	--	--	--
Conditional Random Forest R-square	.40	.50	.39	.33	.44	.47	.47	.33	.36	.37

Note. See text for definitions of Item Design, Reading Selection, Content, Cognitive, and Linguistic response demands.

[2] The importance of [RD ***] in grade 4 Mathematics, not included in this table, is 100.

Source: *z z 1. CC Final Report FINAL TO 07-27-15*; Table 3.9. *Importance Statistics for Predictors of ELA/Literacy Task P-Values* and Table 3.3. *Importance Statistics for Predictors of Mathematics Task P-Values*