Smarter Balanced Assessment Consortium:

Mathematical Reasoning Project Quantitative Analyses Results: Grades 4, 8, and 11

Eva L. Baker, Ayesha Madni,
Joanne K. Michiuye, Kilchan Choi,
and Li Cai

June 2015

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Prepared by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) for submission under contract with the Council of Chief State School Officers for the Smarter Balanced Assessment Consortium.

Mathematical Reasoning Project Quantitative Analyses Results: Grades 4, 8, and 11

Introduction

In the context of the current report, feature analysis is defined as the qualitative rating of tasks against a set of attributes, followed by a subsequent quantitative analysis to determine how these attributes determine task performance. The four main components of the feature analysis process include feature rating, step-by-step analysis, cognitive labs, and quantitative analysis. For the feature analysis for the Mathematical Reasoning Project, we asked the following overarching questions:

1. What particular attributes/features does each item contain?
2. What are the dominant attributes/features across items?
3. What particular attributes/features of items/tasks contribute to increased or reduced item difficulty across items/tests/tasks and why?
4. What feature combinations contribute to increased and reduced difficulty and why?
5. What features are most valid to elicit and assess key learning outcomes?

The current report will focus on the five preceding questions, with particular emphasis on Questions 3, 4, and 5. Essentially, the current report combines item feature analysis across Grades 4, 8, and 11 with item response analysis across the same grades using a logistic linear test model (Fischer, 1973, 2005). The primary purpose of this analysis is to examine the relationship between item difficulty and identified item features on fourth, eighth, and eleventh grade items. This report will summarize and provide an analysis of these findings related to each respective grade.

This report contains select descriptive statistics across feature categories and items per grade and item-based examples highlighting specific feature and item difficulty components. The purpose of the descriptive statistics across feature categories is to determine whether any patterns appear that can inform the item difficulty analysis.

Theoretical Background

One of the earliest references to the idea of feature analysis can be attributed to Gordon (1970) in the *Report of the Commission on Tests: II*. In his brief, Gordon mentions qualitative analysis of assessments in his call to emphasize description and prescription (i.e., the qualitative description of cognitive functions leading to the prescription of the learning experiences required to more adequately ensure academic success). He continues by recommending that the College Entrance Examination Board add descriptive patterns of achievement and function derived from qualitative analysis of existing tests to its reports. Gordon suggests that existing instruments could be examined with a view to categorization and interpretation to determine whether or not the data can be reported in descriptive and qualitative ways in addition to traditional quantitative ways. Gordon for instance mentions that response patterns can be reported differently for information recall or vocabulary, and further refers to features such as problem solving, expression, and information management among many others.

The main rating framework underlying the feature analysis work is derived from Baker and O'Neil's (2002) approach to designing problem-solving assessments, Jonassen's (2000) typology of problems, and CRESST's problem-solving ontology. Baker and O'Neil's approach first characterizes three types of problem-solving tasks: (a) a task in which an appropriate solution is known in advance, (b) a task in which there is no known solution to the problem, and (c) a task that requires an

application of a given tool set to a broad range of topics. These are all features of problem solving that can be rated as part of a particular item and that require a specific associated cognitive demand or process on the part of the learner, student, or user.

Identifying the problem is often one of the most difficult aspects of problem solving (see Baker & O'Neil, 2002). The ambiguity of problem identification may be dependent on the prior knowledge of the learner and the purpose of the assessment. Essentially, an assessment developer can adjust the difficulty of a task or item by stating the problem explicitly or obscuring it in an embedded setting or context, such as within a narrative. These types of adjustments might not only contribute to increased difficulty, but might also require an associated cognitive demand or cognitive process that might be more complex.

Similar to Baker and O'Neil (2002), Jonassen (2000) articulates different problem types with varying attributes that follow a continuum from well-structured to ill-structured tasks. Jonassen's problem typology assumes that there are similarities in the cognitive processes required to solve each problem, that the problem types are not mutually exclusive, and that each problem category varies with respect to abstractness and complexity. Baker and O'Neil's (2002) approach to developing problem-solving assessments is part of an overarching model developed by CRESST to determine functional validity of assessments. Essentially, this model surpasses other forms for providing evidence in support of assessment validity claims by integrating in-depth and detailed feature and step-by-step analysis and modern statistical modeling. This model ensures validity by going beyond simple task descriptions, and by yielding explanations for possible areas of growth, identifying task elements that are suitable for instruction, and lastly, providing a method for comparability and prediction.

Target Features

Only the features that were found across more than four items and less than 20 items within a grade level were included in the current analysis and report, as features on either end would not contribute to significant variance in item performance. Based on the fourth, eighth, and eleventh grade item difficulty analysis and the inclusion criteria above, the following features were represented in the items. Across grades, features were rated based on the problem stimulus, prompt, and problem type.

Problem Stimulus

Problem stimulus refers to what is presented and asked of students as part of a task. For problem stimulus the following features contributed to variance in item performance.

- Math concepts required to solve the problem
- Math vocabulary
- Representation type(s)
- Multiple solutions

Item Prompt

Item prompt refers to the question stem within the task. For item prompt the following feature contributed to variance in item performance.

- Text type

Problem Type (Level of Cognition)

Problem type refers to the cognitive demands and processes associated with completing a task.

- Problem solving
- Strategy use
- Problem interaction type
- Scoring criteria/partial credit
- Cognitive load

Analytic Approach

Our analytic approach to address the research questions is the linear logistic test model (LLTM: Fischer, 1973, 2005; also see De Boeck & Wilson, 2004, for more recent updates to this model). This model addresses a key question: What are the item or item-cluster features that are significantly related to item characteristics? The fitting of this explanatory model not only provides estimates of the percentage of variance explained by the features, but also answers questions related to feedback. Are there particular features as determined by empirical analysis that should be emphasized in providing guidance to students (or perhaps to their teachers) to improve their performance? The explanatory analysis also addresses questions that are design focused, such as are there item or task-related features that are undesirable in that they may be associated with domain variance? The features also shed more light on other classification frameworks. For example, are math items or tasks classified as cognitively complex consistently associated with high cognitive load and narrative text? Do features related to response format explain item difficulty variation?

LLTM is an extension of the Rasch model (RM, Rasch, 1960) which decomposes item parameters into a linear combination of several basic parameters (Fischer, 1973, 2005). First, RM can be specified as follows:

$$P(X_{vi} = 1 \mid \theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \quad , \qquad (1)$$

where $P(X_{vi} = 1 \mid \theta_v, \beta_i)$ is the probability that person *v* gives a correct response to item *i*, given her or his ability θ_v and item difficulty for item *i*, β_i. In the LLTM, item parameters, β_i ($i = 1, \ldots, k$), are decomposed into a weighted sum of basic parameters, η_j ($j = 1, \ldots, p$), and fixed weights, ω_{ij}.

$$\beta_i = \sum_{j=1}^{p} \omega_{ij}\eta_j \qquad i = 1, \ldots, k \qquad (2)$$

The fixed weights are item-specific covariates, which are item features identified in our feature analysis and cognitive lab study. These are all binary indicators that take a value of 1 if item *i* has an identified feature, 0 for otherwise. The basic parameter, η_j, can be interpreted as regression coefficients that are associated with the item difficulty parameter, β_i, and identified item features, ω_{ij}.

First, we estimated item difficulty based on the Rasch model for the fourth, eighth, and eleventh grade item response data. For the Smarter Balanced fourth grade pilot test item response data, the total number of cases was 12,651 with 25 items. For the Smarter Balanced eighth grade field test item response data, 6,000 cases with 25 items were analyzed. Likewise, for the Smarter Balanced eleventh grade field test item response data, a total of 6,000 cases with 25 items were included in the analysis. Among the fourth grade items, 18 items have binary responses, six items have three categories, and one item has four categories. For the eighth grade items, 21 items have binary

responses, three items have three categories, and one item has four categories. Lastly, for the eleventh grade items, 20 items have binary responses and five items have three category responses. Also note that there was a significant amount of missing item response data due to the matrix sampling design during the field testing of the items, so the covariance matrix for items is very sparse. Thus, we chose the Rasch model to fit the item response data.

Item Difficulty Analysis and Results

The one-parameter logistic item response model (e.g., Rasch model) yields an item characteristic curve for each item. It reflects the probabilities that individuals across a range of trait levels are likely to answer each item correctly. For item characteristic curves, the x-axis reflects a wide range of trait levels, and the y-axis reflects probabilities ranging from 0 to 1.0. We can examine an item's curve to find the likelihood that an individual with a particular trait level will answer the item correctly. Item difficulty (b) is defined as the location on latent trait (x-axis) where $p = .5$.

Table 1 contains the item difficulty levels for the fourth grade items. The item difficulty analysis was performed on a total of 25 fourth grade pilot test items with 12,651 cases. As indicated in Table 1, two items were relatively easy (negative b), and some items were relatively difficult (i.e., Items 1, 2, 8, 13, 22, and 25).

Table 1. Item Difficulty Analysis for Fourth Grade

Item	Difficulty (b)		
2	1.50		
5	0.14		
6	0.55		
7	-0.27		
8	1.91		
9	-0.98		
10	0.77		
11	0.48		
13	2.25		
14	1.25		
16	1.48		
17	0.10		
18	0.11		

Item	Difficulty (b)		
19	0.16		
21	1.19		
22	2.66		
23	0.75		
24	1.38		
	Difficulty 1 (b1)	**Difficulty 2 (b2)**	**Difficulty 3 (b3)**
1	1.28	2.53	2.70
3	0.96	1.22	
4	-0.92	0.29	
12	-0.24	0.10	
15	-1.24	0.41	
20	-0.63	0.37	
25	1.51	2.10	

Figure 1. Pilot test item characteristic curve for Grade 4.

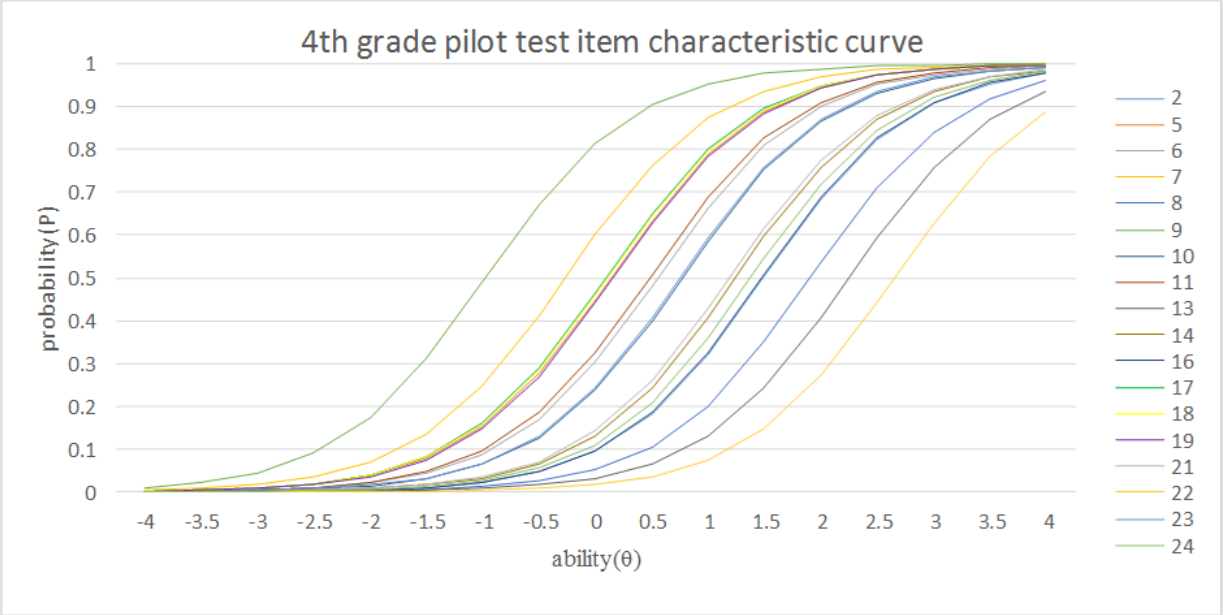


Table 2 contains the item difficulty for the eighth grade items. The item difficulty analysis was performed on a total of 25 eighth grade field test items with 6,000 item performance cases. As indicated in Table 2, Items 1, 6, 16, and 25 were difficult items, while Items 11, 13, 19, and 23 were relatively easy items.

Table 2. Item Difficulty Analysis for Eighth Grade

Item	Difficulty (b)		
1	2.67		
2	0.83		
3	0.49		
4	0.08		
5	1.02		
6	2.39		
7	0.30		
9	1.13		
10	1.41		
11	-0.92		
12	0.46		
13	-0.10		
14	1.05		
15	1.78		
17	1.25		
18	1.55		
19	-0.53		
20	0.51		
21	1.27		
22	1.17		

Item	Difficulty (b)		
25	2.21		
	Difficulty 1 (b1)	Difficulty 2 (b2)	Difficulty 3 (b3)
8	-1.45	0.80	2.81
16	1.81	2.16	
23	-0.46	0.95	
24	0.68	3.09	

Figure 2. Field test item characteristic curve for Grade 8.

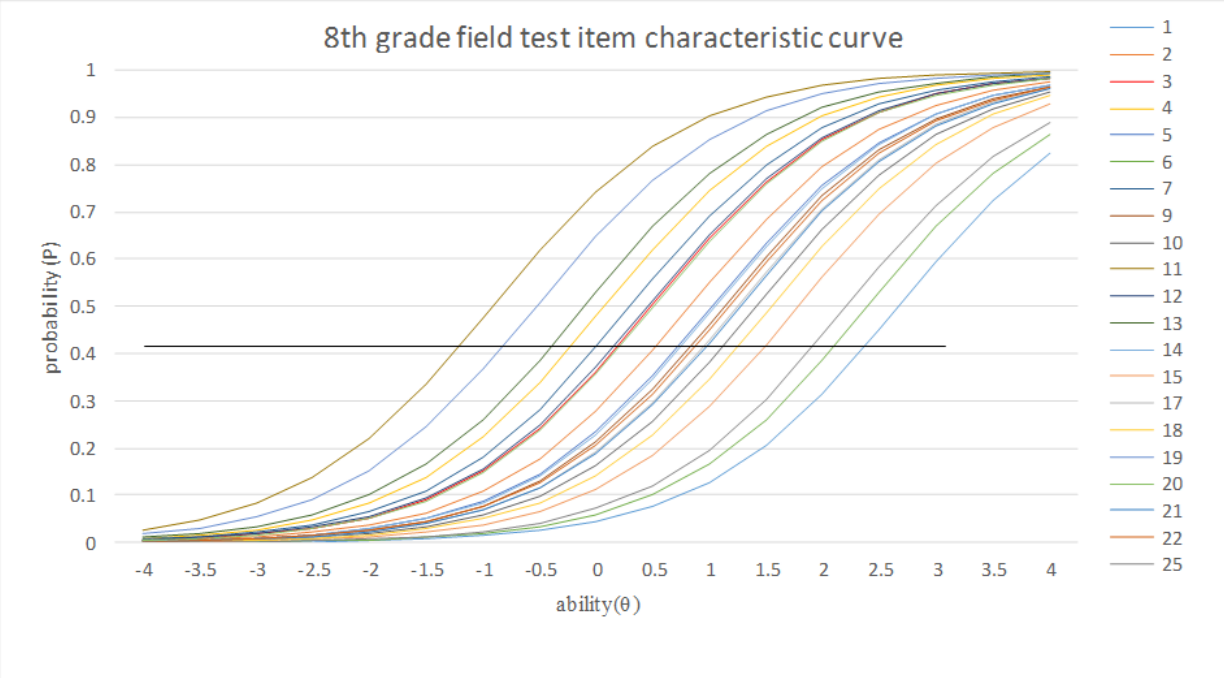


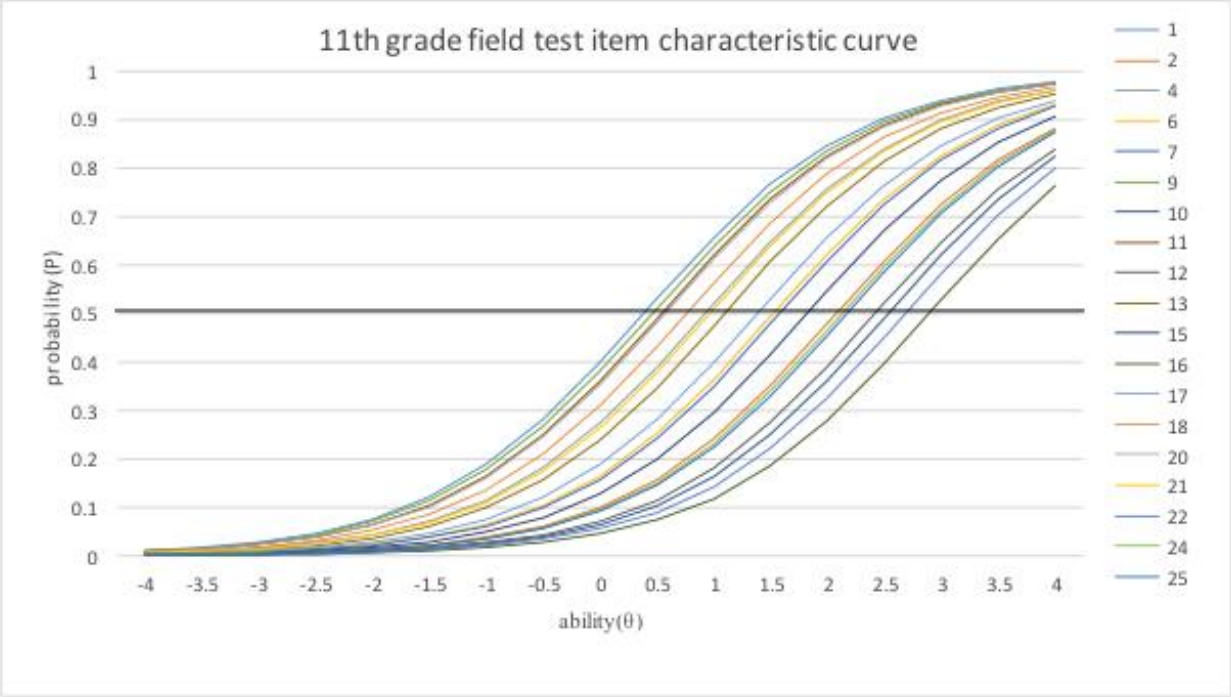
Table 3 contains the item difficulty levels for the eleventh grade items. The item difficulty analysis was performed on a total of 25 eleventh grade field test items with 6,000 item performance cases. As indicated in Table 3, most of the eleventh grade items had positive values, indicating that for the most part, there were no particularly easy items, and seven of the items were particularly difficult (i.e., Items 2, 10, 12, 16, 22, 24, and 25).

Table 3. Item Difficulty Analysis for Eleventh Grade

Item	Difficulty (b)	
1	0.38	
2	2.08	
4	0.92	
6	1.53	
7	1.59	
9	0.45	
10	2.54	
11	0.54	
12	2.43	
13	1.09	
15	1.83	
16	2.90	
17	1.39	
18	0.75	
20	0.57	
21	0.96	
22	2.69	
24	2.14	
25	2.17	
	Difficulty 1 (b1)	Difficulty 2 (b2)
3	-1.89	-0.25
5	-0.32	1.98
8	1.17	3.27
14	-0.31	1.09

Item	Difficulty (b)	
19	2.32	2.96
23	3.36	4.97

Figure 3. Pilot test item characteristic curve for Grade 11.



Linear Logistic Test Model Results: Combining Quantitative Analysis with Feature Analysis

In an LLTM, item features essentially account for the Rasch model's item difficulty. In applying identified features as fixed weight, basic parameters are estimated. As a result, LLTM-based item difficulty parameters are estimated. These parameters should approach the Rasch model-based item difficulty parameters. The closer the LLTM-based item difficulty parameters are to the Rasch model-based item parameters, the better our features, identified qualitatively, have accounted for the "true" (Rasch model based) difficulty parameters. In the following sections we provide the estimated coefficients of identified item features for the fourth, eighth, and eleventh grade items. We also present the percentage of variance in item difficulty parameters accounted for by identified item features.

Fourth Grade Results

For fourth grade, the features related to the problem stimulus included math concepts, math vocabulary, and representation types. Features related to the prompt included text type, and features related to problem type included problem solving, strategy use, problem interaction type, and scoring. Table 4 provides the results from LLTM analysis for the fourth grade item features.

Table 4. LLTM Analysis Results for Fourth Grade Item Features

Number	Feature	Coefficient	*SE*
1	Number representation/conversion	-0.69*	0.12
2	Units and currency	-0.35*	0.13
3	Fractions	1.05*	0.22
4	Decimals/place value	-0.42*	0.10
5	Simple equations	-0.07	0.15
6	Partial credit	1.36*	0.17
7	Drag and drop	-0.32*	0.11
8	Fill in the blank	-2.20*	0.10
9	Select all that apply	-2.40*	0.16
10	Multiple choice	0.54*	0.13
11	Story problem	1.01*	0.11
12	Single sentence	0.53	0.31
13	Multi-sentence text	0.00	0.31
14	Math vocabulary	-0.94*	0.07
15	Equation	2.24*	0.15
16	Strategy use	0.74*	0.13
17	Recall and reproduction	-0.46*	0.21
18	Procedural/algorithmic	-0.70*	0.21
19	Cognitive load for solving problem	-0.30*	0.08
	Intercept	1.20*	0.03

Note. Positive coefficients indicate presence of features that may reduce item difficulty (i.e., features with positive coefficients make items easier). Negative coefficients indicate presence of features that may increase item difficulty (i.e., features with negative coefficients make items more difficult).
$*p < .05$.

As indicated in Table 4, a total of 19 features were included in the analysis. Of those 19, 10 features contributed to increased item difficulty and six features reduced item difficulty. Related to the problem stimulus, four math concepts appeared to increase item difficulty. These math concepts included number representation/conversion, units and currency, decimals/place value, and simple equations. As indicated in Table 4, number representation/conversion contributed the most to item difficulty of the fourth grade math concept features, as it had a larger negative coefficient than units and currency, decimals/place value, and simple equations. On the other hand, one math concept feature appeared to significantly decrease item difficulty with a medium positive coefficient (i.e., fractions).

Related to the problem stimulus, both math vocabulary and representation type, in this case equation, also contributed to variance in item difficulty. Math vocabulary increased item difficulty with a medium negative coefficient and equation significantly decreased item difficulty with the largest positive coefficient of the fourth grade features. These results are consistent with earlier findings from cognitive labs, as students approached problems that required straight computation, as in the case of an equation being provided as the stimulus, with more understanding than problems requiring students to decipher charts and graphs. Moreover, while it seems counterintuitive that items with below grade level standard vocabulary (i.e., math vocabulary) would contribute to increased item difficulty, it became apparent during earlier cognitive labs that students who did not understand or could not recall such terminology had more difficulty with completing items. One such example involved many students not recalling the difference between prime, even, and odd numbers.

In terms of the item prompt, it is unexpected that text type, such as multi-sentence text and story problem, did not contribute to increased item difficulty above single-sentence text, especially since earlier cognitive labs indicated that items which contained multi-sentence text (i.e., more text and detail) were more difficult for students than story problems, which contain more narrative. In particular, story problems contributed to a significant decrease in item difficulty with a medium positive coefficient. Single-sentence text also contributed to a decrease in item difficulty, but on a smaller scale than story problem, which had a smaller positive coefficient. Multi-sentence text, on the other hand, contributed to no variance in item difficulty with a coefficient of zero. During the fourth grade cognitive labs, students indicated that they understood the directions for most problems, which could corroborate the findings for both multi- and single-sentence text.

With respect to problem type, problem solving related to both recall and reproduction and procedural/algorithmic thinking increased item difficulty. However, procedural/algorithmic had a larger negative coefficient (i.e., medium negative coefficient) than recall and reproduction, thereby contributing to a higher increase in item difficulty. These findings are consistent with earlier cognitive lab findings and are aligned with the cognitive processing involved with the two different problem-solving types. Essentially, procedural/algorithmic thinking requires more cognitive processing and has a higher cognitive demand than recall and reproduction. Thus, it is clear that procedural/algorithmic would contribute to a higher increase in item difficulty than recall and reproduction.

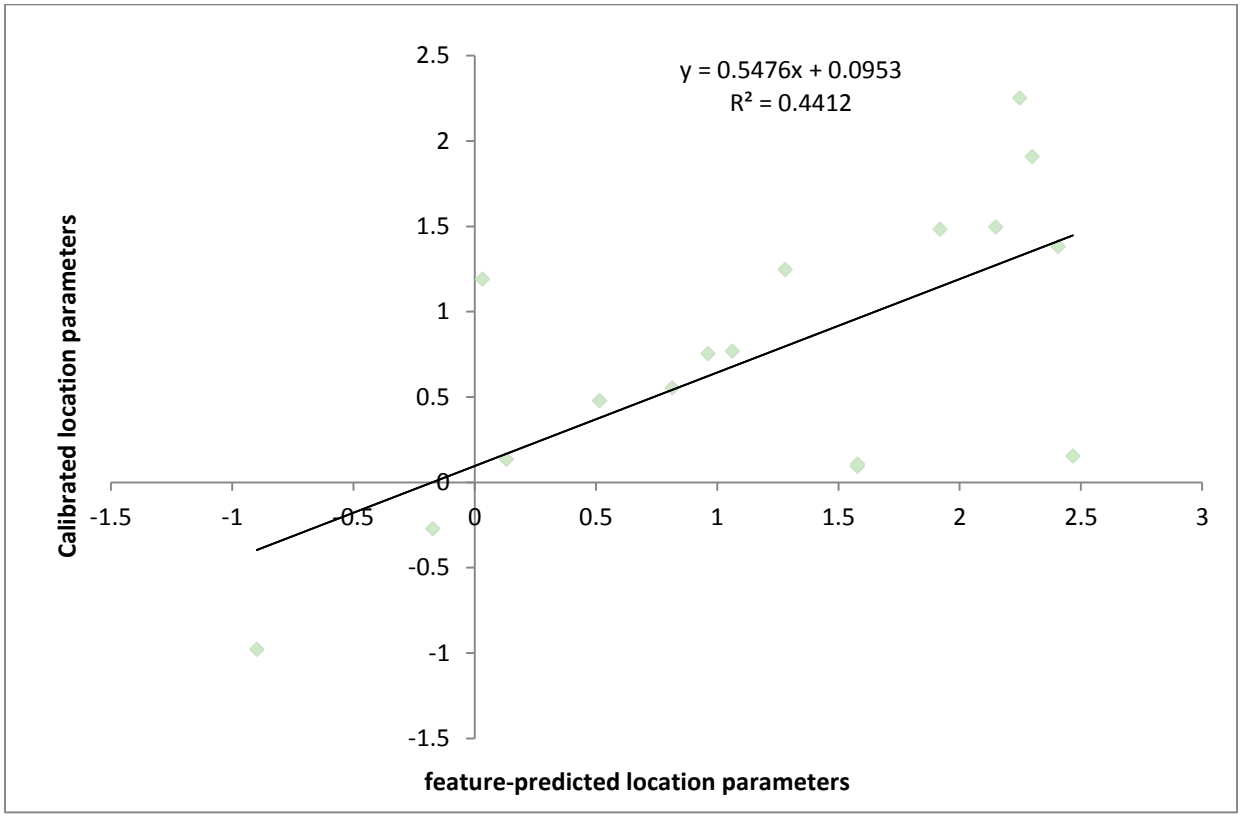
The finding that strategy use and partial credit both contributed to reduced item difficulty is consistent with the fact that with both features students have more opportunities or varying avenues

for receiving points or a complete score on an item. However, of the two, partial credit contributed to significantly more reduction in item difficulty with a larger positive coefficient than strategy use. On the other hand, the problem interaction types including drag and drop, fill in the blank, and select all that apply increased item difficulty, particularly with respect to fill-in-the-blank and select-all-that-apply items, which had the largest negative coefficients of all the fourth grade features. This finding is consistent with earlier cognitive labs, as students experienced significant challenges with items containing these problem interaction types. Students had some difficulty with drag and drop due to the novelty and the technology-based interaction. Students struggled with fill-in-the-blank items, as the items were often unclear as to the extent and expectation for the open-ended responses students could provide. With respect to the select-all-that-apply items, students appeared to be unsure of how many of the options they could select—this problem interaction type was counterintuitive to their familiarity with multiple choice, where students select the best answer. Multiple-choice problem interaction, therefore, reduced item difficulty, containing a small to medium positive coefficient, due to the familiarity of the problem type and response mode. Lastly, cognitive load for solving the problem was found to increase item difficulty for fourth grade, having a smaller negative coefficient.

In summary with respect to degree of difficulty for fourth grade items across features, number representation, units and currency, decimals/place value, simple equation, drag and drop, recall and reproduction, and procedural algorithmic knowledge ranked in the category of smaller negative coefficients, indicating that these features contributed to some increase in item difficulty. Further, math vocabulary had a medium negative coefficient suggesting a slightly higher increase in item difficulty. The remaining features that contributed to a significantly large increase in item difficulty included fill in the blank and select all that apply, which both contained large negative coefficients. As for the features that contributed to reduced item difficulty, equation as representation type contributed to the highest decrease of item difficulty with the largest positive coefficient. Story problem, fractions, and partial credit also contributed to a significant decrease in item difficulty having medium positive coefficients. Moreover, strategy use, single-sentence text, and multiple choice also contributed to a reduction in item difficulty with smaller positive coefficients. Lastly, multi-sentence text did not contribute to variance in item difficulty with a coefficient of zero.

Figure 4 below depicts the item difficulty parameters with calibrated item difficulty by Rasch model on the y-axis and feature-predicted item difficulty using LLTM on the x-axis. Essentially, the features explained a significant portion of variance in item difficulty (R^2 = .44).

Figure 4. LLTM results for Grade 4.



Eighth Grade Feature Results

For eighth grade, the features related to the problem stimulus included math concepts, language complexity, representation types, and multiple solutions. Features related to the prompt included text type, and features related to problem type included problem interaction type. Table 5 provides the LLTM analysis results for the eighth grade item features.

Table 5. LLTM Analysis Results for Eighth Grade Item Features

Number	Feature	Coefficient	*SE*
1	Function	0.12	0.13
2	Order of operations	-0.93*	0.12
3	Fill in the blank	0.26*	0.11
4	Multiple response	0.49*	0.13
5	Multiple choice	1.01*	0.13

Number	Feature	Coefficient	*SE*
6	Story problem	-0.95*	0.10
7	Single sentence	-0.48*	0.17
8	Math vocabulary	-0.12	0.10
9	Equation	-0.14	0.12
10	Geometric shape	-0.64*	0.11
11	Coordinate plane	0.12	0.10
12	Charts/graphs	-0.30*	0.12
13	Multiple solutions possible	0.83*	0.11
14	Cognitive load for solving problem	-0.73*	0.08
	Intercept	0.66*	0.08

Note. Positive coefficients indicate presence of features that may reduce item difficulty (i.e., features with positive coefficients make items easier). Negative coefficients indicate presence of features that may increase item difficulty (i.e., features with negative coefficients make items more difficult).
**p < .05.*

As indicated in Table 5, a total of 14 features were included in the analysis. Of those 14 features, six contributed to increased item difficulty and four reduced item difficulty. Related to the problem stimulus, one math concept (i.e., order of operations) appeared to significantly increase item difficulty having a medium negative coefficient, whereas one item appeared to decrease item difficulty (i.e., function) with a small positive coefficient.

Further related to the problem stimulus, both math vocabulary and representation type, in this case, equation, geometric shape, coordinate plane, and charts/graphs, also contributed to variance in item difficulty. Math vocabulary, equation, geometric shape, and charts/graphs increased item difficulty containing a range of small negative coefficients, and coordinate plane decreased item difficulty also with a small but positive coefficient. These results are consistent with earlier cognitive lab findings, as students appeared to have a clearer understanding of problems that contained coordinate planes as opposed to problems requiring students to decipher charts/graphs and geometric shapes. In contrast to fourth grade, where equation as a representation type significantly decreased item difficulty with the largest positive coefficient of the feature set, for eighth grade equation increased difficulty with a small negative coefficient. This finding could be explained by the eighth grade equations often being more complex and that the eighth grade items often contained more than one equation to solve.

Similar to the fourth grade item features, items with below grade level standard vocabulary (i.e., math vocabulary) contributed to increased item difficulty. However, for the eighth grade items they contributed to a smaller increase in item difficulty (i.e., small negative coefficient). This could potentially be explained by the fact that if students did not fully comprehend the key math terminology when initially taught, that it might persist as a gap in pertinent background knowledge.

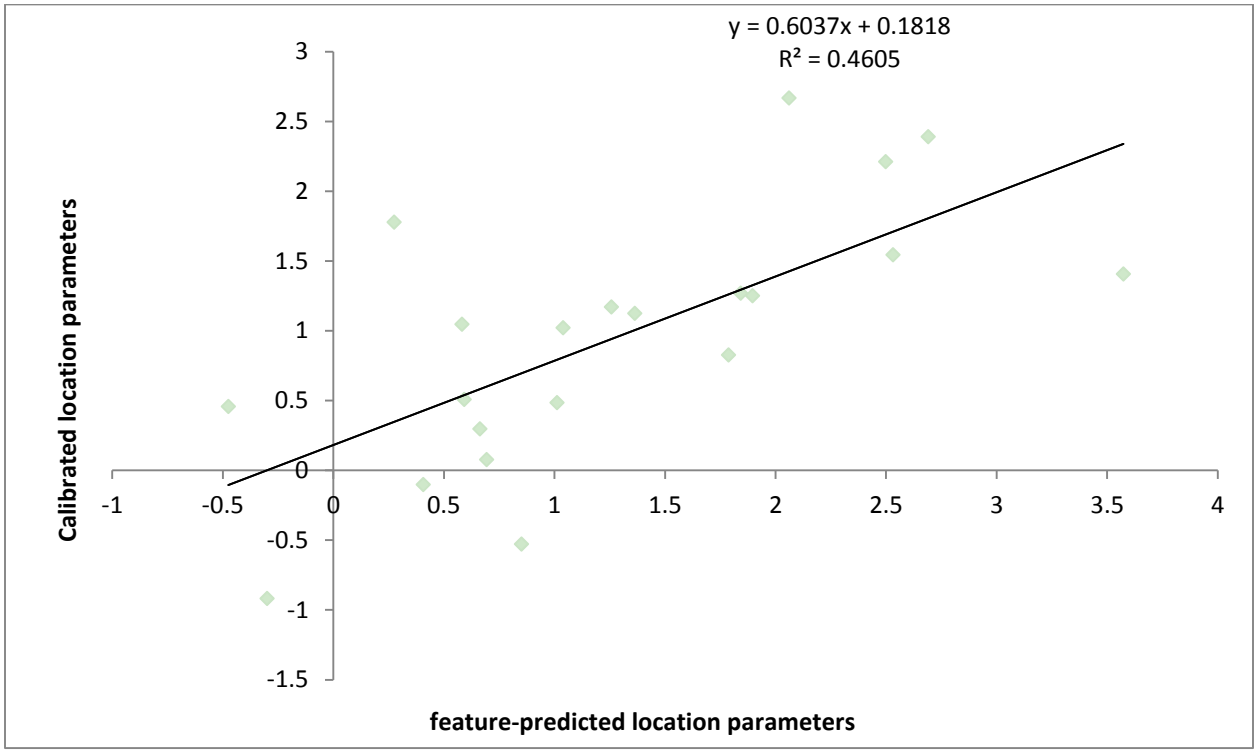
Having multiple solutions possible, on the other hand, was found to significantly reduce item difficulty, as that feature had a medium positive coefficient. This finding is as expected as it provides more opportunities for students to complete an item correctly.

In terms of the item prompt, story problem and single-sentence text contributed to increased item difficulty with story problem containing a medium negative coefficient and single-sentence text having a smaller negative coefficient. This finding is consistent with cognitive labs in that items which contained more narrative and text appeared to be more challenging for eighth grade students. With respect to problem interaction type, items including fill in the blank, multiple response, and multiple choice reduced item difficulty, with fill in the blank and multiple response having smaller positive coefficients, and multiple choice having a medium coefficient. While these results are somewhat divergent from the fourth grade results, it was apparent during cognitive labs that eighth grade students were more comfortable with the various problem interaction types and the corresponding technology than fourth grade students, and subsequently, that the fill-in-the-blank items were clearer in terms of what response was expected. Lastly, the feature of cognitive load for solving the problem increased difficulty for the eighth grade items with close to a medium negative coefficient.

In summary with respect to degree of difficulty for eighth grade items across features, charts/graphs, geometric shape, equation, math vocabulary, and single-sentence text ranged in the category of smaller negative coefficients, indicating that these features contributed to some increase in item difficulty. Further, cognitive load, story problem, and order of operations had medium negative coefficients suggesting slightly higher increases in item difficulty. None of the features for eighth grade contained large negative coefficients. As for the features that contributed to reduced item difficulty, multiple choice and multiple solutions possible contributed to the highest decrease of item difficulty with medium positive coefficients. Function, fill in the blank, multiple response, and coordinate plane decreased item difficulty having small positive coefficients. In contrast to the fourth grade items, the eighth grade item features appeared to fall within a narrower range of item difficulty both on the negative and positive end.

Figure 5 below depicts the item difficulty parameters with calibrated item difficulty by Rasch model on the y-axis and feature-predicted item difficulty using LLTM on the x-axis. In essence, the features accounted for a significant portion of variance in item difficulty (R^2 = .46).

Figure 5. LLTM results for Grade 8.



Eleventh Grade Feature Results

For eleventh grade, features related to the problem stimulus included math concepts, language complexity, and representation types. Features related to the prompt included text type, and features related to problem type included problem interaction type. Table 6 provides the LLTM analysis results for the eleventh grade item features.

Table 6. LLTM Analysis Results for Eleventh Grade Item Features

Number	Feature	Coefficient	*SE*
1	Polynomial	0.83*	0.22
2	Factoring	-0.75*	0.28
3	Creating equations	0.82*	0.15
4	Slope/rate of change	-0.31*	0.19
5	Function	1.81*	0.24
6	Linear equation	-0.17*	0.15

Number	Feature	Coefficient	*SE*
7	Exponentiation	-1.05*	0.15
8	Drag and drop	-2.17*	0.14
9	Fill in the blank	-1.53	0.19
10	Multiple response	-1.61	0.17
11	Story problem	0.42*	0.11
12	Math vocabulary	1.35*	0.12
13	Equation	0.25	0.15
14	Coordinate plane	-1.34	0.25
15	Charts/graphs	-0.38*	0.13
16	Cognitive load for solving problem	-2.27*	0.15
	Intercept	0.99*	0.11

Note. Positive coefficients indicate presence of features that may reduce item difficulty (i.e., features with positive coefficients make items easier). Negative coefficients indicate presence of features that may increase item difficulty (i.e., features with negative coefficients make items more difficult).
**p < .05.*

As indicated in Table 6, a total of 16 features were included in the analysis. Of those 16 features, seven contributed to increased item difficulty and five reduced item difficulty. Related to the problem stimulus, four math concepts (i.e., factoring slope/rate of change, linear equation, and exponentiation) appeared to increase item difficulty having varying degrees of item difficulty. Factoring and exponentiation fell within the range of medium negative coefficients, and slope/rate of change and linear equation had small negative coefficients. On the other hand, three math concepts (i.e., polynomial, creating equations, and function) appeared to decrease item difficulty, with polynomial and creating equations having medium positive coefficients, whereas function significantly decreased item difficulty as it had a large positive coefficient.

Further related to the problem stimulus, both math vocabulary and representation type, in this case equation, coordinate plane, and charts/graphs, also contributed to variance in item difficulty. In contrast to fourth and eighth grades, math vocabulary significantly reduced eleventh grade item difficulty as it had a medium positive coefficient. These findings are consistent with expectations as eleventh grade students have accumulated more background knowledge in key math terminology. Cognitive labs also confirmed eleventh grade students' understanding of prior math terminology. For eleventh grade, equation had a small positive coefficient, similar to fourth grade, where equation had the largest positive coefficient of the fourth grade features. Cognitive labs showed that equation decreased item difficulty. Aligned with eighth grade items, the feature of charts/graphs increased difficulty for eleventh grade items, with a small negative coefficient. However, for eighth grade coordinate plane reduced item difficulty with a small positive coefficient, but for eleventh grade it

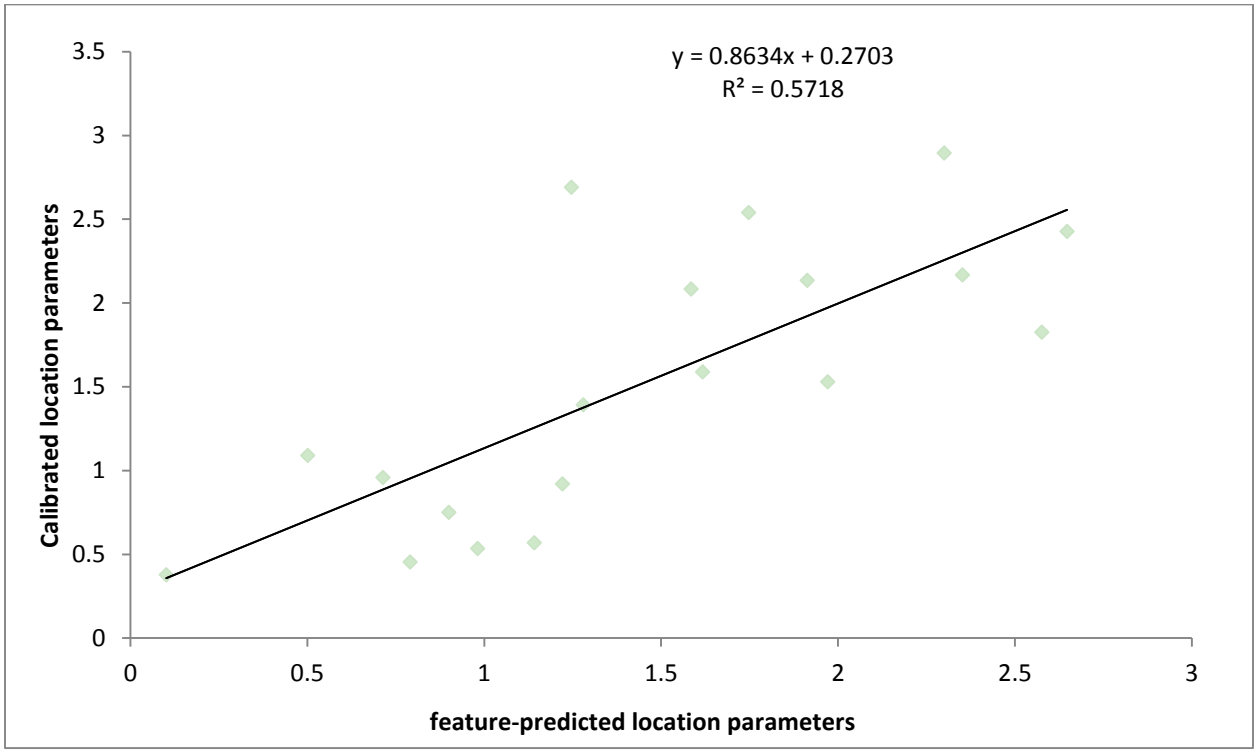
significantly increased item difficulty with a medium to large negative coefficient. From cognitive labs it was apparent that the coordinate planes for the eleventh grade items were more complex.

In terms of the item prompt, story problem was found to reduce item difficulty having a small positive coefficient. This finding is consistent with cognitive labs as items which contained more narrative and text did not appear to pose more of a challenge or be more confusing for eleventh grade students as opposed to fourth and eighth grade students. With respect to problem interaction type, items including drag and drop, multiple response, and fill in the blank significantly increased item difficulty, with all having large negative coefficients. This can be explained by eleventh grade students' previous experience with paper-and-pencil assessments as opposed to computer-based tests. It also became apparent during cognitive labs that it was not consistently intuitive for students to know how they were supposed to respond to multiple response as well as drag-and-drop items. Lastly, cognitive load for solving problems increased difficulty for the eleventh grade items much beyond what was found for the eighth grade items, as it had a large negative coefficient. This finding is as expected, as eleventh grade items are likely to be more cognitively taxing than eighth grade items.

In summary with respect to degree of difficulty for eleventh grade items across features, charts/graphs, linear equation, and slope/rate of change placed in the category of smaller negative coefficients, indicating that these features contributed to some increase in item difficulty. Further, factoring, exponentiation, and coordinate plane had medium negative coefficients suggesting a higher increase in item difficulty. The remaining features that contributed to significantly high increases in item difficulty included drag and drop, fill in the blank, multiple response, and cognitive load, which all contained large negative coefficients. In fact, drag and drop and cognitive load were the features that contributed most to item difficulty for eleventh grade. As for the features that contributed to reduced item difficulty, function contributed to the highest decrease in item difficulty with the largest positive coefficient. Math vocabulary, creating equations, and polynomial also contributed to a significant decrease in item difficulty having medium positive coefficients. Lastly, story problem and equation contributed to reduction in item difficulty containing smaller positive coefficients. In contrast to the eighth grade items, the eleventh grade item features appeared to fall within a wider range of item difficulty both on the negative and positive end, much like the fourth grade items.

Figure 6 below depicts the item difficulty parameters with calibrated item difficulty by Rasch model on the y-axis and feature-predicted item difficulty using LLTM on the x-axis. The features explained a significant portion of variance in item difficulty (R^2 = .57).

Figure 6. LLTM results for Grade 11.



Feature Representation Across Fourth, Eighth, and Eleventh Grade Items

The following sections detail the type of features represented in the various items across Grades 4, 8, and 11, and the percentage for which select overarching features are represented in the items.

Feature Representation Across Fourth Grade Items

For fourth grade a total of 64 features was rated. Overall, the 25 items ranged from having 11% feature representation to 27% representation. Of these, 20 of the items contained 10 or more features, whereas five items contained nine features or less.

With regard to specific feature categories, the 25 items ranged from covering one to four math concepts per item, with 84% of the items containing two or more math concepts. Moreover, with respect to representation types, the 25 items ranged from containing two to four types, with 44% of the items containing three or more representation types and the remaining 56% of items containing two types. Further, for the feature category dealing with tools provided to aid in item completion, almost all 25 items did not have a tool provided. For the text type feature category, the items ranged from containing one to two text types. Of the total 25 items, 48% contained one text type and the remaining 52% contained two text types. In regards to problem solving, the items ranged from containing one type of problem solving to three types of problem solving. The majority (84%) of the items contained one type of problem solving. For the problem interaction type feature category, the 25 items ranged from containing one to two problem interaction types, with the majority (84%) of items containing one problem interaction type.

In addition, 25% of the fourth grade items had partial credit possible, 72% contained math vocabulary, 4% depended on world knowledge, 12% had multiple solutions possible, 8% included scaffolding, 20% could be solved through strategy use, and finally 64% included some degree of cognitive load.

Feature Representation Across Eighth Grade Items

For eighth grade a total of 67 features were rated. Overall, the 25 items ranged from having 12% feature representation to 25% representation. Of these, 21 of the items contained 11 or more features, whereas four items contained 10 features or less. Compared to the fourth grade items, the eighth grade items had a higher density of features overall.

Similar to the fourth grade items, the 25 eighth grade items ranged from covering one to four math concepts per item, with 68% of the items containing two or more math concepts. Moreover, with respect to representation types, the eighth grade items ranged from containing two to four types, with 96% of the items containing three or more representation types. This finding indicates that the eighth grade items contained more representation types than the fourth grade items overall, but also that each eighth grade item likely had a higher frequency of representation types than the fourth grade items.

For the feature category dealing with tools provided to aid item completion, similar to fourth grade, all 25 eighth grade items did not have a tool provided. For the text type feature category, the items ranged from containing one to two text types, and similar to the fourth grade items, 52% contained one text type and the remaining 48% contained two text types. In contrast to fourth grade, the eighth grade items only contained one type of problem solving per item. For the problem interaction type feature category, the 25 items ranged from containing one to two problem interaction types, with the majority (92%) of items containing one problem interaction type.

In addition, 16% of the eighth grade items required students to obtain relevant information, 4% of the fourth grade items had partial credit possible, 52% contained math vocabulary, 4% depended on world knowledge, 8% contained academic math vocabulary, 20% had multiple solutions possible, 16% could be solved through strategy use, and finally 60% included some degree of cognitive load.

Feature Representation Across Eleventh Grade Items

For eleventh grade a total of 59 features were rated. Overall, the 25 items ranged from having 15% feature representation to 25% representation. Of these, 21 of the items contained 11 or more features, whereas four items contained 10 features or less. Compared to the fourth grade items, the eleventh grade items had a higher density of features overall, and also appeared to have a higher frequency and wider range of features per item than both fourth and eighth grade items.

Further, in contrast to both the fourth and eighth grade items, the 25 eleventh grade items ranged from covering one to seven math concepts per item, with 84% of the items containing two or more math concepts. With respect to representation types, the eleventh grade items had a wider range than both fourth and eighth grade items containing two to six types, with 56% of the items containing four or more representation types and 44% containing three or fewer representation types. This finding indicates that the eleventh grade items contained more representation types than the fourth and eighth grade items overall, but also that each eleventh grade item likely had a higher frequency of representation types than the fourth and eighth grade items.

For the feature category dealing with tools provided to aid item completion, similar to both fourth and eighth grade, all 25 eleventh grade items did not have a tool provided. For the text type feature category, the items ranged from containing one to two text types, similar to both fourth and eighth

grade items. Of the 25 eighth grade items, 40% contained one text type and the remaining 60% contained two text types. Similar to fourth grade but in contrast to eighth grade, the eleventh grade items only contained two types of problem solving. However, only one item contained two types of problem solving, whereas the remaining 24 items contained one type of problem solving each. For the problem interaction type feature category, the 25 eleventh grade items all contained one problem interaction type, inconsistent with both fourth and eighth grade findings.

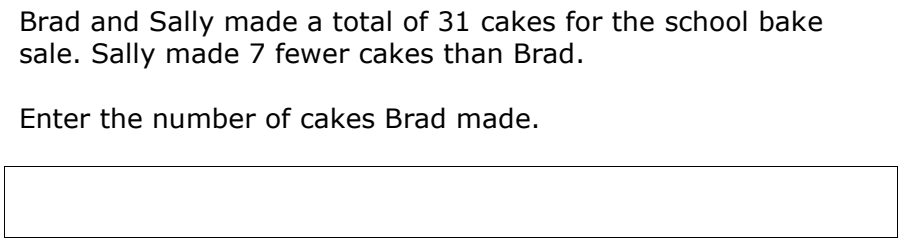
In addition, 12% of the eleventh grade items required students to obtain relevant information, 12% of the eleventh grade items had partial credit possible, 76% contained math vocabulary, 8% could be solved through strategy use, and finally 56% included some degree of cognitive load.

Item-Based Illustrative Examples

The following section provides specific item-based examples highlighting features and item parameter data for select eighth and eleventh grade items.

Eighth Grade Examples

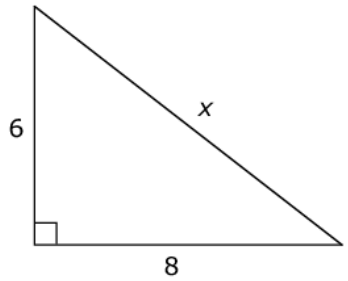
Figure 7. Grade 8 Math Item 5.



Item 5 in Figure 7 obtained a predicted difficulty and a calibrated difficulty that were almost the same, with the predicted difficulty at 1.04 and the calibrated difficulty at 1.02. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was almost identical to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 1.02 is considered somewhat complex, which suggests that this was a fairly difficult item for students. This item contained two features from the quantitative analysis: fill in the blank and story problem. For eighth grade, fill in the blank had a small positive coefficient (0.26), which indicated that it reduced item difficulty somewhat. On the other hand, story problem had a medium negative coefficient (-0.95), which suggested that it contributed to item difficulty. Consistent with the cognitive labs for eighth grade, it was apparent that this item posed difficulty for students due to the story problem aspect (i.e., the narrative). Specifically, the students seemed to be somewhat confused by the prompt and were not clear about how to approach the problem after having read the narrative. The cognitive lab data show that three out of 14 students completed this item correctly.

Figure 8. Grade 8 Math Item 20.

A right triangle is shown.

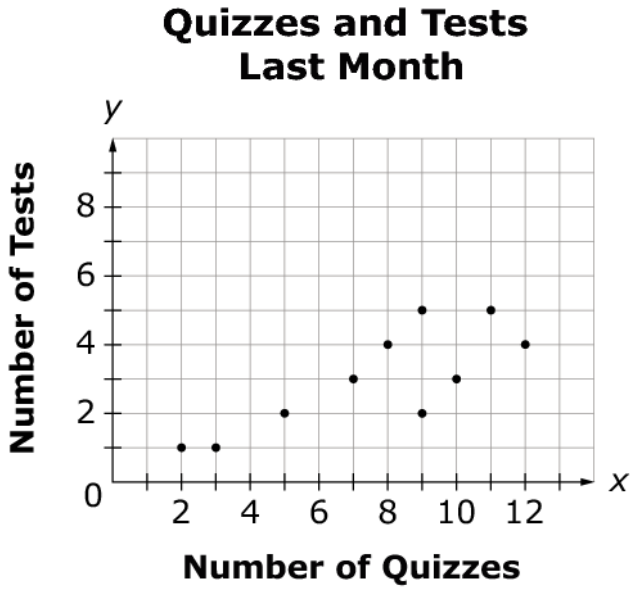


Enter the value of x.

Similar to the preceding item, Item 20 in Figure 8 obtained a predicted difficulty and a calibrated difficulty that were almost the same, with the predicted difficulty at 0.59 and the calibrated difficulty at 0.57. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was almost identical to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 0.57 is considered slightly challenging. This item contained five features including fill in the blank, math vocabulary, geometric shape, multiple solutions, and cognitive load for solving the problem. For eighth grade, fill in the blank (0.26) and multiple solutions possible (0.83) both had positive coefficients which indicated that these features reduced item difficulty. On the other hand, math vocabulary (-0.12), geometric shape (-0.64), and cognitive load (-0.73) all had negative coefficients, which suggested that these features contributed to item difficulty. Consistent with the cognitive labs for eighth grade, it was apparent that items with geometric shapes were more difficult for students, and that items with multiple solution paths were easier for students if they were aware of math-specific strategies, such as in this case the 3-4-5 rule.

Figure 9. Grade 8 Math Item 25.

Amanda made the scatter plot shown below based on the results of a survey of 10 classmates about how many quizzes and tests they each had last month.



The equation for the line of best fit for the data set was $y = 0.36x + 0.27$.

Then, Amanda added a point representing her quizzes and tests last month. She had 12 quizzes last month. After adding her point, she recalculated the equation for the line of best fit and found that the slope had decreased.

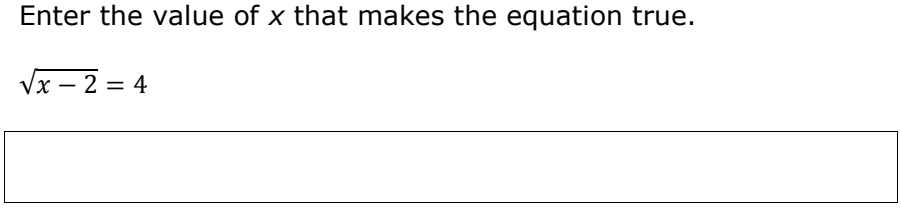
Determine the maximum number of tests Amanda could have had last month.

Item 25 in Figure 9 obtained a predicted difficulty and a calibrated difficulty that were similar, with the predicted difficulty at 2.4 and the calibrated difficulty at 2.21. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was close to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 2.21 is considered complex, which suggests that this was a difficult item for students. This item contained eight features including order of operations, fill in the blank, story problem, math vocabulary, equation (representation type), coordinate plane, multiple solutions possible, and cognitive load. For eighth grade, fill in the blank (0.26), coordinate plane (0.12), and multiple solutions possible (0.83) had positive coefficients, and were found to reduce item difficulty. In

contrast, order of operations (-0.93), story problem (-0.95), math vocabulary (-0.12), equation (representation type) (-0.14), and cognitive load (-0.73) had negative coefficients, thereby increasing item difficulty. Clearly, this item had more features with larger negative coefficients than with large positive coefficients, corroborating its calibrated difficulty. Consistent with the cognitive labs for eighth grade, it was apparent that this item posed some difficulty for students due to the story problem aspect, and because students had difficulty interpreting the representation.

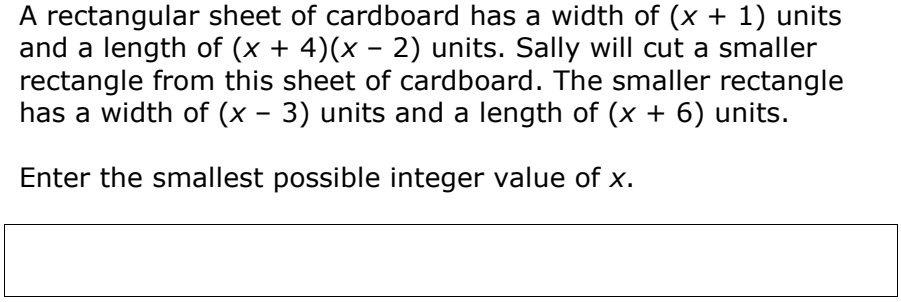
Eleventh Grade Examples

Figure 10. Grade 11 Math Item 1.



Item 1 in Figure 10 obtained a predicted difficulty and a calibrated difficulty that were similar, with the predicted difficulty at 0.1 and the calibrated difficulty at 0.38. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was close to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 0.38 indicates that this was a fairly easy item for students. This item contained four features including linear equation, fill in the blank, math vocabulary, and equation (representation type). For eleventh grade, fill in the blank (-1.53) and linear equation (-0.17) had negative coefficients, with fill in the blank having a large negative coefficient, thereby contributing to increased item difficulty. In contrast, math vocabulary (1.35) and equation (representation type) (0.25) had positive coefficients, with math vocabulary having a large positive coefficient, thereby contributing to reduced item difficulty. Based on the foregoing feature parameters, the calibrated item difficulty for Item 1 is consistent.

Figure 11. Grade 11 Math Item 7.



Item 7 in Figure 11 obtained a predicted difficulty and a calibrated difficulty that were almost the same, with the predicted difficulty at 1.62 and the calibrated difficulty at 1.59. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was almost identical to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 1.59 indicates that this was a significantly difficult item for students. This item contained eight features including polynomial, creating equations, linear

equation, exponentiation, fill in the blank, story problem, math vocabulary, and cognitive load for solving the problem. For eleventh grade, fill in the blank (-1.53), linear equation (-0.17), exponentiation (-1.05), and cognitive load (-2.27) had negative coefficients, with fill in the blank and exponentiation having large negative coefficients, and cognitive load having the largest negative coefficient. These features therefore corroborate the calibrated difficulty of this item. In contrast, polynomial (0.83), creating equation (0.82), story problem (0.42), and math vocabulary (1.35) had positive coefficients, with math vocabulary having the largest positive coefficient of the four features. However, the features that increase difficulty had larger coefficients overall, thereby contributing to complexity of the item. Based on the foregoing feature parameters, the calibrated item difficulty for Item 7 is consistent. Moreover, it was apparent during cognitive labs for eleventh grade that the fill in the blank and cognitive load aspects of this item contributed to student difficulty. Students were not consistently clear about how to approach the problem nor about what information to provide in the fill-in-the-blank box.

Figure 12. Grade 8 Math Item 17.

The function $f(x) = -\frac{3}{25}(x-3)^2 + 73$ can be used to model the temperature, in degrees Fahrenheit, of a spring day in Nashville, where x is time in hours starting at noon.

Select True or False for each statement.

	True	False
The highest daily temperature occurs at noon.	☐	☐
The daily temperature is 73 degrees Fahrenheit at 9 a.m.	☐	☐
The highest daily temperature occurs at 3 p.m.	☐	☐
The daily temperature is the same at 9 a.m. and 3 p.m.	☐	☐

Item 17 in Figure 12 obtained a predicted difficulty and a calibrated difficulty that were very similar, with the predicted difficulty at 1.28 and the calibrated difficulty at 1.39. This finding demonstrates that the difficulty that was predicted based on the item feature analysis was very close to the item difficulty obtained based on student item performance data. Moreover, a calibrated item difficulty rating of 1.39 indicates that this was a difficult item for students. This item contained eight features including function, linear equation, exponentiation, multiple response, story problem, math vocabulary, equation (representation type), and cognitive load. For eleventh grade, linear equation (-0.17), exponentiation (-1.05), multiple response (-1.61), and cognitive load (-2.27) had negative coefficients, with exponentiation and multiple response having large negative coefficients, and cognitive load having the largest negative coefficient. In contrast, function (1.81), story problem (0.42), math vocabulary (1.35), and equation (representation type) had positive coefficients, with function and math vocabulary having the largest positive coefficients of the four features. However, the features that increase difficulty had larger coefficients overall, thereby contributing to the complexity of the item. Based on the foregoing feature parameters, the calibrated item difficulty for Item 17 is consistent. Moreover, it was apparent during cognitive labs for eleventh grade that the multiple response and cognitive load aspects of this item contributed to student difficulty. Students were not consistently clear about how to approach the problem and therefore were also unclear about what options to select out of the multiple responses.

Conclusions and Implications

As indicated in the goals and questions in the preceding section, there are several important implications that result from feature analysis. Combining the qualitative analysis (i.e., feature analysis and cognitive labs) with the quantitative analysis (e.g., item analysis and linear logistic test model, Fischer, 1973, 2005) addresses a key question: What are the item features that are significantly related to item difficulty?

As such, the goal of combining student performance data with the feature ratings is to essentially determine what features contributed to increased and reduced difficulty based on student performance on the particular items and experience with particular features. This analysis not only determines what items were most and least difficult, but also why students performed higher or lower on the items based on particular features and/or feature combinations and their associated difficulty level. For instance, an item or task might be classified as significantly complex due to high cognitive load and narrative text, and in contrast, another item or task might be considered easy due to a simpler target concept or response mode, such as multiple choice.

As indicated in the previous sections, features are conceptually related to problem solving, cognitive demands and processes, and problem types. Features can also be rated or derived according to specific content/concepts, task elements, user interactions, and response modes among many other features dependent on the task and content at hand. Features can therefore also be analyzed across tasks including, but not limited to, game levels, scenarios, instructional activities, computer simulations, tests, and other media, thereby applying across contexts and situations to include instruction and learning, as well as assessment. The fact that the same features can be rated and analyzed across contexts and tasks provides a parsimonious tool to ensure alignment between instructional modes, assessment development, and student learning, but also enables novel validity analyses prior to the data analysis stage. Essentially, validity can be ensured, verified, and "built in" during the development process by incorporating salient (valid) features into target products, whether they be instructional activities, games, assessments, or related to specific student learning behaviors and cognition.

Thus, obtaining data on an item feature basis takes item development and item difficulty to a new level by pinpointing what particular aspects or attributes of items are contributing to increased or reduced difficulty, and by ensuring validity of instruction and assessment during the development process as opposed to during post hoc analysis by comparing predicted validity for features with calibrated validity. This type of information and level of detail are highly valuable to all stakeholders within the educational process including, but not limited to students, teachers, test developers, funding institutions, and policy makers. For example, teachers can use the results to not only understand their students' strengths and weaknesses, but also formatively respond to their students' needs by redesigning and adjusting both their instruction and assessment accordingly. For item and test developers, this conceptual and analytical framework has a high value proposition in customizing items and tests beyond aligning to content and standards. An additional benefit as a test developer is that one could create items that target specific salient features, for example career or college readiness features, that might be problematic for students across consecutive years.

The information obtained from such items is also fundamental to decision makers related to funding and policies. If one can determine the specific areas where a particular school district or school is having difficulties with respect to student performance in one year, then funding can be provided and target test packages can be created including items with salient feature components for the following year to accurately determine student progress and improvement. Reports of these results in common language will provide more tailored information on student performance and

improvement, and thereby more accurate information on what instructional areas need additional attention and where resources need to be provided.

References

Baker, E. L., & O'Neil, H. F., Jr. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18,* 609-622.

De Boeck, P., & Wilson, M. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. H. (2005). Linear logistic test models. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (Vol. 2, pp. 505-514). Amsterdam: Elsevier.

Gordon, E. W. (1970). Toward a qualitative approach to assessment. *Report of the Commission on Tests, II. Briefs* (pp. 42-46). New York: College Entrance Examination Board.

Jonassen, D.H. (2000). Toward a design theory of problem solving. *Educational Technology: Research & Development, 48*(4), 63-85.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.