Incorporating Accessibility and Complexity Concepts into Test Specification and Anchor Set Selection for Alternate Assessments Based on Alternate Achievement Standards

Presented at the annual meeting of the American Education Research Association, Special Interest Group - Inclusion and Accommodation in Educational Assessment

April 17, 2015

Chicago, IL

Anne H. Davidson, Smarter Balanced Assessment Consortium

Sarah L. Hagge, Minnesota Department of Health

Bill Herrera, edCount

Charlene Turner, edCount

Martha L. Thurlow, National Center for Educational Outcomes

Karla L. Egan, Center for Assessment

Rachel Quenemoen, National Center and State Collaborative

The contents of this paper were developed as part of the National Center and State Collaborative under a grant from the U.S. Department of Education (PR/Award # H373X100002), Project Officer, Susan.Weigert@ed.gov. However, the contents do not necessarily represent the policy of the U.S. Department of Education and no assumption of endorsement by the Federal government should be made.

Keywords: Alternate Assessment based on Alternate Achievement Standards, Students with Significant Cognitive Disabilities, Test Specification, Complexity, Accessibility, Anchor Items, Equating

Abstract

The study investigated how accessibility and complexity concepts can be incorporated in test specifications in order to build item banks that appropriately cover a range of complexity for all covered content, including criteria for selection of anchor items for a common-item, non-equivalent groups equating design. The National Center and State Collaborative state and organizational research and content partners articulated models of learning of how students with significant cognitive disabilities build competence in each of the domains tested (i.e., math, reading, writing). These models then informed design specifications for *families* of items to (a) be developed for each priority content target in the testing blueprint, (b) ensure the resulting item pool reflected a range of complexity, and (c) support features in a given family of items all related to the same content target. The design specifications for each level of an item in a family were called *tier* specifications. Building on the SRI PADI system to ensure fidelity of implementation of the design specifications, the goal was to maintain a single construct across a group of items aligned to a given test standard while systematically varying the items' complexity and supports/accessibility features. Results of item trials suggested that the items resulting from the development of tier specifications were functioning in ordinal patterns of empirical item difficulty. Mean tier p-values were consistently ordinal. Given limitations of item trial form designs (e.g., no student took more than one tier of item per trial; broad range of ability of students who participated), it is expected that the full census operational data will allow for better understandings to further refine this approach. This study contributes to literature by investigating performance of new, scaffolded item types to refine test and anchor specifications.

Incorporating Accessibility and Complexity Concepts into Test Specification and Anchor Set Selection for Alternate Assessments Based on Alternate Achievement Standards

Introduction

The National Center and State Collaborative (NCSC) is a consortium of states and national centers building an alternate assessment based on alternate achievement standards (AA-AAS) for students with the most significant cognitive disabilities. The effort is guided by a theory of action that incorporates instructional context, assessment design, intended score interpretation and use, and intended long-term student outcomes. NCSC approached this challenge of developing a comprehensive assessment system by ensuring the design was developed within the broader framework of rigorous and relevant academic standards, curriculum, and instruction.

Using a principled approach to design based on evidence-centered design (ECD) literature, Design Patterns and Task Templates were developed to serve as item specifications. The resulting Task Templates and Design Patterns (tools built into the ECD process that serve as precursors to item development) served as the mechanism by which varying levels of content difficulty were implemented in the family of assessment items measuring a particular aspect of the core academic content in Mathematics, Reading, and Writing. Each Task Template was designed to facilitate the creation of four items, an item *family*. These items are intended to target the range of abilities within the target population, and this approach allowed items developed to be accessible to students with varying levels of cognitive functioning and communication capabilities. This integrated methodology of ECD and Universal Design used an assessment

design process that incorporated the assumption of interaction between content, task, and learner characteristics in the creation of assessment items. Building on the SRI PADI system to ensure fidelity of implementation of the design specifications, the goal was to maintain a single construct across a group of items aligned to a given test standard while systematically varying the items' complexity and supports/accessibility features.

Central to the NCSC program goals is operational test forms that can be scaled and equated using item response theory models (IRT). Equating requires that test forms maintain an equivalent construct and are parallel in terms of content and statistical specifications (Dorans & Holland, 2000; Kolen & Brennan, 2004). This study investigated how accessibility and complexity concepts can be incorporated in test specifications, including the equating design and criteria for selection of anchor items for a common-item, non-equivalent groups equating design.

Three questions were of specific interest in this study:

1) *To what degree did the tier design specification result in content and empirical characteristics of the NCSC items?*
2) *To what extent did items within a given family demonstrate evidence of a singular test construct?*
3) *How should item tier characteristics be represented within a set of anchor items as compared to using item statistics independent of tier assignment?*

Perspectives

Alternate assessments based on alternate achievement standards (AA-AAS) have seen a sustained trend toward greater standardization while maintaining flexibility through universal design and accommodation (Gong & Marion, 2006; Quenemoen, Kearns, Quenemoen, Flowers, & Kleinert, 2010). Since the first of such programs, AA-AAS have employed

strategies to develop confidence in the comparability of scores across students within grades (Schafer & Lissitz, 2009), including standardized test forms, administrator and scorer training, and test accommodation refinement. The advent of multi-state assessment consortia has prompted new expectations for test quality and score comparability, given larger and more representative student samples and the potential for IRT scaling and equating.

Since any assessment must be designed for a target population in order to report an interpretable score, AA-AAS design must address how students interact with content, how they communicate, and how they develop proficiency within an academic domain (Marion & Pellegrino, 2006). The AA-AAS target population (i.e., students with the most significant cognitive disabilities) is heterogeneous both in terms of typical demographic characteristics and learner characteristics (e.g., disability, expressive or receptive communication, classroom setting). Therefore, to design a test for the population, the assessment program must address how students access test content and constructs across all demographic and learner characteristics (Towles-Reeves, Kearns, Flowers, Hart, Kerbel, Kleinert, Quenemoen, & Thurlow, 2012).

Using a principled design approach building on evidence-centered design (ECD, Mislevy, 1996) and the work of the Committee on Assessments that resulted in the book *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), one multi-state and center consortium, NCSC, developed item and test specifications through a logical sequence of test development steps. The ECD steps include the *a priori* development of claims and rationales as representation of student cognition (Pellegrino et al., 2001). Defined as the empirically-based theories and beliefs about how students represent information and develop competence in a particular domain, these claims and resulting assessment targets (a) focus on what students need to know and be able to do in the given content domain and (b) establish hypotheses as to what evidence will

reflect the relationship between a claim and evidence of it within student response data (Mislevy & Risconscente, 2006).

A key outcome of NCSC's principled design process was articulated models of learning of how students with significant cognitive disabilities build competence in each of the domains tested (i.e., Mathematics, Reading, Writing). These models then informed design specifications for *families* of items to (a) be developed for each priority content target in the testing blueprint, (b) ensure the resulting item pool reflected a range of complexity, and (c) support features in a given family of items all related to the same content target. The design specifications for each level of an item in a family were called *tier* specifications. These tiers were used as a tool to specify item and form development (Table 1).

Based on the articulated models of learning, the structure of the tier specifications incorporate concepts of access to academic content/construct, cognitive complexity, and language complexity, and are defined in four levels in each content area. The tiers ranged from test questions designed to allow for the students who are very early in instruction with the academic content to test questions designed to reflect expectations very near/at grade-level. The items were written starting with content standards at grade level then considering how the other items in the family can be translated so that students at different levels of functioning or communication would be able to interact with the construct.

Items based on the tier specifications were developed within *families* and were intended to retain an equivalent construct while varying complexity and scaffolding to address student access (Table 1). Specifications, including item-level complexity notes which documented the characteristics (e.g., number of decimals points in a number, Lexile and length of passages), were systematically controlled to create for a graduated degree of complexity across the family of

items from most to least complex. Additionally, the NCSC assessment blueprints incorporated tiers by specifying the marginal percentage of tiered items per content standard.

While the use of a model of learning prompted its use, tier specifications were developed as a test development tool for creating a full range of accessible items across the range of performance, and investigation of the statistical functioning of item tiers is needed to facilitate interpretation and understanding of test scores. Designing test forms to incorporate selection of items based on tiers was intended to allow the AA-AAS to ensure that students across a broad performance range could show what they know. In addition, however, scaling and equating would need to be considered in order to convert raw scores into a scale score and produce interchangeable forms. One of the most crucial components of a non-equivalent groups equating design is the careful development of anchor item sets that accurately reflect the total test. Much of past equating research has recommended anchor sets that are "mini-tests" of the total test in terms of content and statistical characteristics (Kolen & Brennan, 2004). That is, the selection of a set of common anchor items needs to be both substantively (i.e., in terms of content specification) and statistically equivalent to the test forms being equated. It would follow that when we are incorporating within an item bank design those items that vary in complexity systematically and that are developed with consistent design principles, we would also need to consider how these criteria are consistent with anchor sets used for equating.

Methods

Two item trial pilots were conducted. Pilot 1 was conducted in spring 2014. Student demographic, learner characteristic, and item response data were collected from approximately 5,200 students from 17 U.S. states and territories during item trials in spring 2014. Eight forms per grade (3-8 and 11) and content area (Mathematics and English Language Arts (ELA)) were

administered. These linear, fixed-length forms incorporated tiers in item and test specification. They were administered via computer and one-on-one with trained teacher administrators. Pilot 2 was conducted in fall 2014 and focused on item functioning as well as test structure. Nineteen states and territories participated and over 6000 students participated with their teachers. A two-session test design was used in Pilot 2 to mirror the proposed design for the summative assessments in spring 2015. Ultimately, both pilots served to evaluate whether the items functioned as intended in format and across statistical properties (Standard 4.10, AERA/APA/NCME, 2014).

The two item trials (pilot tests) allowed for the first empirical examination of the item tier design with representative samples of students. In this study, results from the item trials were evaluated using descriptive data analysis to address study questions. Analyses aimed to identify (flag) weaknesses of items for instances when items did not perform as expected in terms of their tiers. First, we looked at tier reversals; next, we evaluated intact families of items in light of their tier specification and student performance. Finally, we evaluated the selection of intact families qualitatively to theorize on the design elements that could have contributed to within-family item performance.

Tier Reversal Analysis

Items were evaluated with respect to empirical item difficulty using classical item statistics by tier. Items were then flagged for instances in which their tier specification and difficulty were reversed. In other words, items were flagged if they were designed at a lower tier

(e.g., Tier 1) but had a higher difficulty than items with adjacent tier designations (e.g., Tier-1 has a higher difficulty than Tier-2 within the same item family).

To evaluate tier reversals, item p-values were calculated and compared across tiers. In ELA, item passages are written in families, where there are four variations of each passage, one for each of Tier 1 to Tier 4. Each tier of a passage has an associated set of items, and the number of items associated with a passage varies by tier and passage. Because there is not a direct one-to-one correspondence of items across the tiers, mean p-values were calculated in order to identify instances of tier reversal in ELA. The mean p-values were calculated by averaging the p-values of all items measuring the same content standard in each passage tier. The mean p-values were then compared across tiers. In Mathematics, an item family was a set of four items, one item per tier. Each content standard in Mathematics had between four and five item families. To calculate tier reversal flags in Mathematics, the p-value was first calculated for each item tier in an item family. The p-values were then compared across tiers within an item family.

Higher p-values indicate easier items and lower p-values indicate more difficult items. It was hypothesized that Tier 1 would be the easiest items and Tier 4 would be the most difficult items. Items were flagged for a Tier 4 reversal if the Tier 4 p-value was greater than any of the other tiers. Similarly, items were flagged for a Tier 3 reversal if the Tier 3 p-value was greater than the p-value for Tier 2 or Tier 1. Finally, items were flagged for a Tier 2 reversal if the Tier 2 p-value was greater than the p-value for Tier 1. Tables 2 and 3, ELA and Mathematics, respectively, represent the number of item families measuring a given content standard, or Core Content Connector (CCC), that contained a tier reversal flag.

Family-Tier Evaluation

Item specification in terms of tier was further evaluated using data analysis and qualitative review of intact item families in the Mathematics assessments across all grades (3-8 and 11). First, all families with complete tier representation (i.e., all four tiers represented) were identified. Next, p-values for item performance in both pilots were inspected. It was hypothesized that items within a given family would demonstrate ordinal relationship between items by tier, with Tier 1 having the highest p-value (easiest item), Tier 2 the next highest, Tier 3 the next highest, and Tier 4 having the lowest p-value (most difficult item). Families that demonstrated this pattern were identified for qualitative inspection to build theory as to what design elements may have contributed to the item-tier pattern within family. Likewise, families that did not demonstrate this pattern were also identified and inspected for relevant design characteristics.

Results

Tier Reversals

Preliminary results from the spring item trial are presented in Tables 2 and 3. The preliminary results provide evidence of item difficulty by item tier specification. Table 2 contains the number of items flagged at each tier in ELA, as well as the mean p-value, or item difficulty for each tier. Table 3 contains the same results for Mathematics. Results were very similar across both ELA and Mathematics. Few items were flagged for having a Tier 2, Tier 3, or Tier 4 p-value greater than the Tier 1 p-value, meaning that Tier 1 was nearly always easier than Tiers 2, 3 or 4. The majority of the tier reversals occurred as a result of Tier 3 being easier than Tier 2, or as a result of Tier 4 being easier than either Tier 3 or Tier 2. Mean Tier 1 p-values are nearly always

higher than any other tier.[1] Mean p-values for Tiers 2, 3 and 4 were often either similar in magnitude or diverged from the expected pattern.

Results of the item trial suggested that tiers were functioning somewhat consistently with empirical item difficulty. In almost all cases, Tier 1 items were easier than the other items in their family. Mean tier p-values were consistently ordinal. However, within individual item families, tiers were reversed more heavily in Tiers 2, 3, and 4. Specifically, in Pilot 1, 59 out of 64 (92.2%) ELA passages or foundational item families were flagged for a tier reversal; 218 out of 294 (74.1%) Mathematics item families were flagged for at least one tier reversal.

Family-Tier Evaluation

For all items in a single item family included in both pilot tests, a comparison of p-values across all tiers in Mathematics was conducted (Table 4). Items for those families that exhibited a tier reversal were selected for a review of the item characteristics including scaffolding features. Sixty (60) mathematics items representing 15 unique families across six (6) grades were reviewed. A summary of this qualitative review is presented in Table 5.

In general, the pilot tests reflected the systematic variability of complexity across items in a family and improvements in accessibility. However, in the isolated cases of tier reversals, researchers discovered the most frequent cause was rooted in the item structure. Specifically, the item features developed to support the student, may have instead contributed to unintended extraneous cognitive load. For example, a grade three item displays a data table twice: first with an interpretation of the cell contents and then a second time as part of the item stem. This repetition of the data table not only increases the cognitive load, but also suggests implications

1 Note that Tier-1 items had two answer choices; Tiers 2-4 items had 3 answer choices. See Discussion.

regarding the display of the item on the online assessment platform. Given the constraints of font size, spacing, and screen resolution, an item with multiple tables, charts and diagrams needs to be scrolled through in order to reach the stem and response options. Hence, the support may not be visible when the item stem is visible.

In some cases, the lack of an ordinal pattern of p-values suggests that students may not have had an opportunity to learn (OTL) the content. Based on the results from the pilot test and the corresponding end of test survey, there were some Mathematics concepts included in the Common Core State Standards - and represented on the assessment - not currently being addressed in instruction. In a well-aligned assessment system, students with significant cognitive disabilities have opportunities for learning academic content that is well matched to what their peers at that grade level are learning and being assessed against (Browder, Spooner, Wakeman, Trela, & Baker, 2006).

Discussion

Results of the study provide evidence to support the incorporation of varying complexity and accessibility features in item and test specification for AA-AAS. The evaluation of items, families, and ultimately the item pool required a starting hypothesis that items would produce patterns of ordinal mean p-values by their tier specification. On average, both item trials resulted in such patterns. In instances of tier reversals, items could be flagged for further evaluation. Further research in this area should continue to test the assumption that tier specifications *should* result in ordinal difficulty patterns, especially as students have additional opportunity to learn grade-level content aligned to standards. Additionally, the relationship between tier and depth of knowledge should be explored.

The study has provided guidance to future test development regarding specific item structural elements within a family of items. The generally ordinal nature of the p-value patterns within families could suggest that the item content design maintains a test construct. The within-family qualitative review of items pointed to specific issues (e.g., item structure) as possible explanations for tier reversal patterns. Construct definition and content alignment were not identified as issues in this preliminary qualitative review. Further research should look at empirical patterns within families and replicate qualitative analyses.

In situations where the p-value patterns did not present in an ordinal fashion, experts identified three primary reason codes: (a) need for specific refinements to Task Templates and Design Patterns, (b) computer-based rendering of items (especially items with more text or graphics; e.g., selected response items), and (c) a need for giving students additional opportunity to learn. These observations could prompt the clarification of language in the items' test directives, refinements to increase appropriateness of the item content and complexity, re-evaluation of the accessibility of the items, the maximization of readability and comprehensibility, and further alignment criteria to precisely defined constructs.

Further research should incorporate additional measures, including item response theory results, to investigate the question of whether tiers should demonstrate ordinal performance patterns. By conducting sample-independent analyses, the relationship between tiers and information functions could lead to a better understanding of item and tier performance within and across families. In addition, residual p-values should be reviewed, given the difference in guessing parameters between Tier-1 items (e.g., two answer choices) and the other items.

Conclusion

In recent years, the inclusion of new item formats has introduced additional considerations for test scoring, scaling and equating. Like these item format types, the item tiers in the NCSC assessment introduce another consideration for ensuring accuracy of student scores. This study contributes to the current literature by investigating the performance of a new, scaffolded item type, and how this item type should be represented within test and anchor specifications in order that student scores accurately reflect what students know and are able to do.

Results of the study suggest that both tiers and families of items varying by tiers are useful content constraints for test specification. Therefore, both concepts should be incorporated in the selection of anchor sets in order to undergird the content validity argument for the AA-AAS scores.

Results from this study also suggest that the tier of an item family is not strictly equivalent to difficulty, and therefore it cannot be represented through statistical specifications alone. Further, though the item tiers are designed to measure the same underlying construct, each tier represents a unique level of scaffolding and student access. Further research should investigate test construct across tiers and the relative equitability (Liu & Dorans, 2014).

References

Browder, D.M., Spooner, F., Wakeman, S., Trela, K., & Baker, J.N. (2006). Aligning instruction with academic content standards: Finding the link. *Research & Practice for Persons with Severe Disabilities, 31*(4), 309–321

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities.* A paper presented at the Large-Scale Assessment Conference, Council of Chief State School Officers. San Francisco, CA.

Kolen, M.J., & Brennan, R.L., (2004). *Test equating, linking, and scaling: Methods and practice* (2nd ed.). New York: Routledge Falmer.

Marion, S.F., & Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice 25*(4), 47-57.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Mislevy, R.J., & Risconscente, M.M. (2006). Evidence-centered assessment design. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). New York: Routledge.

Pellegrino, J.W., Chudowsky, .J., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy of Sciences.

Liu, J., & Dorans, N. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32(*1), 15-22.

Quenemoen, R., Kearns, J., Quenemoen, M., Flowers, C., & Kleinert, H. (2010). *Common misperceptions and research-based recommendations for alternate assessment based on alternate achievement standards* (Synthesis Report 73). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Schafer, W.D., & Lissitz, L.W. (Eds.). *Alternate assessment based on alternate achievement standards: Policy, practice, and potential.* Baltimore: Paul H. Brookes.

Towles-Reeves, E., Kearns, J, Flowers, C., Hart, L., Kerbel, A., Kleinert, H., Quenemoen, R., & Thurlow, M. (2012). *Learner characteristics inventory project report (A product of the NCSC validity evaluation).* Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

Table 1. Item Tiers

Tier*	Content Assessed	Complexity	Scaffolding	Additional Features
1	Essential Understanding of CCC**	Least complex content	Greatest use of non-construct relevant scaffolds	
2	Focal KSA***	Grade level but less complex than Tiers 3 or 4	Non-construct relevant scaffolds	In math, items may use modeling for multi-step problems.
3	Focal KSA	Grade level but less complex than Tier 4	Some items include non-construct relevant scaffolds	In math, items may use modeling for multi-step problems.
4	Focal KSA	On grade level; most complex coverage of focal KSA	Minimal use of non-construct relevant scaffolds	

Notes. *Items assessing a single construct, aligned to one core content connector (content standard), and representing all four tiers are considered a family; **Core content connector (CCC), ***Knowledge/skill/ability of CCC

Table 2. Summary of Tier Flags by Content Standard in ELA, Pilot 1

Grade	CCC	Number of Passages	Tier 4			Tier 3		Tier 2	Tier 1 Mean p-value	Tier 2 Mean p-value	Tier 3 Mean p-value	Tier 4 Mean p-value
			> Tier 3	> Tier 2	> Tier 1	> Tier 2	> Tier 1	> Tier 1				
3	3.RI.H1	4	1	2	0	4	0	0	0.65	0.38	0.47	0.41
	3.RI.H4	4	2	0	0	0	0	0	0.87	0.72	0.61	0.55
	3.RI.I2	4	1	1	0	3	0	0	0.83	0.55	0.49	0.50
	3.RI.K5	2	0	0	0	1	1	0	0.76	0.57	0.60	0.43
	3.RL.H1	4	1	2	0	3	0	0	0.79	0.56	0.63	0.57
	3.RL.I2	4	1	2	0	4	1	0	0.78	0.65	0.72	0.63
	3.RL.K2	4	2	1	0	0	0	0	0.77	0.56	0.49	0.45
	3.RWL.H2	4	1	1	0	1	0	0	0.79	0.51	0.48	0.41
	3.RWL.I2	6	1	1	0	2	0	0	0.85	0.58	0.55	0.50
4	4.RI.H4	4	1	0	0	2	0	1	0.73	0.60	0.51	0.42
	4.RI.I3	4	1	2	0	1	0	0	0.83	0.55	0.52	0.49
	4.RI.L1	4	2	3	1	3	0	0	0.71	0.47	0.55	0.53
	4.RL.I1	4	0	0	0	1	0	0	0.75	0.57	0.54	0.47
	4.RL.K2	4	0	1	0	1	0	0	0.78	0.52	0.53	0.42
	4.RL.L1	4	3	2	0	1	0	1	0.75	0.61	0.54	0.56
	4.RWL.H2	4	0	0	0	2	0	0	0.78	0.49	0.53	0.39
	4.RWL.I2	4	1	2	0	1	0	1	0.77	0.54	0.52	0.50
	4.RWL.J1	2	1	1	0	0	0	0	0.76	0.42	0.34	0.36
5	5.RI.C4	4	1	1	0	1	0	0	0.79	0.51	0.46	0.43
	5.RI.D5	2	2	1	0	1	0	0	0.78	0.37	0.35	0.38
	5.RI.E2	4	2	1	0	3	0	0	0.76	0.47	0.50	0.41
	5.RL.B1	4	0	1	0	1	0	0	0.81	0.62	0.62	0.58
	5.RL.C2	4	1	0	0	0	0	0	0.82	0.57	0.51	0.48
	5.RL.D1	4	0	1	0	2	0	0	0.79	0.52	0.52	0.41
	5.RWL.A2	6	2	1	0	3	1	1	0.75	0.55	0.53	0.48
6	6.RI.B4	2	1	1	0	1	0	0	0.85	0.41	0.41	0.36
	6.RI.C2	2	1	0	0	0	0	0	0.87	0.74	0.42	0.46
	6.RI.G4	4	2	0	0	1	0	0	0.80	0.58	0.44	0.42
	6.RI.G6	4	1	0	0	0	0	0	0.78	0.62	0.44	0.41
	6.RL.B2	4	1	1	0	2	0	0	0.79	0.71	0.68	0.58
	6.RL.B3	4	1	1	0	2	0	0	0.83	0.61	0.61	0.45
	6.RL.C3	4	0	1	0	1	0	0	0.81	0.57	0.51	0.46
	6.RWL.A1	4	1	1	0	2	2	0	0.76	0.69	0.67	0.61

	6.RWL.C1	4	1	3	0	4	0	0	0.79	0.54	0.65	0.57

Table 2. Summary of Tier Flags by Content Standard in ELA, Pilot 1 (continued)

Grade	CCC	Number of Passages	Tier 4			Tier 3		Tier 2	Tier 1 Mean p-value	Tier 2 Mean p-value	Tier 3 Mean p-value	Tier 4 Mean p-value
			> Tier 3	> Tier 2	> Tier 1	> Tier 2	> Tier 1	> Tier 1				
7	7.RI.J1	4	2	1	0	1	0	0	0.80	0.48	0.48	0.46
	7.RI.J5	4	3	3	0	3	0	0	0.82	0.45	0.51	0.49
	7.RI.K4	4	2	2	0	1	0	0	0.76	0.48	0.45	0.44
	7.RI.L1	2	1	2	0	1	0	0	0.75	0.40	0.46	0.47
	7.RL.I2	4	1	2	0	1	0	0	0.76	0.56	0.53	0.49
	7.RL.J1	4	2	0	1	1	1	2	0.72	0.72	0.65	0.53
	7.RWL.G1	8	3	2	0	3	2	1	0.72	0.57	0.51	0.52
8	8.RI.J1	4	2	2	0	2	0	0	0.77	0.53	0.51	0.45
	8.RI.K2	4	3	3	0	3	0	0	0.80	0.39	0.46	0.46
	8.RI.K4	4	1	2	0	1	0	0	0.81	0.56	0.54	0.41
	8.RI.L1	2	0	1	0	1	0	0	0.81	0.57	0.56	0.49
	8.RL.I2	4	2	0	0	0	0	1	0.78	0.61	0.50	0.48
	8.RL.J2	4	2	3	0	3	0	0	0.79	0.56	0.59	0.62
	8.RWL.G1	4	2	1	1	1	1	3	0.63	0.66	0.52	0.57
	8.RWL.I1	4	1	0	0	1	1	0	0.76	0.54	0.51	0.45
11	1112.RI.B1	4	2	0	0	0	0	0	0.84	0.51	0.42	0.39
	1112.RI.B5	4	1	1	0	1	0	0	0.82	0.55	0.49	0.42
	1112.RI.D1	4	0	2	0	3	1	0	0.82	0.55	0.60	0.47
	1112.RI.E1	2	0	1	0	2	1	0	0.85	0.50	0.71	0.52
	1112.RL.B1	4	2	2	0	1	0	0	0.81	0.64	0.52	0.55
	1112.RL.D1	4	1	1	0	3	0	0	0.81	0.67	0.65	0.62
	1112.RWL.B1	4	2	2	0	1	0	1	0.87	0.64	0.61	0.61
	1112.RWL.C3	4	1	1	0	0	1	1	0.75	0.64	0.59	0.49

Table 3. Summary of Tier Flags by Content Standard in Mathematics, Pilot 1

Grade	CCC	Number of Item Families	Tier 4			Tier 3		Tier 2	Tier 1 Mean p-value	Tier 2 Mean p-value	Tier 3 Mean p-value	Tier 4 Mean p-value
			> Tier 3	> Tier 2	> Tier 1	> Tier 2	> Tier 1	> Tier 1				
3	3.DPS.1G1	4	3	1	0	0	0	0	0.71	0.48	0.43	0.48
	3.GM.1I1	4	2	2	0	2	0	0	0.72	0.57	0.60	0.57
	3.ME.1D2	4	1	0	0	1	0	0	0.66	0.52	0.39	0.30
	3.NO.1J3	4	2	3	1	3	0	0	0.62	0.40	0.42	0.44
	3.NO.1L3	4	4	4	1	2	0	0	0.70	0.35	0.40	0.55
	3.NO.2C1	5	3	3	0	2	1	0	0.73	0.38	0.42	0.38
	3.NO.2D3	4	3	1	0	0	0	0	0.65	0.48	0.34	0.43
	3.NO.2E1	4	1	0	0	1	0	1	0.70	0.65	0.45	0.32
	3.PRF.2D1	5	1	0	0	0	1	4	0.55	0.63	0.39	0.35
	3.SE.1G1	4	2	0	0	0	0	0	0.68	0.53	0.38	0.38
4	4.DPS.1G3	5	1	0	0	2	0	0	0.78	0.24	0.18	0.12
	4.GM.1H2	4	0	0	0	0	0	0	0.76	0.62	0.48	0.32
	4.ME.1G2	4	3	2	0	1	0	0	0.73	0.49	0.34	0.40
	4.NO.1J5	4	1	1	0	1	0	1	0.61	0.44	0.39	0.35
	4.NO.1M1	4	1	3	0	3	0	0	0.69	0.26	0.34	0.28
	4.NO.1N2	4	1	0	1	2	4	3	0.39	0.47	0.47	0.27
	4.NO.2D7	4	1	1	0	2	0	0	0.71	0.44	0.45	0.39
	4.NO.2E2	4	0	0	0	3	0	0	0.65	0.36	0.34	0.29
	4.PRF.1E3	5	3	0	0	2	1	1	0.54	0.41	0.34	0.32
	4.SE.1G2	4	1	1	0	2	0	0	0.70	0.32	0.38	0.30
5	5.GM.1C3	5	2	1	0	1	0	0	0.63	0.40	0.22	0.23
	5.ME.1B2	5	2	1	0	2	0	1	0.53	0.38	0.35	0.25
	5.ME.2A1	4	4	2	0	1	0	0	0.81	0.30	0.25	0.34
	5.NO.1B1	4	2	2	0	3	0	0	0.75	0.37	0.39	0.41
	5.NO.1B4	4	1	0	0	3	0	1	0.69	0.52	0.43	0.36
	5.NO.2A5	4	2	2	0	2	0	0	0.74	0.33	0.34	0.41
	5.NO.2C1	4	3	1	0	0	0	0	0.72	0.43	0.30	0.37
	5.NO.2C2	4	3	1	0	0	0	0	0.62	0.44	0.28	0.34
	5.PRF.1A1	4	0	1	0	2	1	0	0.58	0.46	0.46	0.36
	5.PRF.2B1	4	2	1	0	1	0	0	0.67	0.37	0.37	0.32

Table 3. Summary of Tier Flags by Content Standard in Mathematics, Pilot 1 (continued)

Grade	CCC	Number of Item Families	Tier 4			Tier 3		Tier 2	Tier 1 Mean p-value	Tier 2 Mean p-value	Tier 3 Mean p-value	Tier 4 Mean p-value
			> Tier 3	> Tier 2	> Tier 1	> Tier 2	> Tier 1	> Tier 1				
6	6.DPS.1D3	4	1	1	0	1	0	0	0.71	0.44	0.45	0.41
	6.GM.1D1	4	3	0	0	1	0	0	0.82	0.49	0.42	0.41
	6.ME.2A2	4	0	0	0	0	0	1	0.68	0.63	0.43	0.30
	6.NO.1D2	4	2	1	1	1	1	1	0.71	0.67	0.59	0.64
	6.NO.1D4	4	0	0	0	2	0	0	0.68	0.44	0.40	0.34
	6.NO.1F1	5	1	0	0	0	0	0	0.77	0.60	0.54	0.37
	6.NO.2A6	5	1	1	0	4	0	0	0.70	0.33	0.43	0.27
	6.NO.2C3	4	2	2	3	2	3	2	0.55	0.57	0.56	0.56
	6.PRF.1C1	4	3	2	0	2	0	1	0.72	0.57	0.52	0.56
	6.PRF.1D1	4	0	0	0	0	0	1	0.69	0.66	0.45	0.38
7	7.DPS.1K1	4	0	0	0	0	0	1	0.78	0.58	0.51	0.31
	7.GM.1H2	5	1	1	1	0	0	1	0.66	0.57	0.46	0.35
	7.ME.2D1	4	0	0	0	3	0	0	0.68	0.37	0.45	0.28
	7.NO.2F1	4	0	1	0	1	0	1	0.77	0.55	0.43	0.34
	7.NO.2F2	4	1	0	0	0	0	0	0.76	0.58	0.35	0.26
	7.NO.2F6	4	1	2	0	2	0	0	0.80	0.42	0.40	0.43
	7.NO.2I1	5	5	5	1	4	0	0	0.64	0.39	0.40	0.53
	7.NO.2I2	4	3	3	0	3	0	0	0.60	0.37	0.41	0.40
	7.PRF.1F1	4	3	2	1	1	0	0	0.59	0.46	0.34	0.44
	7.PRF.1G2	4	0	1	0	3	1	0	0.64	0.44	0.50	0.40
8	8.DPS.1H1	4	4	2	0	0	0	0	0.77	0.54	0.44	0.52
	8.DPS.1K2	4	2	2	1	2	0	1	0.62	0.44	0.43	0.44
	8.GM.1G1	4	2	3	2	4	1	0	0.58	0.40	0.57	0.52
	8.ME.1E1	4	1	0	0	0	0	0	0.71	0.48	0.41	0.37
	8.ME.2D2	4	2	0	0	1	0	0	0.67	0.51	0.46	0.39
	8.NO.1K3	5	2	3	0	4	0	0	0.74	0.38	0.49	0.45
	8.PRF.1E2	5	3	3	1	2	0	0	0.56	0.44	0.37	0.39
	8.PRF.1F2	4	0	0	0	0	0	0	0.69	0.51	0.43	0.38
	8.PRF.1G3	4	1	3	1	3	1	0	0.51	0.38	0.45	0.44
	8.PRF.2E2	4	3	1	0	1	0	0	0.68	0.47	0.32	0.32

Table 3. Summary of Tier Flags by Content Standard in Mathematics, Pilot 1 (continued)

Grade	CCC	Number of Item Families	Tier 4			Tier 3		Tier 2	Tier 1 Mean p-value	Tier 2 Mean p-value	Tier 3 Mean p-value	Tier 4 Mean p-value
			> Tier 3	> Tier 2	> Tier 1	> Tier 2	> Tier 1	> Tier 1				
11	H.DPS.1B1	4	2	3	0	1	1	1	0.50	0.39	0.41	0.41
	H.DPS.1C1	5	1	2	0	2	0	0	0.78	0.47	0.47	0.42
	H.GM.1B1	4	1	2	0	2	0	0	0.66	0.32	0.33	0.27
	H.ME.1A2	4	0	0	0	2	0	0	0.71	0.46	0.43	0.22
	H.ME.1B2	4	1	0	0	2	0	0	0.70	0.48	0.44	0.33
	H.NO.1A1	5	3	1	1	1	1	2	0.52	0.50	0.33	0.37
	H.PRF.1C1	4	0	1	3	1	3	3	0.45	0.52	0.48	0.39
	H.PRF.2B1	4	2	3	0	1	0	1	0.72	0.44	0.36	0.41
	H.PRF.2B2	4	2	2	1	2	0	0	0.68	0.59	0.50	0.55
	H.PRF.2C1	4	3	1	0	1	0	0	0.73	0.48	0.41	0.46

Table 4. Mean p-values by Grade and Tier for Mathematics, Pilot 2

Grade	Tier	Items (Points)	Mean p-value	SD
3	1	20	0.66	0.07
	2	35	0.50	0.11
	3	35	0.43	0.09
	4	10	0.40	0.10
4	1	20	0.65	0.12
	2	35	0.40	0.12
	3	35	0.39	0.10
	4	10	0.35	0.10
5	1	20	0.70	0.10
	2	35	0.45	0.09
	3	35	0.34	0.08
	4	10	0.38	0.09
6	1	20	0.70	0.07
	2	35	0.50	0.12
	3	35	0.44	0.09
	4	10	0.42	0.13
7	1	20	0.71	0.09
	2	35	0.49	0.11
	3	35	0.42	0.07
	4	10	0.36	0.10
8	1	20	0.66	0.14
	2	35	0.46	0.11
	3	35	0.45	0.10
	4	10	0.37	0.09
11	1	20	0.67	0.10
	2	35	0.46	0.08
	3	35	0.44	0.07
	4	10	0.39	0.08

Table 5. Family-Tier Evaluation Content Review Codes

Grade	Family	Depth of Knowledge	Tier	p-value (Pilot 1)	p-value (Pilot 2)	Content Review Codes					
						Reasonable pattern of performance	Item Structure Issues	Rendering (Scrolling) Issues	Missing OTL	No Explanation	Error in Metadata
3	44	DOK 3	1	0.69	0.54						
		DOK 3	2	0.37	0.45						
		DOK 3	3	0.40	0.35		X				
		DOK 4	4	0.29	0.42						
3	82	DOK 3	1	0.76	0.74	X					
		DOK 3	2	0.75	0.60	X					
		DOK 5	3	0.39	0.39	X					
		DOK 5	4	0.27	0.28	X					
3	91	DOK 3	1	0.59	0.59						
		DOK 3	2	0.65	0.58						
		DOK 3	3	0.38	0.31		X	X			
		DOK 3	4	0.30	0.43		X				
3	103	DOK 2	1	0.73	0.74						
		DOK 3	2	0.50	0.46						
		DOK 3	3	0.43	0.31		X	X			
		DOK 3	4	0.41	0.40						
4	42	DOK 2	1	0.57	0.61						
		DOK 3	2	0.43	0.49					X	
		DOK 3	3	0.42	0.40					X	
		DOK 3	4	0.44	0.47					X	
4	52	DOK 2	1	0.78	0.70						
		DOK 2	2	0.27	0.21						X
		DOK 2	3	0.44	0.31						
		DOK 3	4	0.36	0.34						

Grade	Family	Depth of Knowledge	Tier	p-value (Pilot 1)	p-value (Pilot 2)	Content Review Codes					
						Reasonable pattern of performance	Content Design Issue	Rendering (Scrolling)	Missing OTL	No Explanation	Error in Metadata
4	63	DOK 2	1	0.58	0.49						
		DOK 3	2	0.49	0.40			X			
		DOK 3	3	0.62	0.41						
		DOK 3	4	0.34	0.35						
4	72	DOK 3	1	0.70	0.80						
		DOK 3	2	0.54	0.47			X			
		DOK 5	3	0.52	0.54						
		DOK 5	4	0.47	0.39						
5	23	DOK 2	1	0.80	0.76						
		DOK 3	2	0.35	0.36				X		
		DOK 4	3	0.23	0.37				X		
		DOK 5	4	0.52	0.40				X		
5	31	DOK 4	1	0.78	0.75						
		DOK 5	2	0.23	0.29	X	X		X		
		DOK 5	3	0.34	0.23	X	X		X		
		DOK 5	4	0.47	0.51	X	X		X		
5	33	DOK 3	1	0.78	0.72						
		DOK 5	2	0.37	0.51						
		DOK 5	3	0.25	0.34					X	
		DOK 5	4	0.44	0.32						
6	42	DOK 3	2	0.59	0.59						
		DOK 3	2	0.48	0.53						
		DOK 3	3	0.61	0.52						
		DOK 4	4	0.52	0.67	X	X		X		
6	43	DOK 3	1	0.61	0.55						
		DOK 3	2	0.30	0.46						
		DOK 4	3	0.37	0.25	X	X		X		
		DOK 4	4	0.30	0.31						

Grad	Famil	Depth of	Tier	p-	p-	Content Review Codes

e	y	Knowledge		value (Pilot 1)	value (Pilot 2)	Reasonable pattern of performance	Content Design Issue	Rendering (Scrolling)	Missing OTL	No Explanation	Error in Metadata
7	63	DOK 4	1	0.52	0.54	X					
		DOK 5	2	0.44	0.48	X					
		DOK 5	3	0.37	0.42	X					
		DOK 5	4	0.25	0.23	X					
8	14	DOK 4	1	0.81	0.74						
		DOK 3	2	0.54	0.36		X	X			
		DOK 3	3	0.49	0.46						
		DOK 3	4	0.49	0.47		X				
8	33	DOK 3	1	0.35	0.90					X	
		DOK 2	2	0.29	0.65					X	
		DOK 5	3	0.59	0.53					X	
		DOK 5	4	0.43	0.32					X	
8	102	DOK 4	1	0.70	0.61						
		DOK 4	2	0.57	0.53						
		DOK 4	3	0.27	0.18						
		DOK 3	4	0.23	0.19	X	X				
11	12	DOK 4	1	0.50	0.51	X	X				
		DOK 4	2	0.52	0.30		X	X			
		DOK 4	3	0.40	0.44		X				
		DOK 5	4	0.31	0.42		X				
11	61	DOK 3	1	0.70	0.67						
		DOK 4	2	0.62	0.53		X				
		DOK 5	3	0.55	0.55		X				
		DOK 5	4	0.45	0.55	X			X		