Construct Maintenance When Varying Accessibility Characteristics

Presented at the annual meeting of the National Council on Measurement in Education

April 18, 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

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, Supports, Design Features, Universal Design for Assessment

Abstract

Using a principled approach to test design including evidence-centered design (Mislevy, 1996; Pellegrino, Chudowsky, & Glaser, 2001) and universal design (Rose, Meyer, & Hitchcock, 2005), the study investigated how academic constructs can be maintained while item features vary to meet the diverse needs of a target population. The National Center and State Collaborative (NCSC) developed an alternate assessment based on alternate achievement standards (AA-AAS) for students with the most significant cognitive disabilities. The assessment's development was guided by a theory of action that incorporated instructional context, assessment design, intended score interpretation and use, and intended long-term student outcomes. We assumed that an assessment for students with significant cognitive disabilities must be based on that which they have had a comparable opportunity to learn (Salvia & Ysseldyke, 1998). Items, developed for content targets, were intended to retain an equivalent construct while varying complexity and scaffolding to address student access. These varying features were built into item specification. Study question was, *To what extent did tiers of items designed to measure specific content targets while varying complexity and scaffolding features demonstrate evidence to support a singular construct?*

Data from two item trials in spring and fall 2014 were used with representative samples of 5200 and 6000 students, respectively. Study analyses focused on identifying instances when items within and across families displayed ordinal patterns and whether evidence supported a singular construct in terms of model fit and dimensionality analysis. Consistent with survey results completed by test administrators indicated that, in some instances, students were not taught the assessed content, results of the NCSC AA-AAS must be interpreted within the context

of student opportunity to learn grade-level CCSS academic content as well as the newness of the assessment experience for both the student and the test administrator (e.g., the assessed content; online assessment platform delivery and item presentation; and student engagement and fatigue). With acknowledgement of these identified stipulations, evidence generated by the study suggested that items generally performed as expected across grades and content standards. However, when using IRT, these patterns were not as pronounced for all grade levels (e.g., grade 4). Furthermore, review of mean discrimination suggested that not all tiers demonstrate the same degree of model fit. Tier patterns of discrimination with each content standard suggested that some content standards may contribute more to construct stability.

Construct Maintenance When Varying Accessibility Characteristics

Introduction

The National Center and State Collaborative (NCSC) developed an alternate assessment based on alternate achievement standards (AA-AAS) for students with the most significant cognitive disabilities. The assessment's development was guided by a theory of action that incorporated instructional context, assessment design, intended score interpretation and use, and intended long-term student outcomes. NCSC approached the 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. We assumed that an assessment for students with significant cognitive disabilities must be based on that which they have had a comparable opportunity to learn (Salvia & Ysseldyke, 1998).

Using a principled approach to test design based on evidence-centered design (ECD; Mislevy, 1996), universal design (UDL; Rose, Meyer, & Hitchcock, 2005) and the work of the Committee on Assessments that resulted in the book *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), Design Patterns and Task Templates (tools built into the design process that serve as precursors to item development) were developed to serve as item specifications. The resulting Design Patterns and Task Templates served as the mechanism by which varying levels of content difficulty/complexity were implemented in a group, or *family*, of assessment items measuring the same particular aspect of the core academic content in Mathematics, Reading, and Writing. Each Task Template was designed to create four *tiers* of items in each family, all aligned and written to the same content target. The item pool was designed to target the full range of learner characteristics within the target population, and the approach prompted the development of an item pool that reflected the full range of student ability. This integrated methodology used an assessment design process that incorporated the assumption of interaction between content, task, and learner characteristics in the creation of assessment items.

The NCSC principled design approach to item development was a highly structured and complex process for designing and implementing alternate assessments as envisioned under an ECD framework for summative purposes (Mislevy, 1994; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003). Under the NCSC approach, claims were first developed to the intended content area. Next, task models were developed that provided evidence about the claims. Design Patterns and Task Templates were then created from the student cognition model to produce a framework for replicating the assessment tasks (items). Finally, the NCSC item design process included several iterations of review to inform multiple decision points to ensure that the Design Patterns and Task Templates best reflected the principles for the assessed population. The in-depth, iterative process was consistent with the assumption that, if the Design Patterns and Task Templates were initially designed with adherence to accessibility issues, efficiencies in the item development process could be realized.

Design specification information in the Task Templates for NCSC (Table 1) included:

1. Decisions regarding specific content to assess in a task;
2. Variable features selected for attaining the appropriate level of complexity, depth of knowledge, scope, and degree of support for the task; and

3. Variable features selected to support the multiple means of representation, expression, and engagement of students (as operationalized in the six principles of UDL) (Rose & Meyer, 2006).

Consistent with the NCSC theory of action, we defined *access* as students' having a way of getting at the test construct in order to demonstrate their knowledge, skill and ability. Access should not to be confused with performance on the test. Rather, access is the idea that, principled design approaches can allow for minimizing barriers caused by construct-irrelevant variance (Sireci, Scarpati, & Li, 2005).

Given the fact that each family of items is intended (a) to contribute to the same test construct and (b) to allow for access to that construct for all diverse learners in the target population, this study aimed to investigate how the test construct can be maintained when accessibility needs are incorporated into the design specifications. The study question was, *To what extent did tiers of items designed to measure specific content targets, while varying complexity and scaffolding features, demonstrate evidence to support a singular construct?*

Theoretical Framework

A central tenet of universal design for assessment (Thompson, Johnstone, & Thurlow, 2002) is that test constructs are clearly defined. Consistent with general expectations of fairness in testing scenarios (AERA/APA/NCME, 2014), AA-AAS for students with the most significant cognitive disabilities provide students opportunity to demonstrate grade-level academic knowledge and skill. To do so, the measurement of the test construct must not vary from student to student, despite the array of diverse learner characteristics which students in the target population display. These learner characteristics necessarily prompt the use of diverse communicative and cognitive access features, including assistive technologies (e.g., screen readers), alternate response modes (e.g., eye gaze), or specific accommodations (e.g., alternate setting). Likewise, the item bank, having had UDL applied, must also represent the full range of student performance for the target population. Therefore, students' access to the test construct is dependent on the clarity of the construct definition and the relative success of the test instrument itself to reliability measure that construct with the target population.

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 (Pellegrino, Chudowsky, & Glaser, 2001). The NCSC target population is heterogeneous both in terms of demographic characteristics and learner characteristics (Towles-Reeves, Kearns, Flowers, Hart, Kerbel, Kleinert, Quenemoen, & Thurlow, 2012). Therefore, the assessment must address how students access test content and constructs across all relevant subgroups. The ECD process includes the *a priori* development of claims and rationales as representation of student cognition (Pellegrino, et al., 2001), and in the case of NCSC, included families of items aligned to the same content standard and ostensibly testing the same construct with varying features (e.g., scaffolding) and complexity in order to address access to that construct (Table 1).

NCSC's principled design 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). Building on the SRI PADI system to ensure fidelity of implementation of the design specifications, a NCSC goal then was to maintain a single construct across a group of items aligned to a given test standard while systematically varying the items' complexity and scaffolding/accessibility features.

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, ELA). 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 (i.e., depth of knowledge), and

(c) provide features in a given family of items to span the access needs of the full population.

Based on these articulated models of learning, the structure of the test specifications incorporate concepts of access to academic content/construct, cognitive complexity, and language complexity. Items were therefore 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. A second category of information consisted of a detailed description of the suite of four exemplar assessment items of graduated complexity in the template and includes:

1. Scripted information that will be communicated to the student, including specific directives/instructions for the examiner,
2. Stimulus materials that will be presented to the student,
3. Response options that will be presented to the student,
4. The correct response, and
5. Materials required for examiners to administer the task.

Each task contains a set of four items (Task Family) that vary systematically in complexity of the content standard-based item. Additionally, the NCSC assessment blueprints incorporated tiers by specifying the marginal percentage of tiered items per content standard.

The item specifications were defined in four levels, termed *tiers*, 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 could be translated so that students at different levels of functioning or communication would be able to interact with the construct. The item family therefore, provides an avenue for supporting access to the assessment for students with unique learning needs.

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. Investigation of the statistical functioning of item tiers within and across families was deemed needed in order 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 and are able to do academically and at grade level.

Method

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). Table 2 contains descriptive statistics based on items from the Pilot 2 Learner Characteristics Inventory (LCI; Towles-Reeves, et al., 2012) administered in both item trials; overall descriptive statistics for demographic and LCI variables were similar to those for Pilot 1.

For Pilot 1, eight linear forms per grade and content area were developed. In order to field as many newly-developed items as possible, minimal numbers of common items were included in mathematics, though the number of common items varied by form and were not representative of the content blueprint. Mathematics forms consisted of three sessions (Session 1, Session 2, and Session 3) with approximately 10 items per session. Each form contained a mix of items, designed to cover a mix of all tiers and cover the overall operational blueprint.

In Pilot 2, the Mathematics pilot forms consisted of two sessions with 20 items per session. Session 1 was a common (anchor) session across all forms within a grade. Items across both Session 1 and Session 2 consisted of a mix of tiers. As a result of Pilot 1 quantitative and qualitative data and observations, tier representation was closely attended to for the Pilot 2 pool. Mathematics tier distributions were as follows: Tier 1, 20% representation; Tier 2, 35% representation; Tier 3, 35% representation, Tier 4, 10% representation. These tier percentages were determined through input from TAC and NCSC state and vendor partners. In addition, test forms were selected to be similar in difficulty.

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 Mathematics item trials were used to conduct preliminary descriptive data analysis of classical item statistics and to establish whether the baseline score distributions were normally distributed. Rasch IRT calibration was conducted for Pilot 2 using WINSTEPS software (Linacre, 2002). In the Pilot 2 test design phase, a minimum sample size of 100 per item was determined necessary for calibration. Nearly all forms reached the minimum target sample size of 100 per item. Concurrent calibration was used to place items on the same scale. Session 1 items were common across all forms, and they were used to anchor the Session 2 items to a common scale. Model-data fit was monitored using Mean Square (MS) infit and MS outfit statistics, which indicate the degree of accuracy and predictability with which the data fit the IRT model (Linacre, 2002). To assess dimensionality, WINSTEPS includes a principal component analysis of the residual variation that is used to assess the unidimensionality assumption.

Study analyses focused on identifying instances when items or families displayed ordinal patterns, from lowest tier to highest, and then whether evidence supported a singular construct in terms of model fit and dimensionality analysis. First, we looked at "tier reversals," illustrated here using Pilot 1 Mathematics items. Items were flagged for instances in which their tier and difficulty were reversed (i.e., item designed at a lower tier had a higher difficulty than items with adjacent tier designations), indicating a departure from the original construct design. It was expected 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. Table 3 presents the number of item families measuring a given content standard, or Core Content Connector (CCC), that contained a tier reversal flag. Next, we reviewed IRT item measures and discrimination for evidence of model fit by item tier, summarizing the measure and discrimination by grade and by tier in Table 7.

Results

Using Pilot 1, tier reversal results suggest that tiers functioned somewhat consistently with empirical item difficulty in that means presented an ordinal pattern, lowest to highest (Table 3). In almost all cases, Tier 1 items were easier than the other items in the family, and mean tier p-values were consistently ordinal. However, within individual item families, tiers were reversed more heavily in Tiers 2-4.

Data from Pilot 2, the fall item trial, were used for IRT analyses. First, Table 4 summarizes p-values by tier. Similar to the preliminary tier reversal analysis from Pilot 1, p-values were ordinal across grades and families. Table 5 summarizes Rasch item measures, and Table 6 contains a summary of the principal component analysis of residual variation, as produced by WINSTEPS. The first column of the table indicates the grade. The second column of the table indicates the number of items, and the remaining three columns indicate the percent of variation associated with each dimension. As expected, the first dimension associated with the Rasch model accounts for the majority of score variation.

Results illustrate ordinal patterns across grades for mean measure by tier (Table 7); Figure 1 demonstrates this pattern across the grades. Some exceptions to the pattern were noted. For example, in grade 5, mean Rasch measures for Tiers 3 and 4 were reversed (0.51 and 0.37, respectively for Tiers 3 and 4). Also, Tier 4 discrimination differed from 1.00 more than other tiers, indicating that Tier 4 items were not fitting the Rasch model as well as other tiers.

Table 8 presents IRT results by content standard. Here discrimination is sometimes relatively similar across tiers for a given standard, whereas for other standards, discrimination varies. For example, mean discrimination across the four tiers for Standard H.PRF.2b1 ranged from 1.05 to 1.33 (difference=0.28), whereas mean discrimination for H.ME.1a2 ranged from 0.31 to 1.41 (difference=1.10).

Discussion

Evidence generated by the study suggests that items generally demonstrate ordinal patterns by tier across grades and content standards (Figure 1). Review of mean discrimination estimates suggests that not all tiers demonstrate the same degree of Rasch model fit. Tier patterns of discrimination with each content standard were even less distinct and suggest that some content standards may contribute more or less to the maintenance of the overall construct across tiers.

Limitations of the study included numerous and complex assessment challenges. For example, an end-of-test-survey of test administrators revealed that tested content was not always addressed in the classroom, affecting students' opportunity to learn. Furthermore, the relative newness of the online testing platform and Common Core-aligned test content contributed to questions about opportunity to learn and construct irrelevant variance thrroughout the testing process. Additionally, the fact that item-level student responses were not available due to data privacy constraints limited the analyses conducted to evaluate construct (e.g., factor analysis; population invariance analyses).

Assuming opportunity to learn and testing familiarity improve over time, future analytic work could focus on strategies to better understand construct stability across content standards and families/tiers. Also, next steps could look at evidence of internal test structure through convergent and discriminant analyses. Generally, it would be expected that, for example, Reading item measures would be highly correlated with Reading scores and not so highly correlated with Math scores. Similarly, we would expect that items within a family or within a content standard would be more highly related to each other than to other items within the same tier if they are measuring the same construct (measuring content construct not a "tier construct"). Additional work should examine correlations and exploratory factor analyses of interrelationships among item tiers and item families, item tiers and content standards, and item tiers and subject areas. It is expected that relationships would be stronger within item families or content areas as opposed to across item tiers. This would provide evidence of a consistent construct measured across item tiers, as specified by the design.

The study provides insight into an application of principled design approach based on ECD and UDL that endeavors to maintain a test construct while varying supports and complexity for equitable access for diverse students. Results from this study provide validity evidence of the extent to which a test construct can be maintained across different levels of item supports and for diverse examinee subgroups. Increased standardization of test administration protocols coupled with tests designed to meet the access needs of the student population should improve reliability of scores and reduce the risk that inappropriate supports are used.

References

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

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.

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. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 24*(4), 6-20.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

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.

Rose, D., Meyer, A., & Hitchcock, C. (Eds.). (2005). *The universally designed classroom*. Cambridge, MA: Harvard Education Press.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment*. Boston: Houghton-Mifflin.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: an analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457-490.

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.

Thompson, S.J., Johnstone, C.J., & Thurlow, M.L. (2002, June). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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. *Core content connector (content standard); **Knowledge/skill/ability of CCC

Table 2. Student Sample Learner Characteristics Inventory Descriptive Statistics Overall

Variable	Category	N	%
Primary Disability	Intellectual Disability/Mental Retardation (includes Mild, Moderate, and Profound)	2877	47.02
	Other Health Impairment	346	5.65
	Orthopedic	93	1.52
	Deaf	27	0.44
	Specific Learning Disability	276	4.51
	Multiple disabilities	629	10.28
	Autism	1554	25.40
	Speech/language Impairment	69	1.13
	Hearing Impairment	41	0.67
	Visual Impairment	30	0.49
	Traumatic Brain Injury	44	0.72
	Emotional Disability	58	0.95
	Deaf/Blind	5	0.08
	Other	70	1.14
EL Status	No	5147	84.29
	Yes	959	15.71
Classroom Setting	Special school	481	7.86
	Regular school, self-contained special education classroom, some special inclusion (students go to art, music, PE) but return to their special education class for most of school day.	4220	68.97
	Regular school, primarily self-contained special education classroom, some academic inclusion (students go to some general education academic classes (reading, math, science, in addition to specials) but are in general education classes less than 40% of the school day).	799	13.06
	Regular school, resource room/general education class, students receive resource room services, but are in general education classes 40% or more of the school day.	440	7.19
	Regular school, general education class inclusive/collaborative (students based in general education classes, special education services are primarily delivered in the general education classes) – at least 80% of the school day is spent in general education classes.	179	2.93

Table 2. Student Sample Learner Characteristics Inventory Descriptive Statistics Overall (continued)

Variable	Category	N	%
Expressive Communication	Uses symbolic language to communicate: Student uses verbal or written words, signs, Braille, or language-based augmentative systems to request, initiate, and respond to questions, describe things or events, and express refusal.	4739	77.45
	Uses intentional communication, but not at a symbolic language level: Student uses understandable communication through such modes as gestures, pictures, objects/textures, points, etc., to clearly express a variety of intentions.	1160	18.96
	Student communicates primarily through cries, facial expressions, change in muscle tone, etc., but no clear use of objects/textures, regularized gestures, pictures, signs, etc., to communicate.	220	3.60
Augmentative Communication System	No	5348	89.01
	Yes	660	10.99
Receptive Language	Independently follows 1-2 step directions presented through words (e.g. words may be spoken, signed, printed, or any combination) and does NOT need additional cues.	3311	54.11
	Requires additional cues (e.g., gestures, pictures, objects, or demonstrations/models) to follow 1-2 step directions.	2509	41.00
	Alerts to sensory input from another person (auditory, visual, touch, movement) BUT requires actual physical assistance to follow simple directions.	259	4.23
	Uncertain response to sensory stimuli (e.g., sound/voice; sight/gesture; touch; movement; smell).	40	0.65
Oral	No	1011	18.70
	Yes	4396	81.30
Braille	No	5464	99.53
	Yes	26	0.47
Vision	Vision within normal limits	4321	72.25
	Corrected vision within normal limits.	1421	23.76
	Low vision; uses vision for some activities of daily living.	147	2.46
	No functional use of vision for activities of daily living, or unable to determine functional use of vision.	92	1.54
Hearing	Hearing within normal limits.	5772	95.15
	Corrected hearing loss within normal limits.	65	1.07
	Hearing loss aided; but still with a significant loss.	89	1.47
	Profound loss, even with aids.	91	1.50
	Unable to determine functional use of hearing.	49	0.81
Motor	No significant motor dysfunction that requires adaptations.	5463	90.6
	Requires adaptations to support motor functioning (e.g., walker, adapted utensils, and/or keyboard).	283	4.69
	Uses wheelchair, positioning equipment, and/or assistive.	161	2.67
	Needs personal assistance for most/all motor activities.	123	2.04

Table 2. Student Sample Learner Characteristics Inventory Descriptive Statistics Overall (continued)

Variable	Category	N	%
Engagement	Initiates and sustains social interactions.	3859	63.92
	Responds with social interaction, but does not initiate or sustain social interactions.	1831	30.33
	Alerts to others.	292	4.84
	Does not alert to others.	55	0.91
Health	Attends at least 90% of school days.	5550	92.95
	Attends approximately 75% of school days; absences primarily due to health issues.	338	5.66
	Attends approximately 50% or less of school days; absences primarily due to health issues.	28	0.47
	Receives Homebound Instruction due to health issues.	9	0.15
	Highly irregular attendance or homebound instruction due to issues other than health.	46	0.77
Reading	Reads fluently with critical understanding in print or Braille (e.g., to differentiate fact/opinion, point of view, emotional response, etc.).	174	2.84
	Reads fluently with basic (literal) understanding from paragraphs/short passages with narrative/informational texts in print or Braille.	1466	23.96
	Reads basic sight words, simple sentences, directions, bullets, and/or lists in print or Braille.	2774	45.33
	Aware of text/Braille, follows directionality, makes letter distinctions, or tells a story from the pictures that is not linked to the text.	1175	19.20
	No observable awareness of print or Braille.	530	8.66
Math	Applies computational procedures to solve real-life or routine word problems from a variety of contexts.	274	4.48
	Does computational procedures with or without a calculator.	3089	50.48
	Counts with 1:1 correspondence to at least 10, and/or makes numbered sets of items.	1791	29.27
	Counts by rote to 5.	565	9.23
	No observable awareness or use of numbers.	400	6.54

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 (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. Summary of p-value by Tier in Mathematics, Pilot 2

Grade	Tier	Items (Points)	Mean p-value	SD p-value
3	1	20	0.663	0.072
	2	35	0.497	0.114
	3	35	0.434	0.092
	4	10	0.398	0.102
4	1	20	0.653	0.119
	2	35	0.398	0.117
	3	35	0.393	0.095
	4	10	0.354	0.098
5	1	20	0.697	0.095
	2	35	0.446	0.086
	3	35	0.344	0.080
	4	10	0.375	0.093
6	1	20	0.698	0.070
	2	35	0.498	0.119
	3	35	0.440	0.094
	4	10	0.418	0.127
7	1	20	0.713	0.088
	2	35	0.491	0.112
	3	35	0.420	0.071
	4	10	0.358	0.097
8	1	20	0.662	0.142
	2	35	0.462	0.106
	3	35	0.452	0.104
	4	10	0.374	0.092
11	1	20	0.665	0.100
	2	35	0.456	0.077
	3	35	0.437	0.066
	4	10	0.394	0.078

Table 5. Pilot 2 Mathematics Item Measure Summary by Grade

Grade	Statistic	Measure	SE	InMS	InZStd	OutMS	OutZStd
3	Items	100	100	100	100	100	100
	Mean	0.00	0.19	1.00	-0.08	1.00	-0.05
	Standard Deviation	0.65	0.05	0.12	1.96	0.18	1.77
	Min	-1.40	0.10	0.66	-6.21	0.59	-5.11
	10th	-0.84	0.10	0.84	-2.69	0.77	-2.22
	25th	-0.53	0.16	0.92	-1.19	0.86	-1.21
	50th	0.08	0.22	0.99	-0.06	0.99	-0.07
	75th	0.50	0.22	1.09	1.14	1.12	0.94
	90th	0.84	0.24	1.17	2.35	1.22	2.50
	Max	1.42	0.25	1.35	4.49	1.48	3.93
4	Items	100	100	100	100	100	100
	Mean	0.00	0.18	1.00	-0.10	1.00	-0.09
	Standard Deviation	0.71	0.05	0.11	1.89	0.15	1.77
	Min	-1.71	0.09	0.79	-5.39	0.74	-4.66
	10th	-0.99	0.09	0.86	-3.00	0.81	-2.57
	25th	-0.31	0.15	0.93	-1.27	0.89	-1.28
	50th	0.14	0.19	1.00	0.08	1.00	-0.04
	75th	0.42	0.22	1.07	0.99	1.09	1.11
	90th	0.74	0.24	1.15	2.07	1.20	1.96
	Max	1.60	0.30	1.30	5.26	1.48	4.64
5	Items	100	100	100	100	100	100
	Mean	0.00	0.20	1.00	-0.01	1.00	-0.04
	Standard Deviation	0.72	0.06	0.08	1.10	0.12	1.23
	Min	-1.74	0.10	0.82	-3.41	0.77	-3.04
	10th	-1.13	0.10	0.92	-1.32	0.86	-1.39
	25th	-0.34	0.15	0.94	-0.67	0.92	-0.85
	50th	0.11	0.22	0.99	-0.03	0.99	-0.09
	75th	0.54	0.24	1.04	0.69	1.06	0.67
	90th	0.83	0.26	1.09	1.41	1.14	1.38
	Max	1.19	0.31	1.41	2.97	1.50	3.91

Table 5. Pilot 2 Mathematics Item Measure Summary by Grade (continued)

Grade	Statistic	Measure	SE	InMS	InZStd	OutMS	OutZStd
6	Items	100	100	100	100	100	100
	Mean	0.00	0.19	0.99	-0.18	0.98	-0.18
	Standard Deviation	0.68	0.05	0.11	1.69	0.15	1.55
	Min	-1.89	0.10	0.79	-3.54	0.72	-2.83
	10th	-0.95	0.10	0.87	-2.47	0.80	-2.15
	25th	-0.54	0.16	0.92	-1.34	0.87	-1.40
	50th	0.09	0.21	0.99	-0.18	0.98	-0.23
	75th	0.44	0.23	1.04	0.60	1.05	0.52
	90th	0.92	0.24	1.17	1.82	1.18	1.83
	Max	1.52	0.28	1.32	5.29	1.45	4.74
7	Items	100	100	100	100	100	100
	Mean	0.00	0.18	1.00	-0.12	0.99	-0.11
	Standard Deviation	0.72	0.05	0.10	1.53	0.16	1.54
	Min	-2.11	0.09	0.81	-2.85	0.74	-2.75
	10th	-1.10	0.09	0.89	-1.94	0.83	-1.87
	25th	-0.48	0.16	0.92	-1.19	0.88	-1.29
	50th	0.20	0.18	0.99	-0.10	0.97	-0.35
	75th	0.47	0.21	1.05	0.71	1.09	1.12
	90th	0.84	0.22	1.13	1.88	1.22	2.07
	Max	1.34	0.27	1.31	3.94	1.46	3.80
8	Items	100	100	100	100	100	100
	Mean	0.00	0.19	0.99	-0.10	0.99	-0.11
	Standard Deviation	0.70	0.06	0.08	1.35	0.16	1.25
	Min	-2.56	0.09	0.84	-3.28	0.73	-2.72
	10th	-0.86	0.10	0.91	-1.55	0.86	-1.40
	25th	-0.31	0.15	0.95	-1.06	0.91	-0.93
	50th	0.12	0.20	0.99	-0.24	0.96	-0.33
	75th	0.43	0.23	1.03	0.49	1.05	0.42
	90th	0.66	0.25	1.11	1.56	1.16	1.77
	Max	1.63	0.38	1.22	5.38	1.80	3.70
11	Items	100	100	100	100	100	100
	Mean	0.00	0.19	1.00	-0.09	0.99	-0.06
	Standard Deviation	0.57	0.06	0.10	1.64	0.15	1.51
	Min	-1.55	0.10	0.79	-3.70	0.71	-2.89
	10th	-0.95	0.10	0.89	-2.12	0.83	-1.86
	25th	-0.28	0.15	0.92	-1.08	0.88	-1.02
	50th	0.12	0.20	0.98	-0.24	0.97	-0.34
	75th	0.40	0.22	1.04	0.69	1.06	0.73
	90th	0.62	0.28	1.15	1.99	1.19	1.83
	Max	1.07	0.30	1.37	5.37	1.71	4.93

Table 6. Principal Component Analysis in Mathematics

Grade	Number of Items	Percent First Dimension	Percent Second Dimension	Percent Third Dimension
3	100	33.3	2.5	1.8
4	100	27.0	3.0	2.1
5	100	23.5	3.7	2.8
6	100	27.5	3.0	2.2
7	100	32.6	3.3	2.5
8	100	32.7	2.1	1.6
11	100	32.8	2.2	1.8

Table 7. Rasch Measure and Discrimination by Grade and Tier, Pilot 2 Mathematics

Grade	Tier	N Items	Mean Rasch Measure	SD	Min	Max	Mean Error	Mean Discrimination	Mean PLPBSE
3	1	20	-0.82	0.374	-1.40	-0.04	0.19	0.93	0.29
	2	35	0.00	0.547	-1.13	0.95	0.19	1.12	0.38
	3	35	0.32	0.478	-0.71	1.17	0.19	0.98	0.33
	4	10	0.50	0.505	-0.43	1.42	0.19	0.89	0.30
4	1	20	-0.99	0.570	-1.71	0.41	0.18	0.97	0.22
	2	35	0.22	0.566	-0.94	1.60	0.18	1.06	0.24
	3	35	0.23	0.440	-0.66	1.16	0.18	1.08	0.24
	4	10	0.39	0.499	-0.21	1.33	0.18	0.80	0.14
5	1	20	-1.12	0.468	-1.74	0.07	0.20	1.06	0.23
	2	35	0.03	0.401	-0.58	0.94	0.19	1.08	0.23
	3	35	0.51	0.401	-0.52	1.19	0.20	0.96	0.18
	4	10	0.37	0.433	-0.29	0.83	0.19	0.89	0.15
6	1	20	-0.90	0.399	-1.89	-0.11	0.19	1.07	0.28
	2	35	0.06	0.563	-1.11	1.19	0.18	1.15	0.31
	3	35	0.33	0.462	-0.53	1.52	0.18	0.94	0.23
	4	10	0.44	0.596	-0.71	1.03	0.19	1.13	0.29
7	1	20	-1.04	0.502	-2.11	-0.31	0.19	1.03	0.26
	2	35	0.03	0.529	-0.98	0.96	0.17	1.20	0.34
	3	35	0.37	0.349	-0.26	1.18	0.17	0.94	0.25
	4	10	0.69	0.443	-0.05	1.34	0.18	0.80	0.14
8	1	20	-0.87	0.752	-2.56	0.60	0.20	0.98	0.25
	2	35	0.15	0.470	-0.78	1.63	0.19	1.13	0.32
	3	35	0.19	0.488	-0.88	1.52	0.19	1.02	0.27
	4	10	0.57	0.450	-0.01	1.62	0.19	0.95	0.23
11	1	20	-0.85	0.476	-1.55	0.12	0.20	1.06	0.29
	2	35	0.14	0.370	-0.59	0.98	0.19	1.13	0.33
	3	35	0.22	0.311	-0.39	0.74	0.19	0.98	0.28

4	10	0.44	0.396	-0.22	1.07	0.19	0.74	0.16

Table 8. Rasch Measure and Discrimination by Content Standard and Tier, Pilot 2 Mathematics

Content Standard	Tier	N Grade	Mean Rasch Measure	Mean Discrimination	Mean INMSQ
H.DPS.1b1	1	2	-0.28	1.01	1.00
	2	3	0.51	0.80	1.07
	3	4	0.11	0.47	1.15
	4	1	0.31	0.45	1.16
	Total	10	0.17	0.67	1.10
H.DPS.1c1	1	2	-1.16	1.28	0.89
	2	4	0.15	1.17	0.95
	3	3	-0.03	1.16	0.95
	4	1	0.58	0.88	1.05
	Total	10	-0.12	1.16	0.95
H.GM.1b1	1	2	-0.54	0.85	1.05
	2	3	0.32	0.63	1.09
	3	4	0.50	0.95	1.02
	4	1	0.54	0.80	1.07
	Total	10	0.24	0.82	1.05
H.ME.1a2	1	2	-1.08	1.41	0.86
	2	4	0.19	1.25	0.93
	3	3	0.22	1.34	0.92
	4	1	1.07	0.31	1.37
	Total	10	0.04	1.22	0.96
H.ME.1b2	1	2	-1.15	1.19	0.94
	2	3	0.22	1.29	0.93
	3	4	0.20	0.68	1.07
	4	1	0.72	0.55	1.21
	Total	10	-0.01	0.95	1.02
H.NO.1a1	1	2	-0.29	0.93	1.01
	2	3	0.09	1.45	0.89
	3	4	0.58	1.01	1.00
	4	1	0.00	1.23	0.96
	Total	10	0.20	1.15	0.96
H.PRF.1c1	1	2	-0.64	0.74	1.05
	2	3	0.13	1.26	0.94
	3	4	0.16	0.68	1.08
	4	1	0.65	0.40	1.26
	Total	10	0.04	0.84	1.05
H.PRF.2b1	1	2	-1.41	1.11	0.94
	2	4	-0.14	0.98	1.01
	3	3	-0.14	1.48	0.90
	4	1	0.06	0.92	1.02

	Total	10	-0.37	1.15	0.97

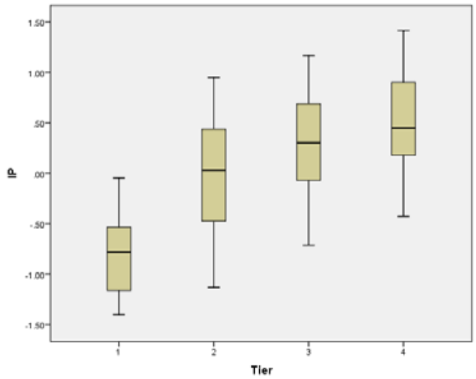
Table 8. Rasch Measure and Discrimination by Content Standard and Tier, Pilot 2 Mathematics

(continued)

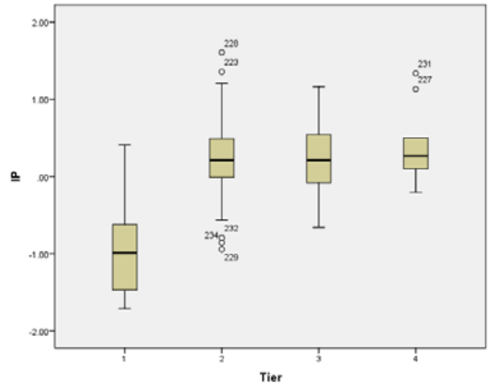
Content Standard	Tier	N Grade	Mean Rasch Measure	Mean Discrimination	Mean INMSQ
H.PRF.2b2	1	2	-1.06	1.05	0.98
	2	4	-0.34	1.33	0.93
	3	3	0.01	1.21	0.95
	4	1	-0.22	1.33	0.93
	Total	10	-0.37	1.24	0.94
H.PRF.2c1	1	2	-0.89	1.07	0.97
	2	4	0.39	1.08	0.97
	3	3	0.49	1.15	0.95
	4	1	0.69	0.56	1.19
	Total	10	0.19	1.05	0.98
Grand Total		100	0.00	1.02	1.00

Figure 1. Tier Rasch Measure Box and Whisker Plots by Grade

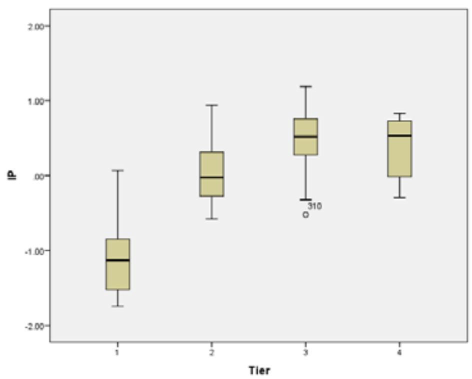
Grade 3



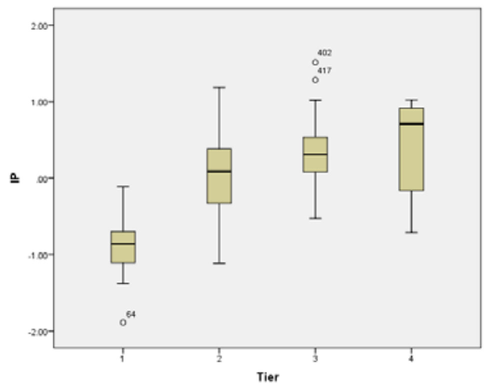
Grade 4



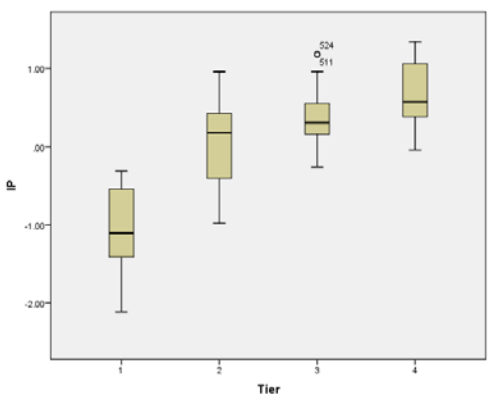
Grade 5



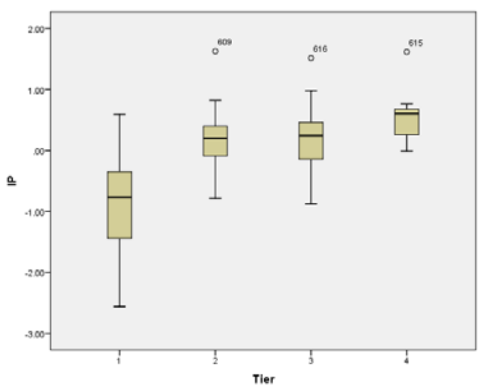
Grade 6



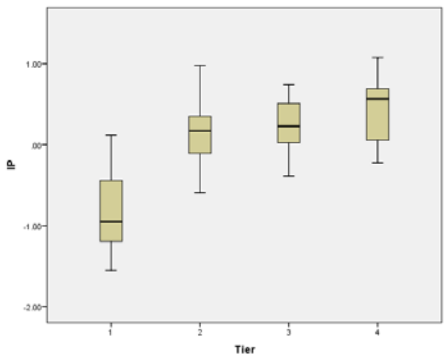
Grade 7



Grade 8



Grade 11


1.00
.00
-1.00
-2.00
IP
1
2
3
4
Tier