Savvas Mathematics Diagnostic and Screener Assessment (MSDA) Technical Report

Submitted to:

Savvas

Submitted by:

Sonya Powers
Matthew N. Gaertner
WestEd

August 2021

WestEd
WestEd.org

WestEd

excellence in research development, and service

Table of Contents

Introduction

Background

In April 2020, Savvas Learning Company LLC contracted with WestEd to develop academic screener and diagnostic assessments for grades K-8 in Mathematics. The screener and diagnostic assessments are intended to work together to identify learning challenges that require additional screening, to identify students' strengths and areas for improvement, and to provide data-driven connections to instructional supports. The first year of work focused on item development and the construction of field-test forms. This report documents elements of the assessment lifecycle including test design and development, field testing, psychometric and statistical analyses, norming studies, planning for subsequent validation studies, and reporting features.

Purpose

The Savvas Mathematics Screener Assessment is a relatively short assessment administered early in the school year to students in kindergarten through Grade 8. The Screener assesses prerequisite skills for the student's current grade. For example, the Grade 2 screener assesses the Grade 1 standards that prepare students to succeed in learning Grade 2 standards. The Screener is designed to:

- Identify serious learning challenges that require additional screening;
- Provide a snapshot of readiness for grade level instruction; and
- Indicate which Savvas Diagnostic Assessment would be the most appropriate for each student.

The Savvas Mathematics Diagnostic Assessment works in tandem with the Screener to provide more granular diagnostic information for students. The Diagnostic Assessment can be used to identify students' strengths and weaknesses relative to grade-level content and provide associated connections to instructional supports. The Diagnostic Assessment is intended to be given early in the school year, shortly after the Screener. Because it is an assessment of the on-grade level content knowledge students are expected to master by the end of the year, many students are anticipated to have knowledge gaps. The goal is to leverage student performance data from the Diagnostic to personalize instruction for each student so that gaps in knowledge and skills are addressed over the course of the school year.

The first operational administration of the Savvas Mathematics Screener and Diagnostic Assessments will be in Fall 2021. The Diagnostic Assessments were field-tested in Spring 2021 and subsequent statistical and psychometric analyses were conducted to support the Diagnostics' technical foundations. The Savvas Mathematics Screeners will be statistically linked to the Savvas Mathematics Diagnostic Assessments following the Fall 2021 administration.

Test Design and Development

Literature Review Supporting Item Development

The learning of mathematics at the elementary and middle grades forms the foundation for achievement in high school, college, and for the range of mathematical skills used in the workplace. Number recognition, counting facility, pattern recognition[1], and measurement with nonstandard units are key predictors of student mathematics performance in elementary school[2] and are essential for continued learning in secondary school. Fluency with whole numbers and fractions and aspects of geometry and measurement are foundational to algebraic learning, often considered the gateway to college success.[3] Geometry instruction should include opportunities for students to explore shapes and their attributes, spatial relationships, transformations, and visualization.[4] Measurement serves as a foundation for thinking about quantities, connecting the mathematical relationships experienced in the world to numerical expressions as students think about how a quantity relates to the attribute being measured or to the unit of measurement as they construct and analyze graphical representations of real world data.[5]

Many early elementary mathematics screeners focus on number sense. Although number sense is a critical component of students' early mathematics learning, it provides only a limited understanding of students' mathematical conceptions. In addition to number sense,

---

[1] The selected standards for the Kindergarten screener do not include patterns, although it is given as a foundational skill for student mathematical success. This is due to the absence of pattern standards in K-2 in the Common Core State Standards for Mathematics, from which the selection of standards occurs for all other grades.

[2] Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record, 115*, 1-29.

[3] National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

[4] Allsopp, D. H., Kyger, M. M., & Lovin, L. H. (2008). *Teaching mathematics meaningfully: Solutions for reaching struggling learners*. Baltimore, MD: Paul H. Brookes Publishing Co.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.
National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: The National Academies Press.

[5] Van de Walle, J. A., Karp, K. S., & Williams, J. M. B. (2007). *Elementary and middle school Mathematics: Teaching developmentally*. Boston: Pearson.

including measurement and geometry on the screener provides a broader window through which teachers may understand students' mathematical reasoning.[6] Therefore, early mathematics screeners should also assess components of measurement and geometry deemed foundational to student learning, to create a more complete picture of what students know and are able to do.[7] In consideration of the research on foundational mathematics skills and early screeners, the selection of standards for the Kindergarten screener focused on number competencies, informal measurement, and spatial reasoning.

Content Standards Assessed

WestEd and Savvas subject matter experts worked together to identify the relevant standards to be assessed at each grade level and the numbers of items to be developed for each standard for the Screener and the Diagnostic Assessments. Supported by the findings of the literature review, standards for kindergarten through grade 8 were selected from the associated grades' Common Core State Standards for Mathematics (CCSS-M)[8]. The selected content includes the following domains: Number and Operations/The Number System, Algebra and Algebraic Thinking, Geometry, and Measurement and Data. Fluency is assessed using the fluency standards in Operations and Algebraic Thinking (grades K, 1, 2, 3); Number and Operations in Base Ten (grades 2, 3, 4, 5); and The Number System. Calculator usage is permitted only for students in grades 6-8. The content limits set forth by the standards were supplemented with clarifications provided in the Partnership for Assessment of Readiness for College and Careers (PARCC) Evidence Statement Tables[9]. Each of the items was also linked to the proprietary Savvas skill spines that correspond to Savvas curriculum materials.

The Screener standards represent the most critical pre-requisite skills necessary for on-grade level success. These standards were carefully selected from the CCSS-M at the grade level below "on grade level". Student performance on these off-grade level items provides a quick indicator for teachers at the beginning of the school year of how prepared students are for on-grade level instruction. Because the CCSS-M do not include pre-kindergarten standards, California pre-kindergarten learning and development guidelines were adapted for use with the kindergarten Screener. The content limits set forth by the standards were supplemented with clarifications provided in the PARCC Evidence Statement Tables. The standards and item counts used for the Screener and Diagnostic Assessments are included in Appendix A and Appendix B, respectively.

---

[6] National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: The National Academies Press.

[7] Brendefur, J. L., Johnson, E. S., Thiede, K. W., Strother, S., & Severson, H. H. (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: the primary mathematics assessment. *Early Childhood Education Journal, 46*(2), 153–157.

[8] Retrieved from: http://www.corestandards.org/Math/

[9] Retrieved from: https://resources.newmeridiancorp.org/math-test-design/

Accessibility, Universal Design, Bias and Sensitivity

The Screener and Diagnostic Assessments were developed to be accessible for all students to support score interpretations and valid inferences about mathematics achievement for all students. Our item development guidelines and item writer training on accessibility, universal design, and bias and sensitivity helped to ensure that test items and supporting materials were as free as possible from unnecessary access barriers that might limit the demonstration of student achievement.

Specifically, all item writers received training on WestEd's accepted practices in universal design. They received additional training on methods to avoid bias and sensitive content in item development. The item development process was iterative with many rounds of review at WestEd and Savvas to ensure that the final item pool adhered to the approved guidelines.

Further, alternative text was developed for all graphics used on operational forms. Text-to-speech was also used for the majority of text included on the Savvas Mathematics Screener and Diagnostic Assessments with a few exceptions. Specifically:

- All question stems are read.
- Answer choices are read unless:
    - reading an answer choice gives away the answer,
    - an answer choice is a numerical value (e.g. "15"),
    - an answer choice is a numerical value plus a unit of measure (e.g. "15 inches"),
    - an answer choice is an algebraic expression or equation (e.g. "3x + 5 = 17"),
    - an answer choice is a table of values, or
    - an answer choice is an image that will already have alternative text.

Item Pool Development

The WestEd item development team created items and associated scoring specifications for the Savvas Assessments that meet or exceed assessment industry standards and that adhere to the *Standards for Educational and Psychological Testing*[10]. Processes and quality control procedures were implemented to support the development of Screener and Diagnostic items that:

- measure the subject-specific standards with a consistent interpretation of the standards across item writers, content specialists, and reviewers;

[10] American Educational Research Association (AERA), American Psychological Association (APA), and the National Council for Measurement in Education (NCME). (2014). Standards for Educational and Psychological Testing.

- adhere to the PARCC Evidence Statements[11] and Style Guide[12];
- meet the requirements of the item specifications and blueprints;
- include only one correct answer;
- have the specified number of answer choices that are balanced grammatically and structurally;
- have plausible, but incorrect options;
- do not clue other items;
- are developmentally appropriate for the given grade by engaging content specialists and reviewers familiar with the knowledge, skills, and abilities of students at the tested grades;
- maximize accessibility for students through the application of universal design principles;
- engage student interest using multiple item formats and interesting grade-appropriate contexts;
- are clear, concise and free from idiomatic expressions, and grammatical errors;
- exclude bias, stereotyping information, and sensitive content; and
- exhibit strong alignment to the intended construct.

Each item was documented in a standardized item template (see Appendix A), which included item metadata (e.g., standard alignment, difficulty, depth of knowledge) and item-specific scoring information (e.g., distractor rationales, scoring rubrics). Technology-enhanced item types included:

- fill in the blank, allowing for numerical or text entry;
- hotspot, with either a single response or multiple responses;
- graphic gap match, with either single-use or reusable draggers; and
- inline choice.

Savvas Sample Set Review. WestEd provided a sample of items to Savvas prior to development of the full item bank. Development of the full item bank was guided by Savvas feedback from the sample set of items, as well as the principles of Universal Design for Assessment, bias and sensitivity guidelines, and accessibility guidelines. Items were delivered in the standardized item template which included item metadata and native graphic files.

Item Review Process. All developed items underwent WestEd's systematic quality review process. Reviewers provided recommended edits and feedback to the team of item writers.

[11] https://resources.newmeridiancorp.org/math-test-design/
[12] http://parccinc.org/wp-content/uploads/2014/07/PARCCStyleGuidev2-4.pdf

This continuous feedback cycle ensured ongoing monitoring and support for item quality and inflight process improvement as item writers continued to fulfill item development requirements. After a thorough content review, a proofreader reviewed each item to:

- check spelling,
- check grammar,
- verify adherence to the style guide,
- confirm graphics were appropriate with respect to correct size, scale, and format, and
- review the presentation of the item for a computer-based administration.

All issues identified by the proofreader were reviewed and reconciled by the content lead and updates were verified by the proofreader. Edit verification between the content lead and proofreader supported quality assurance and minimized the possibility of introducing errors at this stage of item development.

Final-Eye Review. As a final step prior to Savvas review, all items underwent a final-eye review as a concluding confirmation that the items were technically sound with respect to best practices for the development of items for high-stakes assessments. Items that did not meet the final-eye standard were edited as needed and underwent an additional final-eye review, and ultimately, sign-off. During the final-eye review, the content lead confirmed that each item conforms to the following requirements:

- aligns to the assigned content standard,
- meets the requirements for assigned cognitive complexity, and difficulty levels,
- is grade-appropriate,
- has correct scoring and correct response information,
- does not contain information in the stem that clues the correct answer and does not contain information that clues the correct answer to another item,
- adheres to universal design principles and is free of bias or sensitivity issues,
- adheres to the style guide, and
- does not contain content errors.

WestEd's Director of Test Development and the Mathematics Content Managers audit the work performed by the content leads during this final-eye review. The Director of Test Development provides the final sign-off on the item set as ready for external review or operational use.

In summary, all newly developed K-8 items went through WestEd's rigorous item development and review process including two rounds of content editing, two rounds of

proofreading, and the final-eye review before the items were submitted to Savvas. Savvas content experts then completed their own round of reviews and submitted requests for revisions as needed.

Test Construction

The Savvas Mathematics Diagnostic field-test forms and operational Screener forms were developed in Fall of 2020. Early in 2021, the Diagnostic Assessments were field-tested. Field-test data were used to evaluate and select items for the operational Diagnostic Assessment. The operational Diagnostic forms were completed in spring of 2021. Both the Screener and the operational Diagnostic Assessments will be administered for the first time in Fall 2021.

Computer Administered Fixed Form Screener Assessments

The Screener is a computerized fixed form multiple-choice assessment based on a vertical articulation of the Common Core State Standards. There is one form per grade, with 20-30 items per operational form, as specified in Table 1. The Screener was not field-tested in Spring 2021; instead, the Screener assessments will be administered operationally along with the Diagnostic assessments in Fall 2021.

Table 1. Number of Screener Items on Operational Forms

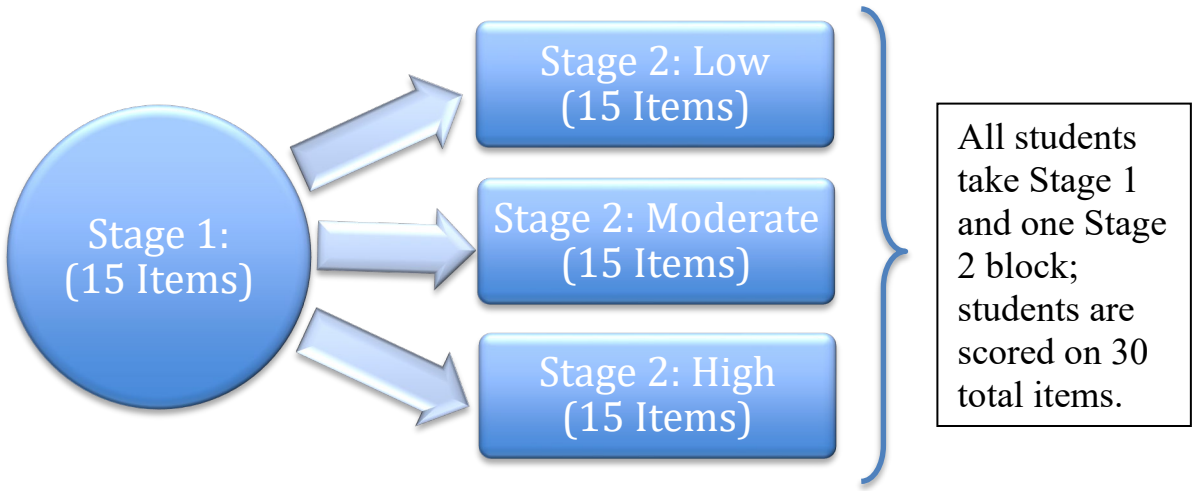
Grade	Number of Items Per Operational Screener
K	20
1	20
2	22
3	25
4	25
5	25
6	30
7	30
8	30

Computer Administered Multi-Stage Adaptive Diagnostic Assessment

The Diagnostic Assessments are multi-stage adaptive assessments designed to identify each student's strengths and weaknesses with connections to instructional supports. As indicated in Figure 1, each form contains a common 15-item routing set in Stage 1 that is used to compute a preliminary ability estimate. This is followed by a Stage 2 15-item set tailored to the Stage 1 ability estimate. Specifically, Stage 1 high, moderate, and low proficiency students are routed to high, moderate, and low difficulty Stage 2 blocks,

respectively. This allows for more precise measurement of students' knowledge and skills, targeted to their current level of mathematics achievement. Across the two item sets, each student receives a total of 30 items. The Diagnostic Assessment includes both multiple-choice and technology-enhanced item types.

Figure 1. Diagnostic Multi-Stage Design



Spring 2021 Field Test. A field test was conducted Spring 2021 to support two primary purposes. First, a vertical scale study was conducted using field-test data to place all Diagnostic items, within and across grades, on a common measurement scale. Second, test and item analyses were conducted to identify high quality items to support the selection of operational forms with superior measurement attributes. A total of 60 items are required for the operational forms including 15 Stage 1 items and 45 Stage 2 items (15 items for each block, including the Low, Moderate, and High difficulty blocks). To support the selection of high-quality forms, approximately twice as many items were field tested as were needed for the operational forms.

The field test followed a randomized form design where forms were randomly assigned to students at the classroom level (i.e., each student within the same classroom received the same, randomly assigned form). Several forms were field tested to get a large pool of items from which to develop the operational forms. Each grade's forms consisted of 15 common items and 15 unique items. The unique items were randomly assigned to forms prior to the test administration and were not based on student performance on the first 15 items. Appropriate coverage of the relevant content standards was supported for each form.

Common across-grade items in each form supported the development of a vertical scale. Specifically:

- Kindergarten forms included 6 items that are in common with the associated Grade 1 forms
    - three Kindergarten common items were included on both the Kindergarten and Grade 1 forms
    - three Grade 1 common items were included on both the Kindergarten and Grade 1 forms
- Grades 1-7 forms included three items from the associated below-grade form and three items from the associated above-grade form, along with six on-grade items used for vertical scaling. For example:
    - Grade 3 forms contained three Grade 2 common items, three Grade 4 common items, and six Grade 3 common items that were used for vertical scaling;
        - three of the Grade 3 common items were included on the Grade 2 forms
        - three of the Grade 3 common items were included on the Grade 4 forms
- Grade 8 forms included six items that are in common with the associated Grade 7 form
    - three Grade 8 common items were included on both Grade 7 and Grade 8 forms
    - three Grade 7 common items were included on both Grade 7 and Grade 8 forms

Operational Form Assembly. Following the field test, items were selected for the operational forms to maximize the standard coverage, that is, to include as many unique content standards as possible. However, statistical criteria were also used to select items, avoiding items with extreme difficulty values (e.g., p-value < 0.05), poor correlations between the item score and the total test score (e.g., point-biserial < 0.15), and poor fit to the Rasch model (e.g., Rasch Infit less than 0.5 or greater than 1.5). In additional, when possible, field test anchor items (common items across forms) were selected as items for the Stage 1 block of the operational assessment and as many field test anchor items as possible were used for operational forms because these items had a larger number of student responses contributing to their item parameter estimates.

The operational forms were built block-by-block because of the different range of difficulty that each block of items (Low, Moderate, High) requires.

Stage 1 Assembly. The Stage 1 block is a common set of items administered to all students. This block was selected to have 15 items with an appropriate distribution of item difficulties to support the assessment of students with a broad range of abilities. The most difficult item in the pool with good statistical characteristics was often selected for Stage 1, along with the easiest item in the pool with good statistical characteristics. The other 13 items were selected to create a reasonably uniform distribution of item difficulty between the extremes—in other words, to have Rasch difficulty values that were approximately equally spaced across the scale range. Designing a Stage 1 block with items representing a very wide range of ability supports the estimation of achievement (theta) for students of various mathematics achievement levels. This is important because these 15 items determine which Stage 2 block a student receives.

Stage 2 Assembly. Three Stage 2 blocks were developed to support more precise measurement targeted for students who demonstrated low, moderate, and high achievement on Stage 1. To guide the development of each block, field test data were used to identify the range of student achievement in each grade. Table 2 provides the range of ability on the theta scale at seven ability reference points per grade—the minimum and maximum abilities achieved by students in the field test and grade, and the theta values associated with the 25th, 33rd, 50th (median), 67th, and 75th percentile ranks.

The field test theta estimates in Table 2 were used as a guide to select items for each Stage 2 block (Low, Moderate, High) in each grade. Specifically, the Low block was selected to include items that provide precise measurement for students primarily below the 33rd percentile. Similarly, the High block was selected to include items that provide precise measurement for students primarily above the 67th percentile. The Moderate block contains a range of item difficulties appropriate for students between the 33rd and 67th percentiles. The Low and Moderate blocks have items that overlap in difficulty and likewise the Moderate and High blocks contain items that overlap in difficulty so that students who may be incorrectly routed to a Stage 2 block from Stage 1 will likely still receive items appropriate to their achievement level.

Once a candidate set of items was selected for each Stage 2 block, the information curves were compared to determine whether the blocks differed sufficiently in difficulty from one another to show appropriate and differentiable measurement precision for different locations on the scale. For example, the test information curves for the three Stage 2 blocks created for Grade 3 are shown in Figure 2. The curves overlap but have maximums at different regions of the score scale, as desired.

Once the Stage 1 and Stage 2 items were selected based on psychometric considerations, a WestEd content development expert reviewed the item set and worked with the WestEd psychometrician to make adjustments to the forms that appropriately balanced content considerations with psychometric considerations. For example, if the Stage 1 block contained too many items with a particular answer key, either some of the items were

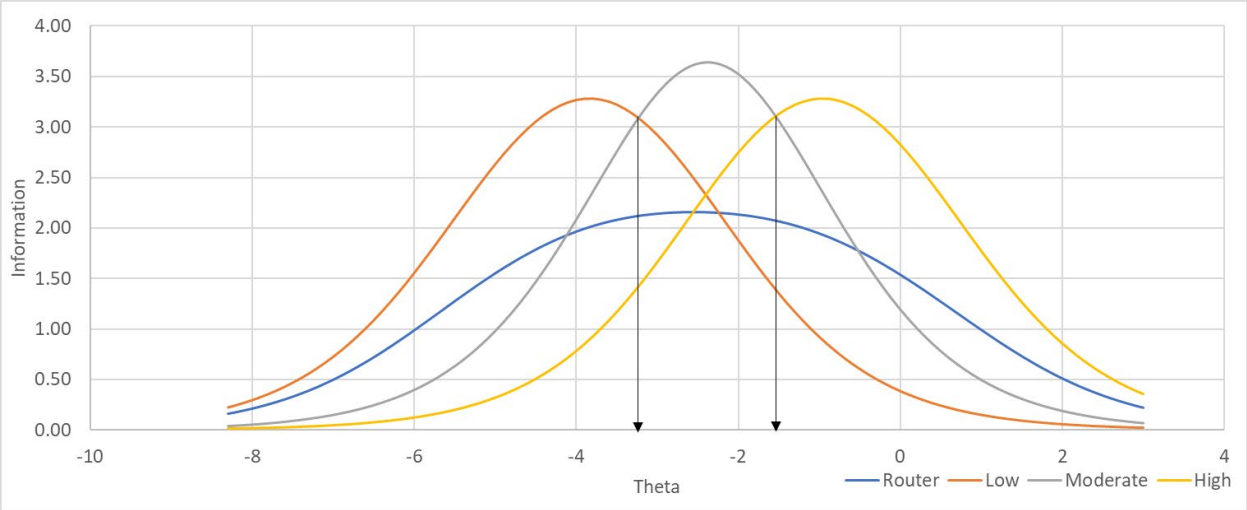
replaced with similar items with different answer keys, or sometimes a Stage 1 item was swapped with a Stage 2 item.

Items were typically ordered by difficulty with the easiest item appearing first in a block and the hardest item appearing last. Occasionally small deviations to the ordering were made to avoid having very similar items side by side or too many items with the same answer key next to one another.

Table 2. Theta Distribution by Grade on the Vertical Scale

Theta	K	1	2	3	4	5	6	7	8
Lowest	-11.99	-10.47	-9.49	-8.29	-6.06	-6.30	-5.38	-5.11	-4.80
25th %ile	-6.45	-5.06	-4.36	-3.39	-2.95	-2.58	-2.26	-1.87	-1.63
33rd %ile	-6.22	-4.81	-4.06	-3.17	-2.74	-2.38	-2.08	-1.69	-1.44
Median	-5.67	-4.25	-3.53	-2.62	-2.28	-2.01	-1.68	-1.30	-1.06
67th %ile	-5.10	-3.77	-3.11	-2.06	-1.83	-1.54	-1.25	-0.91	-0.71
75th %ile	-4.74	-3.51	-2.90	-1.74	-1.63	-1.29	-1.00	-0.64	-0.38
Highest	-1.08	0.51	1.34	3.02	3.09	4.61	4.16	2.58	4.98

Figure 2. Information Curves for the Savvas Diagnostic Assessment Blocks, Grade 3



Routing. After operational forms were developed, routing rules were required to determine how performance on Stage 1 is used to determine the appropriate Stage 2 block. Routing cut scores were determined by identifying the location on the theta scale where the Low and Moderate block information curves crossed and where the Moderate and High block information curves crossed. These locations represent the transition points between where students are optimally measured by the Low, Moderate, and High blocks of items, respectively. The example in Figure 2 indicates that the Low and Moderate information curves cross at approximately -3.3 on the theta scale and the Moderate and

High information curves cross at approximately -1.5 on the theta scale. These theta values are the cut scores that determine how students are routed to the most appropriate Stage 2 block. Table 3 provides the ranges of scale score values resulting from students' performance on the Stage 1 router block that route students to one of the three Stage 2 blocks. The scale score values are a direct transformation of the theta cut scores to the Savvas Diagnostic scale. The theta to scale score transformation is described in the Vertical Scale and Reporting sections of this report.

The percentage of students estimated to receive each of the three forms, based on field-test data is provided in the last three columns of Table 3. Ideally, approximately equal numbers of students would be routed to each form to optimize the adaptive nature of the forms and maximize the number of students taking each item. However, due to the difficulty of the item pools relative to the ability levels of the students, this was not always possible. Additionally, having three forms that measured distinct regions of the scale was prioritized over equal percentages. The data show that although the percentages are not equivalent, especially at the higher grade levels, a sizeable proportion of students will be routed to each form at all grade levels.

Table 3. Scale Score Ranges used to Route to Stage 2

	Routing			% Routed to Low	% Routed to Moderate	% Routed to High
Grade	**Low**	**Moderate**	**High**			
K	1000-1329	1330-1384	1385-2000	25%	40%	35%
1	1000-1389	1390-1459	1460-2000	20%	45%	35%
2	1000-1434	1435-1504	1505-2000	26%	48%	26%
3	1000-1484	1485-1574	1575-2000	28%	51%	21%
4	1000-1509	1510-1579	1580-2000	30%	51%	19%
5	1000-1539	1540-1604	1605-2000	40%	46%	15%
6	1000-1564	1565-1629	1630-2000	49%	38%	12%
7	1000-1589	1590-1649	1650-2000	55%	35%	10%
8	1000-1609	1610-1674	1675-2000	62%	28%	10%

Psychometric Calibration and Equating

Rasch Model

Item Response Theory (IRT) is a family of statistical models used to associate item and test data with examinee (i.e., student) performance. The core components of IRT are the item data, the test takers' latent or unobserved ability, and observed performance on the items. In IRT, performance on the test is considered an estimate of the test takers' ability rather than an absolute measure of their ability.

The Rasch model considers the examinee ability estimates, referred to as theta estimates or θ, as a function of the item difficulty and the student's performance on the items resulting from their true ability. Proponents of the Rasch model assert that (a) item difficulty estimates are independent of the tested sample and (b) student ability estimates are invariant to the items administered.

The Rasch model is a logistic regression model based on a single parameter, the item difficulty parameter, b. The Rasch model describes the relationship between item difficulty, b, and student ability, θ, in terms of the following logistic equation:

$$P(U_i = 1 \mid \theta) = P(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}},$$

where b_i is the difficulty parameter for item *i* and θ is the student ability parameter. The expression $P(U_i = 1 \mid \theta)$ represents the probability of a student of ability θ answering item *i* correctly. Higher values of θ are associated with test takers of higher ability. Similarly, higher b values are associated with more difficult items. The IRT Rasch difficulty parameter b_i is expressed on the same scale as the test taker's ability parameter, θ.

The initial calibration results in a Rasch θ value on the native logit scale, centered at 0 and extending in both the positive and negative directions. The native logit scale is linearly transformed, as is common industry practice, to result in the Savvas scale that ranges between 1000 and 2000, as detailed in the Reporting section of this report.

Estimating Item Difficulty and Student Ability Parameters

Calibrations were conducted using 2021 field test data and the IRT software Winsteps®.[13] The Rasch calibration results, including item characteristic curves, student ability frequency distributions, test characteristic curves, test information curves, and conditional standard errors of measurement, are provided graphically in Appendix C.

Vertical Scale

A vertical scale is a common cross-grade scale score system that allows for the direct comparison of student test scores across grade levels. Vertical scaling is the process of placing test scores that measure similar content at different grade levels onto a common scale.

Vertical Scaling Study. The data collection design used to develop the Savvas vertical scale was the common-item non-equivalent groups design in which students in adjacent grade levels respond to both common and unique items, thereby allowing direct comparison of item difficulties across grades.[14] This design allows the entire Savvas item pool to be placed

[13] Linacre, J. M. (2021). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com
[14] Kolen, M. J., & Brennan, R. L. (2004). Test equating: Methods and practices (2nd ed.). New York: Springer.

on the same scale. The common items between adjacent grades determine the scaling relationship between tests in adjacent grades.

Because the vertical scale study depends explicitly on the quality of the cross-grade common items, it is important to verify that the items have adequate technical quality in both grades. P-values, item-total correlations, and Rasch infit statistics were evaluated. Poor performing items may lead to inaccurate or unstable vertical scale results. Although some of the items were flagged based on their statistics, an evaluation of the scaling constants with and without flagged items indicated that the results were very similar. Therefore, because the common item set was relatively small, all cross-grade common items were retained in the vertical scaling analysis.

Nine separate WINSTEPS item calibrations were used to create nine separate scales, one for each grade. Because the starting place for the scale is arbitrary, the grade 8 scale was selected as the base. By default, each separate scale has a mean of 0 and a standard deviation of 1 on the theta scale, including the grade 8 scale. To place grade 7 on the grade 8 scale, the common items are used to calculate a vertical scale constant between grade 8 and grade 7. This constant is added to the grade 7 mean of 0 to adjust for the difference in difficulty between the two scales. As shown in Table 4, the grade 7 constant was -0.3323. For grade 6, a scaling constant had to be calculated based on the items common between grades 6 and 7. This scaling constant (-0.4215) was added to the scaling constant obtained for the grade 7 to grade 8 relationship (-0.3323), resulting in a combined vertical scale constant of -0.7538 to put grade 6 on the grade 8 scale. This chaining of scaling constant was continued all the way to Grade K, where a vertical scale constant of -6.1771 in Table 4 reflects the sum of 8 adjacent grade mean differences.

Table 4. Grade-to-Grade Linking Constants

Grade	Vertical Scale Constant
K	-6.1771
1	-4.9097
2	-3.9681
3	-2.4881
4	-2.0516
5	-1.3147
6	-0.7538
7	-0.3323
8*	0.0000

* Grade 8 was set as the base scale which by default has a mean of 0.

As described more thoroughly in the Reporting section of this report, the Savvas score scale was created to span a range of 1000 to 2000 across grades K to 8. Based on the

vertical scale results, a linear transformation of the theta scale was developed using a slope of 50 and an intercept of 1650 to achieve a score scale range of approximately 1000 to 2000. Therefore, to transform a given theta on the logit scale to a Savvas Diagnostic scale score, theta is multiplied by 50 and 1650 is added to the result.

Reliability

In the sections below, reliability and measurement error estimates for the Savvas Mathematics Diagnostic Assessments are provided. These values are based on an IRT framework rather than a classical test theory framework. Thus, marginal reliability, standard errors of measurement derived from marginal reliability estimates, and conditional standard errors of measurement are all estimated based on the Rasch-based theta scale. Once the operational forms are administered in Fall 2021, common classical reliability indices including coefficient alpha[15] will be provided. Likewise, reliability information will be estimated for the Screener during the first administration in Fall 2021.

Marginal Reliability

Marginal reliability is a measure of the overall reliability of a test based on the average conditional standard error, estimated across the scale range. The IRT marginal reliability coefficients are similar to coefficient alpha under classical test theory but presented within the IRT framework. In fact, the calculation of marginal reliability is based upon the classical test theory conceptualization of observed scores as a function of true scores and error.

Reliability is then estimated as the variance of true scores minus the error variance, divided by true score variance. True score variance was operationalized as the variance of theta estimates by grade. Error variance was calculated as the reciprocal of test information, calculated for each examinee, and summed across all examinees. Error variance is therefore based on a sample-weighted average of measurement error across the scale. The Savvas Mathematics Diagnostic Assessment marginal reliability estimates are good across all grades, as indicated in Table 5, reflecting reliabilities that are acceptable and expected for tests of this length and purpose.

[15] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334

Table 5. Marginal Reliabilities by Grade for Savvas Diagnostic Assessments

Grade	Marginal Reliability
K	0.89
1	0.86
2	0.84
3	0.88
4	0.81
5	0.84
6	0.81
7	0.81
8	0.83

Marginal reliability is a function of test information. For the Rasch model, item information is defined as:

$I_i(\theta) = P(\theta)\mathrm{Q}(\theta)$,

where $P(\theta)$ was defined in the Rasch Model section above and $\mathrm{Q}(\theta) = 1 - P(\theta)$.

Test information is then the sum of item information across all items within a test form. Test information curves for the Savvas Diagnostic Assessments are presented graphically in Appendix C.

Standard Error of Measurement

The Standard Error of Measurement (SEM) is a measure of score precision computed using the formula $SEM = SD\sqrt{1-r}$, where SD is the standard deviation total scores across examinees and r is the estimate of reliability. SEM is constant across the scale, providing a single numeric description of expected fluctuations in examinees' observed scores due to error. Larger SEM values indicate lower score precision. The values in Table 6 are on the scale score metric (1000-2000). Field test data indicate the SEM patterns conform to expectations.

Table 6. Standard Error of Measurement by Grade

Grade	SEM
K	36
1	32
2	28
3	33
4	24
5	24
6	23
7	21
8	23

Conditional Standard Error of Measurement

The Conditional Standard Error of Measurement (CSEM) is a more precise measure of score error that varies across the score distribution. The CSEM indicates how much a student's true score might vary from the reported score. Measurement is more precise for scores with small CSEMs. The CSEM on the theta scale is presented graphically for each grade and Diagnostic form (Low, Moderate, High) in Appendix C. Multiplying the CSEM on the theta scale by 50 (the slope scaling constant), will transform the CSEM to the Savvas scale score metric (i.e., 1000-2000).

Validity

Validity is considered the most fundamental consideration in developing and evaluating tests and is supported by evidence indicating that appropriate inferences can be made from test scores for specified purposes.[16] The collection of validity evidence occurs over time, beginning with the theory and practices supporting the test design and development process and continuing throughout the assessment lifecycle. No single study or piece of evidence renders an assessment valid. Validity evidence is based on multiple criteria, as detailed next.

Evidence Based on Test Content

Evidence based on test content includes the basis for the development of item specifications and artifacts from the item alignment process. For the Screener and Diagnostic Assessments, this evidence includes the following elements:

- Framing of content selection from the research literature. WestEd conducted a literature review to identify the foundational mathematics skills necessary for

[16] American Educational Research Association (AERA), American Psychological Association (APA), and the National Council for Measurement in Education (NCME). (2014). Standards for Educational and Psychological Testing.

success in elementary and middle school. The results of the literature review were used to guide item specifications and is detailed in the Test Design and Development section of this report. Evidence of appropriate standards coverage for each form is provided in Appendix A and B of this document.

- Evidence for appropriate training of item writers. WestEd supported item writer adherence to the PARCC Item Specifications and Style Guide through targeted training and feedback through the item development process. The training procedures also included guidelines for universal design, bias, and sensitivity, as described more thoroughly in the Test Design and Development section.

- Evidence of item to standard alignment via item writing templates and item review procedures. Item writing templates were adapted from WestEd's standard template designed to support the development of high-quality items written to align to specific content. An example of an item template is included in Appendix D. The inclusion of the item content along with the standard alignment provided a mechanism for tracking and evaluating the appropriateness of the item as a measure of the specified content standard. Content experts verified the identified common core standard alignment during the item review process. Savvas also created a crosswalk between the proprietary skill spine and the common core standards such that all items had an appropriate alignment to a Savvas skill code.

Evidence Based on Response Processes

Evidence based on response processes supports the association between the test construct (i.e., performance on Mathematics skills) and the responses elicited from the examinees (i.e., the students). Evidence based on response processes is typically compiled through cognitive interviews, focus groups, process data (e.g., mouse-clicks, response time), and in some cases, eye-tracking studies. Support for validity based on response processes will be gathered during the operational 2021 administration.

Evidence Based on Internal Structure

Analyses of the internal structure of the assessments provide evidence that the test items appropriately represent the constructs of interest. This evidence typically includes analysis of operational test data. Validity evidence based on the internal structure of the Diagnostic Assessments was provided through item analyses that included item-total score correlations. Item-total correlations were used to select items that indicated an appropriate relationship between the item and the overall construct. The field-test data provide evidence supporting the appropriate internal structure of the assessments, with item-total correlations (point-biserials) ranging from 0.19 to 0.72 with an average of 0.42. In addition, fit statistics were evaluated to verify that all operational items had reasonable fit to the Rasch model (i.e., infit within 0.5 to 1.5).

Additional evidence is provided by the test informative curves and theta frequency distributions provided in Appendix C of this report, which show that the ability of the student and the information yielded by the test forms are well matched. That is, the test information curves and test characteristic curves for the Low, Moderate, and High forms tend to shift from lower to higher achievement on the test scale as would be expected and desired and the highest measurement precision corresponds to the region of the scale associated with most students' achievement level (i.e., theta value).

Evidence Based on Relations to Other Variables

Analysis of the relationship to other variables may take different forms. For instance, it would be expected to observe moderate to high positive correlations between students' Mathematics Diagnostic scores and their associated scores on their state summative mathematics assessments. Predictive validity studies, that demonstrate the association between the Savvas Diagnostic Assessments and state summative assessments are planned for the 2021-22 school year.

Evidence Based on Test Consequences

Consequential validity evidence can be provided in multiple ways. Savvas has developed an MSDA User Guide to help test users understand the purposes of the assessments and the intended score interpretations. The company also monitors feedback from the field to verify that educators understand how to administer the assessments and use the resulting score information. Long term, efficacy studies can be conducted to show that students using the MSDA tend to demonstrate growth and improved achievement on aspects of mathematics that have been identified by the assessments as areas for improvement for students.

Reporting

Generating the Savvas Scale

Scale scores for tests on a common scale are used to report consistent information about student achievement to assessment score users regardless of the form administered, the group taking the assessment, or the date on which the assessment was taken. This is in contrast to other types of scores with interpretations that vary depending on the specific form taken such as number correct scores.

There are two types of scales commonly used with educational assessments: horizontal and vertical scales. Horizontal scales support the consistent interpretation of scores (e.g., students' scores, cut scores, etc.) within grade. Vertical scales support the consistent interpretation of scores across grades. For example, a score of 1500 can be interpreted similarly for a student in fourth grade and for a student in fifth grade. Differences in vertical scale scores across time provide a metric of growth in student performance. For example, if a student's mathematics score was 1450 in fourth grade and 1500 in fifth grade, then the student's observed mathematics achievement increased by 50 scale score points

from fourth to fifth grade. Because vertical scales provide a straightforward gauge of student growth, they are commonly used with educational assessment systems that span multiple grade levels.

Choosing a Scale. The choice of scale is to a large extent arbitrary which is why various assessment programs use different score scales. The Savvas vertical scale was constructed to have a range wide enough to support at least nine different grade levels (kindergarten through grade 8). It is not uncommon for vertical scales that span a wide range of grades to have a wide range of scores. The Savvas vertical scale was established between 1000 to 2000 for the following reasons:

- it is unlikely to be confused with other publishers' scales,
- the scale score range of 1,000 to 2,000 provides a sufficient number of scale score points to allow for adequate score precision across the full continuum of student performance from the lowest performing Kindergarteners to the highest performing eighth graders,
- it is recommended to avoid scale scores in the range of 0-100 because they are easily confused with percentile ranks and percent correct scores, and
- establishing the Savvas vertical scale within the bounds of 1000 and 2000 guarantees that every scale score will be four digits. The scale will not dip below 1000, potentially creating scoring complexity due to varying numbers of digits in scores across the scale.

Estimating National Percentile Ranks

National percentile ranks were estimated for the Diagnostic Assessments using data from the 2021 field test. To account for any lack of national representativeness of the field-test sample, the sample was weighted to match the demographic profile of students nationally. The sample weights were based on relevant strata. A stratum is a variable that can be associated with each student in the sample. The specific strata that were used to weight the Savvas sample included school-level race (percent white, percent Asian, percent black, percent Hispanic), gender (percent male, percent female), and socio-economic status (percent of students receiving free and reduced price lunch). Strata that were underrepresented in the norming sample relative to national data received larger sample weights, while strata that were overrepresented in the norming sample received smaller weights. After the strata and associated strata sampling weights were established, national percentile norming tables were developed for each Mathematics Diagnostic Assessment.

The use of customer data to support norming studies has two key advantages: it requires no additional data collection, and it uses data from real-world operational assessment conditions, which reduces motivation effects. Developing national norms based on field test data supported the ability to report percentile scores for the fall 2021 reports. Because

the field test sample is likely small relative to the operational sample testing in the 2021-2022 academic year administration, the norms will be updated using operational data following the 2021-2022 academic year to improve the representativeness of the national percentile ranks for use in 2022-2023 and beyond.

Predicting Performance on Summative Assessments

Predictive validity studies will be conducted using matched data (matched student Diagnostic and state summative test scores) after the 2021 operational assessment administration. The study will be used to identify the scale scores on the Diagnostic Assessments associated with passing the statewide summative mathematics assessment for each grade that the state summative is administered. These studies will be state specific. Score users in participating states will be able to gauge student readiness for success on their state's assessment based on student performance on the Savvas Diagnostic Assessments.

Embedded Standard Setting

Standard setting is a systematic process used to identify cut scores that categorize test scores into performance levels. Performance levels are presented as intervals of test scores with the cut score being the lowest score in the interval. The use of performance levels supports test score interpretation by giving meaning to test scores by means of performance level descriptors. Performance level descriptors describe the range of knowledge, skills, and abilities of students in each performance level.
Embedded Standard Setting[17] was proposed and used to support the establishment of cut scores, performance levels, and performance level descriptors (PLDs) for the Savvas assessments. Embedded Standard Setting (ESS) is a research-based, peer-reviewed standard setting method based on a principled assessment design framework. ESS offers key advantages for assessments designed to be used in multiple states:

- It supports the efficient development of customized PLDs and cut scores for each state, aligned with the state's summative assessment;
- It supports the estimation of criterion-referenced scores based on the Common Core State Standards that can be adjusted to reflect state-specific standards;
- It supports the establishment of a modular system of core Savvas PLDs and cut scores, which can be leveraged quickly to create new state-specific cut scores;
- It results in a coherent relationship between standards, items, performance levels, and cut scores.

The ESS process for the Savvas Mathematics Diagnostic Assessments led to two key deliverables:

[17] Lewis, D. & Cook, R. (2020). Embedded Standard Setting: Aligning Standard-Setting Methodology with Contemporary Assessment Design Principles. Educational Measurement: Issues and Practice, 39: 8-21. Retrieved from: https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12318

1. Cut scores defining four levels of performance:
    Level 1: Does Not Meet Expectations
    Level 2: Approaching Expectations
    Level 3: Meets Expectations
    Level 4: Exceeds Expectations
2. PLDs that support score interpretation by explicating the knowledge, skills, and abilities attributed to students in each performance level.

Embedded Standard Setting was implemented for the Savvas Diagnostic assessments in 4 steps:

1. Align test items to preliminary performance level descriptors
2. Conduct ESS analyses to estimate preliminary cut scores
3. Conduct vertical articulation to support a coherent system of cross-grade cut scores
4. Develop custom Savvas performance level descriptors

1. **Align test items to preliminary performance level descriptors.** The first step of ESS is an alignment activity that has elements in common with the Item-Descriptor (ID) Matching standard setting method (Ferrara & Lewis, 2012), specifically, the alignment of test items to PLDs. The Savvas Mathematics Screener and Diagnostic Assessment items were written to align to the CCSS. A representative set of items from the Diagnostic item pool was selected for each grade and assembled into ordered item books of approximately 60 items per grade and the items were presented to subject matter experts (SMEs) in order of IRT item difficulty. The PARCC performance level descriptors were used to support this purpose and were considered preliminary performance level descriptors specifically for the purpose of establishing ESS cut scores. That is, for each item, the relevant PARCC performance level descriptors were presented to the subject matter experts. Then, the SMEs selected the PARCC performance level evidence statements that were best aligned to the given item. The SMEs completed the alignment judgment for each item in the ordered item book.

    Because PARCC does not have PLDs for grades K through 2, the CCSSO K-2 Mathematics PLDs[18] were used, which are aligned to the CCSS for these grades.

2. **Conduct ESS analyses to estimate preliminary cut scores.** After all SME item-PLD alignments were established, the ESS-Weight algorithm (Lewis, Lee, & Choi, 2021) was

[18] Retrieved from: https://ccsso.org/resource-library/performance-level-descriptors-pld-grades-k-2-mathematics

used to identify the cut scores that optimized the relationship between the SME Item-PLD alignments and empirical data. These cut scores were considered preliminary cut scores; final cut scores were established after assuring appropriate vertical articulation, described next.

3. **Conduct vertical articulation to support a coherent system of cross-grade cut scores.** The purpose of vertical articulation is to align cut scores across grade levels on the vertical scale and assure that cut scores for a given performance level increase as grade level increases. A common approach to smoothing is to use information from neighboring observations when the stability of a data point is unclear. Each vertically articulated cut score is a function of:
   - The raw cut estimated from the embedded standard setting procedure based on ESS-Weight.
   - The corresponding cut scores in adjacent grades (e.g., the grade 3 Level 2 cut score is informed by the grade 2 Level 2 cut score and the grade 4 Level 2 cut score).

The resulting system of vertically articulated cut scores and associated data (e.g., estimated percentage of students performing at each level) were shared with Savvas for review and discussion. The final, vertically articulated cut scores are provided in Table 7 and are presented graphically in Figure 3. Figure 4 provides the resulting estimated percent of students in each performance level across the grades. These estimated percentages are based on field test data. Operational percentages may vary.

Table 7. Vertically Articulated Cut Scores by Grade

Grade	Level 2	Level 3	Level 4
K	1325	1365	1520
1	1370	1412	1542
2	1413	1455	1564
3	1452	1493	1586
4	1489	1528	1610
5	1517	1550	1634
6	1539	1572	1658
7	1561	1594	1676
8	1579	1610	1699

Figure 3. Vertically Articulated Cut Scores by Grade

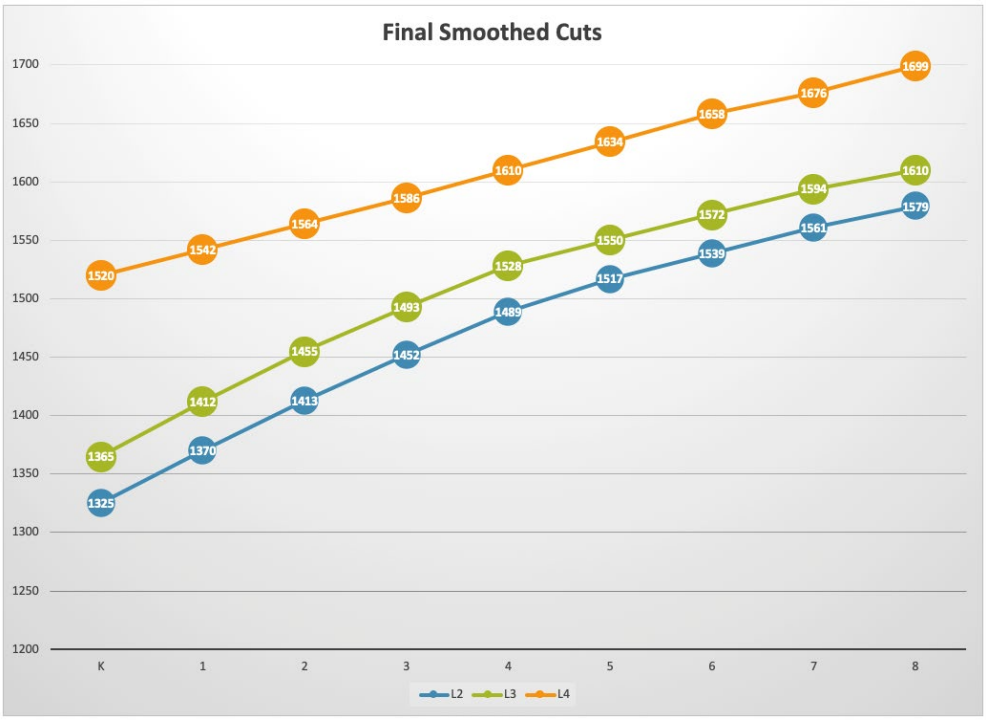
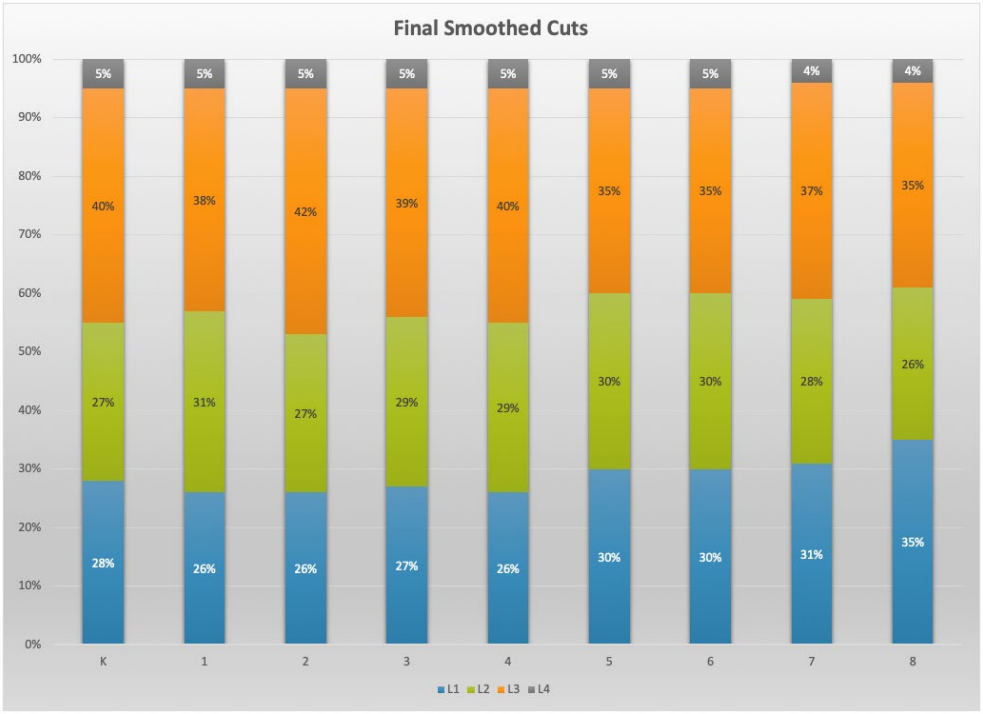


Figure 4. Percentage of Students in each Performance Level by Grade



4. **Develop Custom PLDs for the Savvas Diagnostic Assessments.** PLDs communicate the characteristics of the knowledge of students performing at each level and the ways that this knowledge deepens as the performance levels increase.

   After vertically articulated cut scores were established and adopted, each item included in the Spring 2021 Savvas Mathematics Diagnostic field test was associated with operational performance levels as follows:

   - **Level 1** items are items with difficulty values (expressed as scale scores) from the lowest possible scale score to one less than the Level 2 cut score.
   - **Level 2** items are items with scale score difficulty values from the Level 2 cut score to one less than the Level 3 cut score.
   - **Level 3** items are items with scale score difficulty values from the Level 3 cut score to one less than the Level 4 cut score.
   - **Level 4** items are items with scale score difficulty values from the Level 4 cut score to the highest possible scale score.

   The attributes of the items were used to develop very detailed range PLDs based on all of the skills measured by items within each performance level. That is, WestEd SMEs reviewed the items associated with a given performance level and grade, developed item-level descriptors based on the measurement attributes of each item in the level, and integrated the item-level descriptors for items in the given performance level and grade. The resulting descriptors form the custom Savvas Diagnostic PLDs for that grade. The level of detail included in the custom Savvas Diagnostic range PLDs would allow for targeted item development such that items could be written to assess specific content standards at specific performance levels. The range PLDs also support test score interpretation and communication of the mathematical concepts measured in each performance level in each grade. This level of detail may be important to states and districts who are interested in a detailed alignment of the Savvas PLDs to the state PLDs.

   To support score interpretation for individual score users, a briefer set of reporting PLDs were developed that describe at a high level what each performance level means across all of the Savvas Mathematics Diagnostic Assessments, K-8. The reporting PLDs are:

   - **Level 1: Does Not Meet Expectations**

     Students performing at the **Does Not Meet Expectations** level demonstrate minimal understanding of the concepts, skills, and procedures of the grade-level mathematics standards.

- **Level 2: Approaching Expectations**

  Students performing at the **Approaching Expectations** level demonstrate some understanding of the concepts, skills, and procedures of the grade-level mathematics standards.

- **Level 3: Meets Expectations**

  Students performing at the **Meets Expectations** level demonstrate expected understanding of the concepts, skills, and procedures of the grade-level mathematics standards.

- **Level 4: Exceeds Expectations**

  Students performing at the **Exceeds Expectations** level demonstrate sophisticated understanding of the concepts, skills, and procedures of the grade-level mathematics standards.

State-Specific PLDs and Cut Scores

The core PLDs can be adjusted to align with a specific state's expectations for summative achievement. The first step in this process is to gather state-specific user data from the operational administration in the state. Next, relevant strata are identified to weight the state's user data. Stratum may include school-level race (percent white, percent Asian, percent black, percent Hispanic), gender (percent male, percent female), and socio-economic status (percent of students receiving free and reduced price lunch). Other relevant strata may be used.

Strata that were underrepresented in the user data relative to state data received larger sample weights, while strata that were overrepresented in the user data received smaller weights. After the strata and associated strata sampling weights are established, weights are applied to achieve state representative impact data. Following weighting, the user data will match the demographics of the target state on the various strata. Next, equipercentile methods are applied to adjust the core Savvas cut scores to achieve state-specific cut scores.

Alternatively, state-specific cut scores can be obtained via predictive validity studies where Savvas Mathematics Diagnostic assessment scores and state summative mathematics test scores are statistically linked.

Once state-specific cut scores on the Savvas Mathematics Diagnostic Assessment vertical scale are determined, state-specific performance level descriptors can be developed using the same techniques described for the core PLDs. The reporting PLDs already adopted by the state can be leveraged for Mathematics Diagnostic assessment reporting PLDs and the custom range PLD database can be updated with the state-specific cut scores to realign the range PLDs with the state-specific cut scores.

Screener Cut Scores

Number correct scores are the reporting metric for the Screener Assessments. Because the Screener is not linked to the Diagnostic vertical scale, the Diagnostic Assessment performance levels and performance level descriptors described previously not apply to the Screener. However, one of the purposes of the Savvas Screener Assessments is to signal to teachers which of the Savvas Diagnostic Assessments would best measure the mathematical achievement of a given student. Therefore, a cut score is needed on the Screener that indicates when, for example, performance on the Grade 3 Screener is low enough that the teacher should consider administering a Grade 2 Diagnostic Assessment. Recall that the Grade 3 Screener measures skills from Grade 2 that are pre-requisites for successfully learning Grade 3 content. However, if a student has not mastered much of the Grade 2 pre-requisite skills, then the Grade 2 Diagnostic assessment will likely provide the teacher with better information about the student's strengths and areas for improvement than the Grade 3 Diagnostic.

Although the difficulty of the Mathematics Diagnostic Assessments was established during a field-test conducted early in Spring 2021, the difficulty of the Mathematics Screener Assessments has not been determined. Because the two assessments share no common items, the two assessments cannot be statistically linked until a common set of students takes both assessments.

However, relevant Screener information is available, such as the number of items and the number of response options per item. The Screener Assessments are comprised of multiple-choice items only. In grades K and 1 there are three response options so a "chance score" would be the number of items on the assessment divided by three (e.g., 10 out of 30). In grades 2-8 there are four response options so a "chance score" would be the number of items on the assessment divided by four (e.g., approximately 8 out of 30). Students with a total score on the screener at or below a "chance score" are clearly struggling with the prerequisite skills needed to be successful with on-grade level content. These students are clear candidates for taking a below-grade Diagnostic Assessment. However, it may be appropriate for some students who score above the chance score to also take a below-grade Diagnostic Assessment.

Students likely to be classified as Does not Meet Expectations on the on-grade Diagnostic Assessment are struggling with on-grade level content. Some of these students score above a chance score but may still be good candidates for taking a below-grade Diagnostic Assessment. Although the scores on the Screener that would map to the Does not Meet Expectations range of the Diagnostic scale is unknown, the following assumptions were used to create a Screener cut score higher than the chance score:

- The Screener item difficulty will be similar to (and perhaps even less difficult than) the Low Diagnostic form item difficulty
- The percent correct score on the Low difficulty Diagnostic block associated with the transition from Does not Meet Expectations to Approaching Expectations serves as the percent correct on the Screener that is used to route student to the grade level Diagnostic as opposed to the adjacent below-grade Diagnostic Assessment
- The associated National Percentile Ranks on the Diagnostic form provide a ballpark estimate of the numbers of students that would be routed to the below-grade Diagnostic Assessment

Table 8 shows the number of items on each of the Screener Assessments along with the number of items that represent a chance score (and the associated % correct), as well as the number of items (and percent correct) associated with the score at the bottom of the Approaching Expectations performance level on the Low form of the on-grade level Diagnostic Assessment. From Table 8, it is clear that using the bottom of the Approaching Expectations performance level as a threshold for routing students results in a higher cut score and more students who would be recommended to take a below-grade Diagnostic Assessment than the chance level scores. These percentages are based on the National Percentile Ranks calculated using the Diagnostic Assessment field-test data. If the screener items are actually easier than the Low form items on average, then the number of students routed to the below-grade level Diagnostic Assessment would be lower than the numbers suggested by the percentiles. For example, instead of 33% of students routed to the Kindergarten Diagnostic Assessment (based on the last column, first row of Table 9), if the Screener is much easier than the Low difficulty form of the Diagnostic Assessment, perhaps only 20% would be routed down.

Table 8. Number Correct, Percent Correct, and Percentile Ranks of Potential Screener Cuts

Grade	Number of Screener Items	Chance Percent Correct	Chance Number Correct	Chance Percentile Rank	Bottom of Approaching Percent Correct	Bottom of Approaching Number Correct	Bottom of Approaching Percentile Rank
1	20	33%	7	14	57%	11	33
2	22	25%	6	4	47%	9	31
3	25	25%	6	8	47%	10	27
4	25	25%	6	7	40%	10	27
5	25	25%	6	14	40%	10	30
6	30	25%	8	11	37%	9	37
7	30	25%	8	14	40%	12	35
8	30	25%	8	18	37%	11	36

Note. Kindergarten is not included because there is no lower grade level Diagnostic Assessment for low scoring students to take.

After considering multiple approaches to setting the Screener cut score, the number correct score that represents the average of the two methods (last column of Table 9) was selected as the operational cut score for Fall 2021. Therefore, the adjacent, below-grade Diagnostic Assessment will be recommended for students who answer fewer than 8 items correctly on the Screener in Grades 1-5, or fewer than 10 items in Grades 6-8. This is a fairly conservative approach, resulting in most students being routed to the on-grade level Diagnostic Assessment. Matched data from the Fall 2021 administration for the Screener and the Diagnostic Assessments will be used to refine the Screener cut scores for future administrations.

Table 9. Comparison of Number Correct Screener Cut Scores

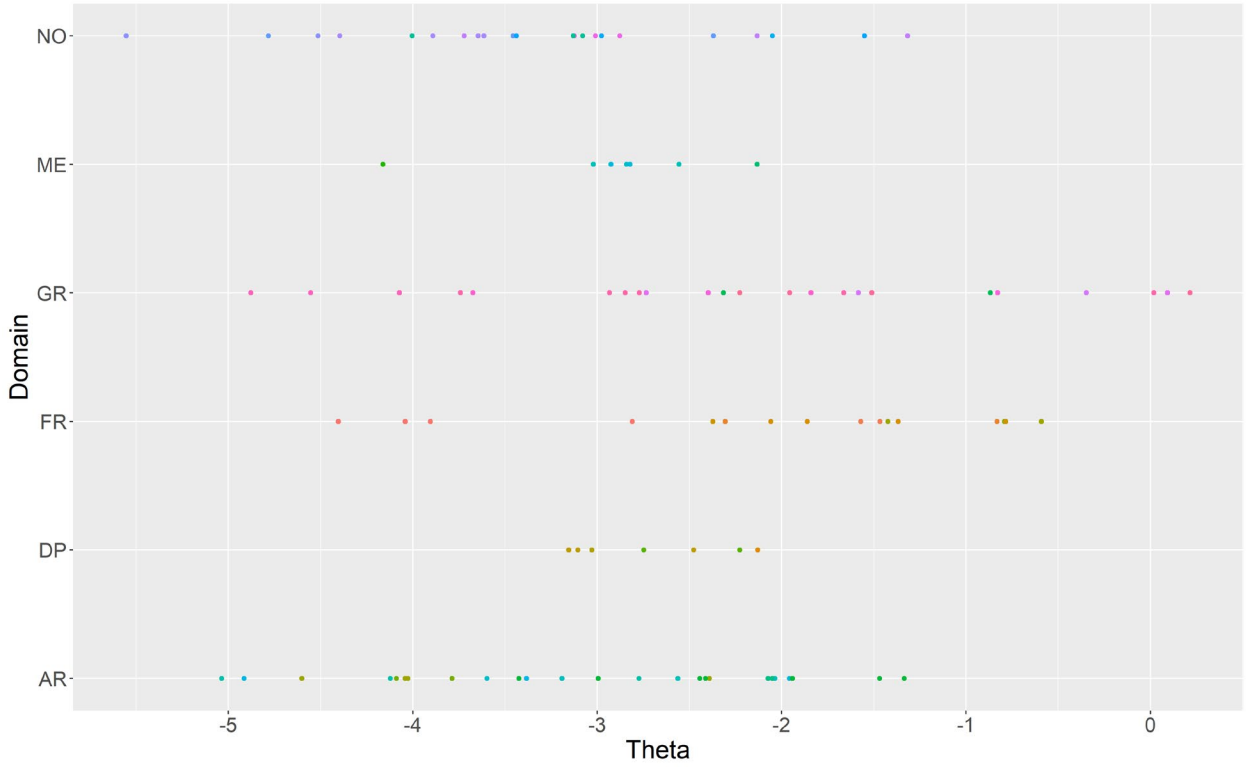
Grade	Number of Screener Items	Chance Number Correct Score	Bottom of Approaching Number Correct Score	Average Number Correct Score
1	20	7	9	**8**
2	22	6	10	**8**
3	25	6	10	**8**
4	25	6	10	**8**
5	25	6	9	**8**
6	30	8	12	**10**
7	30	8	11	**10**
8	30	8	11	**10**

Item Mapping

All items administered during the Spring 2021 Savvas Mathematics field test were considered for inclusion in an item mapping process. Item mapping shows the difficulty of each item on the underlying Savvas vertical scale. Items with point biserial correlations (correlations of item scores with total scores) below 0.2 were excluded from the item mapping process, as were items that had poor fit to the Rasch model (Infit or Outfit values below 0.5 or above 1.5). Through this process, items across all grades and all forms can be included in one large item map.

Figure 5 provides the Grade 3 items mapped to the Savvas vertical scale by domain. Colors indicate the specific skill associated with each item. Domains are indicated along the vertical axis, the Rasch theta scale (used to derive the Savvas scale scores) is represented along the horizontal axis. The further to the right on the scale an item/skill is, the harder it is. Some skills are observed in a relatively narrow region of the scale whereas others are more dispersed across the scale. As indicated in Figure 5, the Data and Probability (DP) and Measurement (ME) domains have relatively few items/skills and are mostly of moderate difficulty. The other four domains have a very wide spread of difficulty, allowing for measurement of students across a wide range of achievement. Figure 5, is for Grade 3 only; including all 900+ items from Grades K-8 would illustrate the much larger range of student ability measured by the Savvas Mathematics Diagnostic Assessment system.

Figure 5. Grade 3 Item Map



AR = Algebraic Reasoning; DP = Data & Probability; FR = Fractions; GR = Geometric Reasoning; ME = Measurement; NO = Number & Operations

Each item measures one or more skills. This provides a mechanism for translating an item map to a skill map to visualize the difficulty of skills along the performance continuum. The skill map was used to classify skills into three categories for each student: "Strengths", "Areas for Improvement" or "Stretch Goals" based on the items' IRT response probability values. That is, by comparing the difficulty of the item to the achievement level of the student, we can estimate how likely they are to respond correctly to an item measuring a specific skill. For example, for a moderately difficult item measuring a geometry skill, a very low performing student will have a low probability of answering the item correctly—this low response probability indicates that the geometry skill measured by the item is Stretch goal for the student—a skill that the student needs to work on to improve their understanding and demonstrate growth. A very high performing student will have a high probability of answering the item correctly—this is a Strength for the student, indicating that they have likely mastered the geometry skill associated with the item and they are ready to move on to more challenging geometry concepts.

Skills were considered within domain, rather than at the overall test level. Therefore, each student will have a set of Strengths, Areas for Improvement, and Stretch Goals for each of the domains measured by the assessment.

Defining Strengths. Strengths are skills measured by items for which a student has a high likelihood of a correct response. That is, the student's achievement level is high relative to the difficulty of the skill being measured by the item(s). Of course, skills were often measured by more than one item in the Savvas mathematics item pool. If the skill is measured by items that are dispersed across the scale, a student might have a 0.75 likelihood of correctly answering one item and a 0.45 likelihood of correctly answering another item, where both items measure the same skill. A conservative approach was taken to determine whether such a skill would be considered a Strength, Area for Improvement, or Stretch Goal. Specifically, the most difficult item within a skill was selected to represent the skill. Therefore, this skill would not be classified as a strength for the student (with only a 0.45 likelihood of success on the most difficult item measuring the skill). This conservative criterion ensures that the skill is a strength for the student based on the most difficult item measuring the skill.

Up to 3 skills per domain were selected as strengths for each student. Strengths were those skills having items with response probabilities closest to 0.67 and ranging from 0.55 to 1. A response probability of 0.67 was selected because it represents a reasonably high likelihood of success on an item measuring the skill and is commonly used to represent mastery in standard setting methods such as the Bookmark Procedure. A detailed description of the adoption of response probability 0.67 as a mastery criterion is provided by Lewis, Mitzel, Mercado & Schulz[19]. A 0.67 response probability indicates that a student would correctly answer items measuring the skill about 2 out of 3 times. The lower boundary of a 0.55 response probability was used to provide some flexibility in selecting skills, but does not allow the likelihood of success on items measuring skills classified as Strengths to become unreasonably low.

For low performing students, there may be no on-grade level skills with items meeting the response probability criterion. In these cases, when possible, the response probabilities of below-grade level items associated with the same domain were evaluated. As an example, if only one skill was found at grade 3 that could be classified as a Strength, then grade 2 items were evaluated. If two more skills were found with a response probability of at least 0.55, then the set of 3 skills were complete. Otherwise, items at grade 1 were evaluated.

After searching the below-grade items, if fewer than three skills are found meeting the criteria for Strengths, then less than three Strength skills would be included on the score report. Below-grade-level skills were only included if they were within two grade levels below on-grade level (e.g., for Grade 4, below-grade skills could be selected from Grades 2 and 3).

[19] Lewis, Mercado, Mitzel, & Schulz (2012). The Bookmark Standard Setting Procedure. Chapter in *Setting Performance Standards: Concepts, Methods, and Perspectives, Second Edition*. (ed: G. J. Cizek), Lawrence Erlbaum.

Above grade level skills were not considered for Strengths because students may not have had the opportunity to learn above-grade level skills and may be confused by reports suggesting that they have achieved above-grade level skills.

Strengths are not included for students with a raw score of 0 on a given domain. Without having answered any of the items within a domain correctly, they have not demonstrated any strengths in the domain.

Defining Areas for Improvement. A skill is identified as an Area for Improvement when the items measuring the skill have relatively low response probabilities, indicating the student is not likely to demonstrate the skill by correctly answering the items. A low response probability indicates that the student's achievement level is low relative to the difficulty of the skill being measured by the item(s). For Areas of Improvement, items measuring the skill have a target response probability of 0.33 and a possible response probability range of 0 to 0.45. The easiest item among those meeting these criteria was selected to represent the skill. This is again a conservative mechanism for ensuring that skills are not classified as *Areas for Improvement* when a student has a moderate to high probability of correctly answering some of the items measuring the skill. Up to two different skills were selected for Areas for Improvement. As an example, a skill measured by three items with response probabilities of 0.46, 0.33, and 0.21 would not be classified as an Area for Improvement because 0.46 is outside the 0 to 0.45 response probability range. The target response probability of 0.33 was selected because it indicates that the skill is challenging for the student—given 3 items measuring the skill, they would likely only answer one of them correctly.

If only one skill was found at grade 3 that could be classified as an Area for Improvement, then grade 2 items were evaluated. If an appropriate skill could not be identified in grade 2, then grade 1 skills were evaluated. If after searching two grade levels below on grade level, less than two skills appropriate for Areas for Improvement could be found, then less than two Areas for Improvement were included on the score report.

Above grade levels skills were not considered for Areas for Improvement because students may not have had the opportunity to learn above grade level skills and may be confused by reports suggesting that they need to improve on above-grade level skills.

Areas for Improvement are not included for students who answered all items correctly within a domain. By correctly answering all items, they have not demonstrated any Areas for Improvement.

Defining Stretch Goals. Stretch Goals are skills that are a little harder and require more effort for students to achieve than Areas for Improvement. The response probabilities for these items are therefore very low. For Stretch Goals, items measuring the skill have a target response probability of 0.25 and a possible response probability range of 0 to 0.4.

The easiest item among those meeting these criteria was selected to represent the skill. One skill per domain was selected as a Stretch Goal. Because this response probability range overlaps with the Areas for Improvement range, the Stretch Goal skill was required to have a lower response probability than the lowest response probability of the Areas for Improvement skills.

Stretch Goals can be on-grade level skills, below-grade level skills for low performing students, or above-grade level skills for high performing students. Above-grade level skills are not to be interpreted as requirements for high performing students, but rather as enrichment opportunities.

Stretch goals are not provided for students who answered all items within a domain correctly.

Appendix A: Screener Standards and Item Counts

Table 10. Screener Standards and Item Counts

Grade Level	*Standard*	*Number of Items*
Kindergarten	COG.1	*3*
	COG.2	*3*
	COG.3	*7*
	COG.4	*3*
	COG.5	*3*
	COG.7	*3*
Grade 1	K.CC.A.1	2
	K.CC.A.3	3
	K.CC.B.4	3
	K.OA.A.2	4
	K.NBT.A.1	2
	K.MD.A.2	2
	K.MD.B.3	2
	K.G.A.2	2
	K.G.B.6	2
Grade 2	*1.OA.A.1*	3
	1.OA.B.4	2
	1.OA.C.6	3
	1.OA.D.7	2
	1.NBT.A.1	3
	1.NBT.B.2	4
	1.MD.A.2	4
	1.G.A.1	3
Grade 3	*2.OA.A.1*	3
	2.OA.B.2	3
	2.OA.C.4	5
	2.NBT.A.2	4
	2.MD.A.1	3
	2.G.A.1	3
	2.G.A.3	6
Grade 4	*3.OA.A.3*	4
	3.OA.B.6	2
	3.OA.C.7	3
	3.NBT.A.2	3
	3.NF.A.1	3
	3.NF.A.3	6
	3.MD.C.7b	3

Grade Level	Standard	Number of Items
	3.G.A.1	3
Grade 5	*4.NBT.B.4*	3
	4.NBT.B.5	3
	4.NBT.B.6	3
	4.NF.A.1	3
	4.NF.B.3	3
	4.NF.C.6	4
	4.MD.A.3	3
	4.G.A.1	3
	4.G.A.2	2
Grade 6	*5.OA.A.1*	4
	5.OA.A.2	4
	5.NBT.B.7	4
	5.NF.A.1	4
	5.NF.B.4	4
	5.NF.B.7	5
	5.MD.A.1	3
	5.G.A.2	4
Grade 7	*6.RP.A.3*	5
	6.NS.A.1	4
	6.NS.C.6	5
	6.EE.A.3	4
	6.EE.B.7	4
	6.G.A.1	4
	6.G.A.2	4
	6.SP.A.3	2
Grade 8	*7.RP.A.2*	7
	7.NS.A.3	5
	7.EE.A.1	5
	7.EE.B.4	5
	7.G.B.5	5
	7.G.B.6	5

Appendix B: Diagnostic Clusters and Item Counts

Table 11. Diagnostic Clusters and Item Counts

Grade Level	*Cluster*	*Number of Items*
Kindergarten	K.CC.A	*15*
	K.CC.B	*18*
	K.CC.C	*9*
	K.OA.A	*30*
	K.NBT.A	*6*
	K.MD.A	*9*
	K.MD.B	*6*
	K.G.A	*18*
	K.G.B	*18*
Grade 1	1.OA.A	15
	1.OA.B	12
	1.OA.C	9
	1.OA.D	6
	1.NBT.A	9
	1.NBT.B	15
	1.NBT.C	15
	1.MD.A	12
	1.MD.B	6
	1.MD.C	9
	1.G.A	21
Grade 2	*2.OA.A*	9
	2.OA.B	6
	2.OA.C	9
	2.NBT.A	24
	2.NBT.B	21
	2.MD.A	18
	2.MD.C	12
	2.MD.D	12
	2.G.A	18
Grade 3	*3.OA.A*	15
	3.OA.B	9
	3.OA.C	9
	3.OA.D	15
	3.NBT.A	12
	3.NF.A	20
	3.MD.A	9
	3.MD.B	12

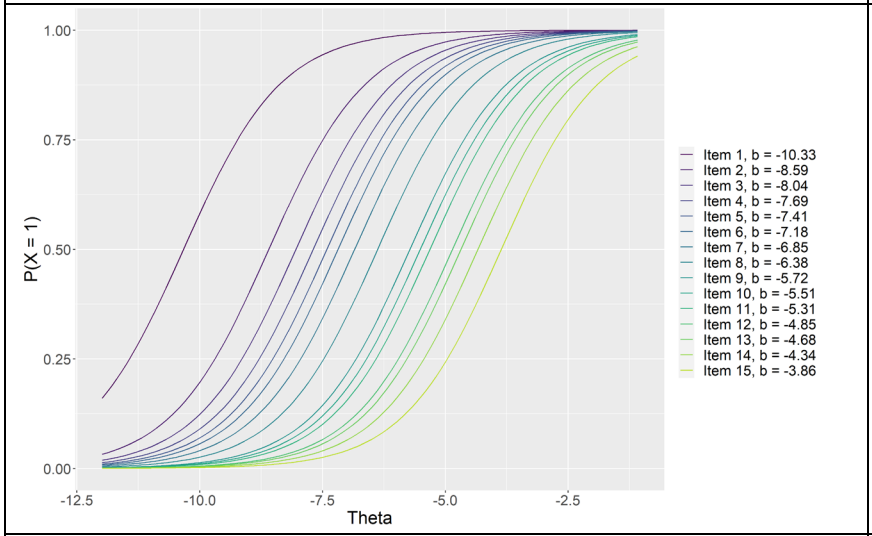
Grade Level	Cluster	Number of Items
	3.MD.C	16
	3.MD.D	6
	3.G.A	6
Grade 4	*4.OA.A*	15
	4.OA.B	6
	4.OA.C	6
	4.NBT.A	9
	4.NBT.B	24
	4.NF.A	12
	4.NF.B	12
	4.NF.C	9
	4.MD.A	9
	4.MD.B	3
	4.MD.C	14
	4.G.A	9
Grade 5	*5.OA.A*	9
	5.NBT.A	19
	5.NBT.B	21
	5.NF.A	18
	5.NF.B	27
	5.MD.A	6
	5.MD.B	6
	5.MD.C	13
	5.G.A	3
	5.G.B	6
Grade 6	*6.RP.A*	21
	6.NS.A	6
	6.NS.B	3
	6.NS.C	23
	6.EE.A	15
	6.EE.B	21
	6.G.A	18
	6.SP.A	6
	6.SP.B	15
Grade 7	*7.RP.A*	21
	7.NS.A	30
	7.EE.A	9
	7.EE.B	18
	7.G.A	6
	7.G.B	20
	7.SP.A	3

Grade Level	Cluster	Number of Items
	7.SP.B	6
	7.SP.C	15
Grade 8	*8.NS.A*	9
	8.EE.A	12
	8.EE.B	8
	8.EE.C	21
	8.F.A	12
	8.F.B	9
	8.G.A	18
	8.G.B	9
	8.G.C	9
	8.SP.A	21

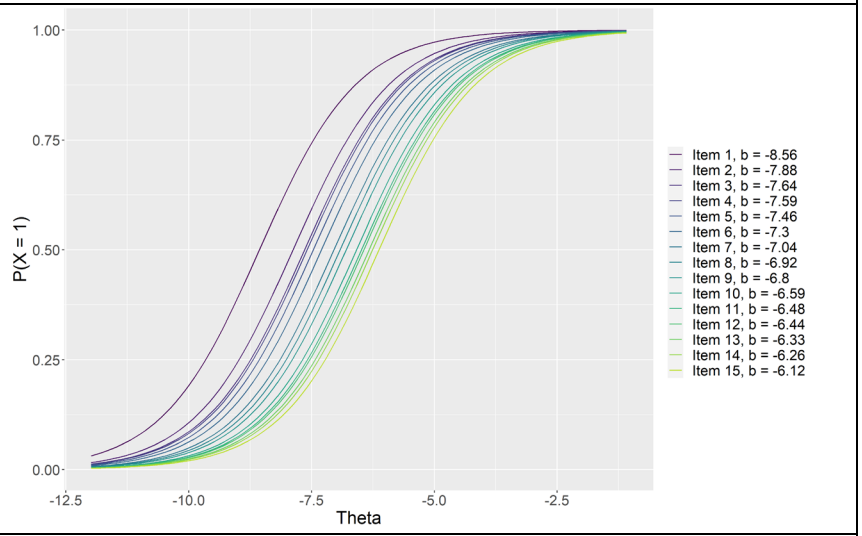
Appendix C: Rasch Calibration Results for the Diagnostic Assessments by Grade

Figure 6. Diagnostic Assessment Item Characteristic Curves, Grade K

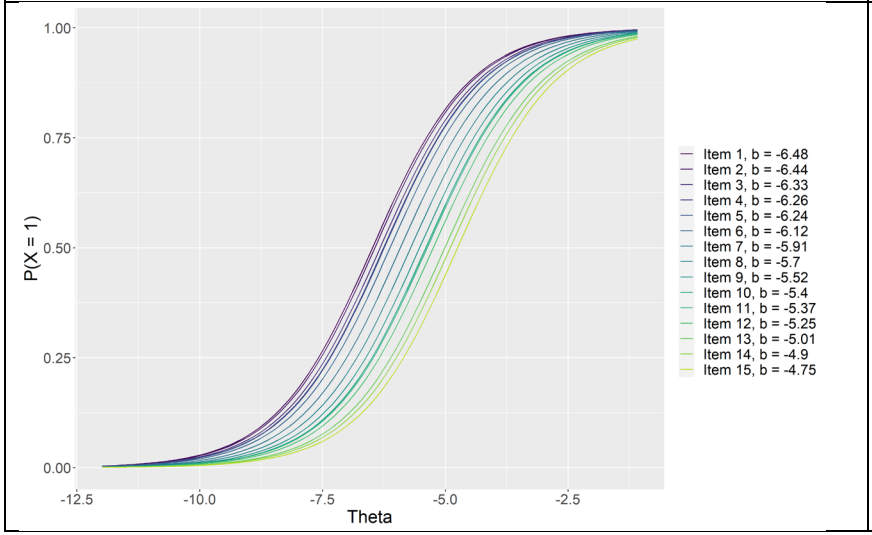
Router



Low



Moderate



High

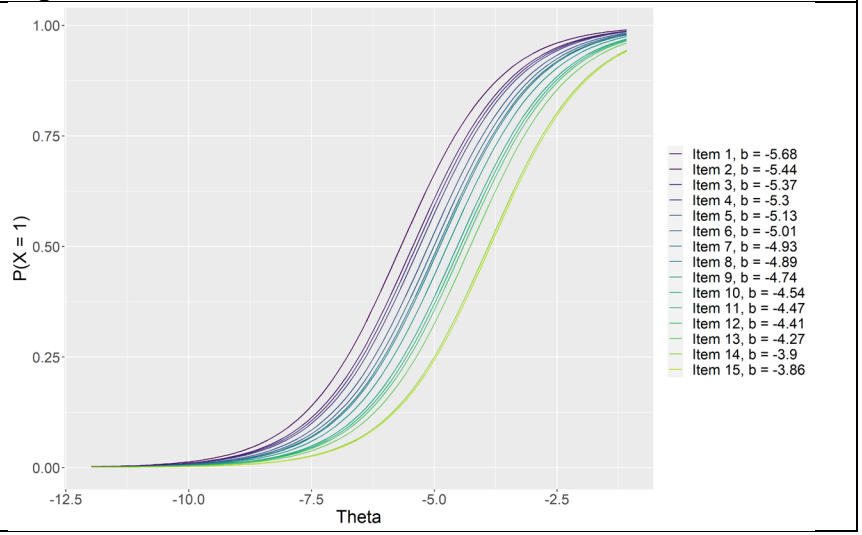


Figure 7. Diagnostic Assessment Item Characteristic Curves, Grade 1

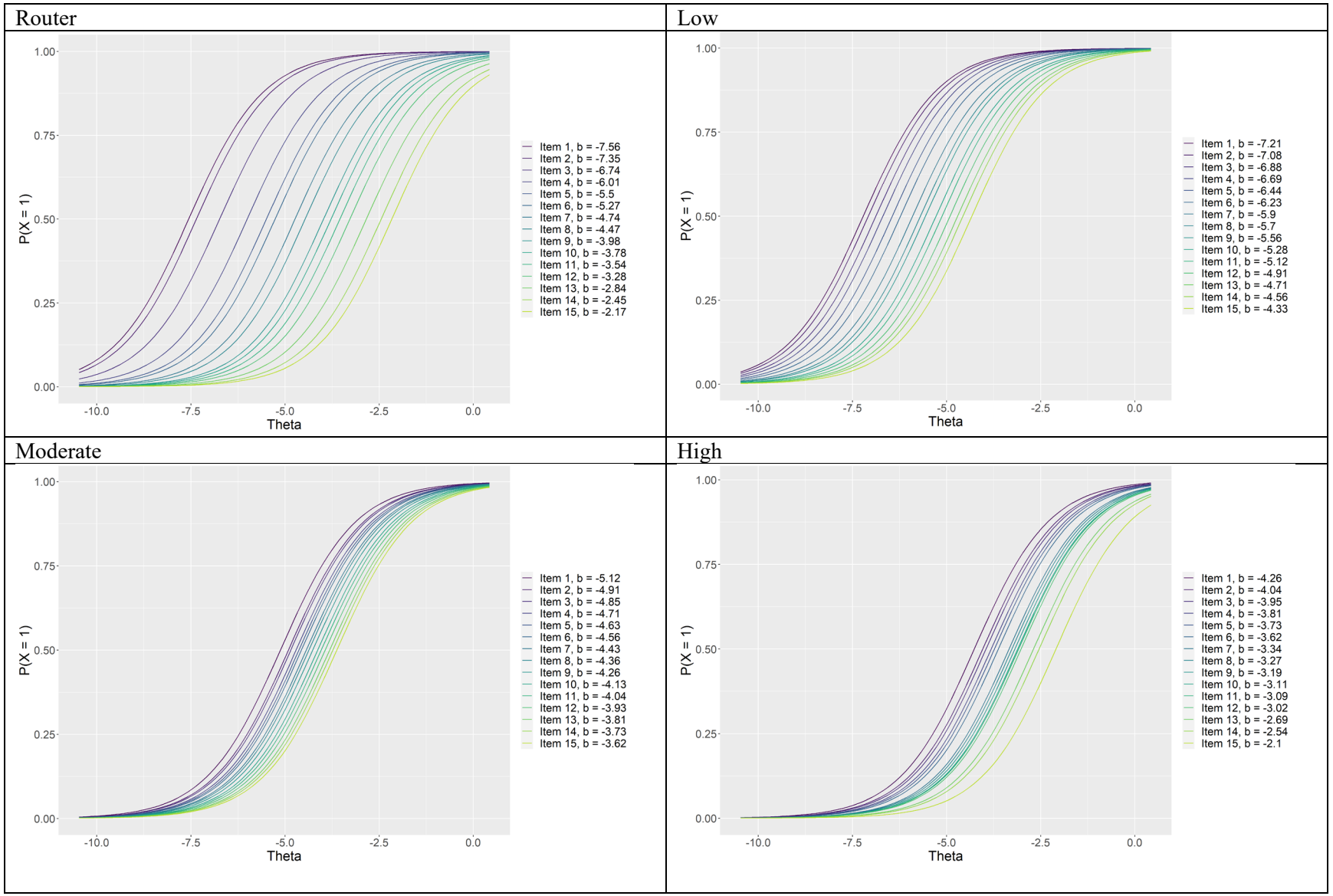
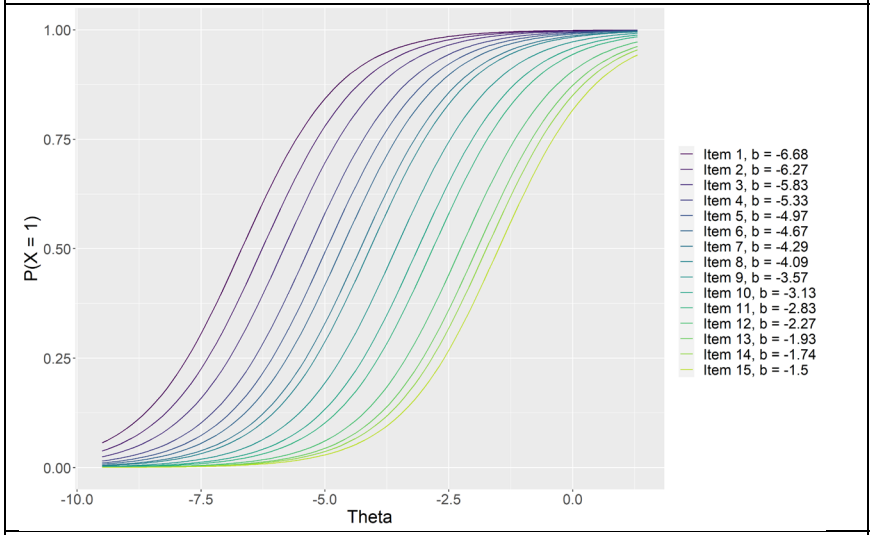
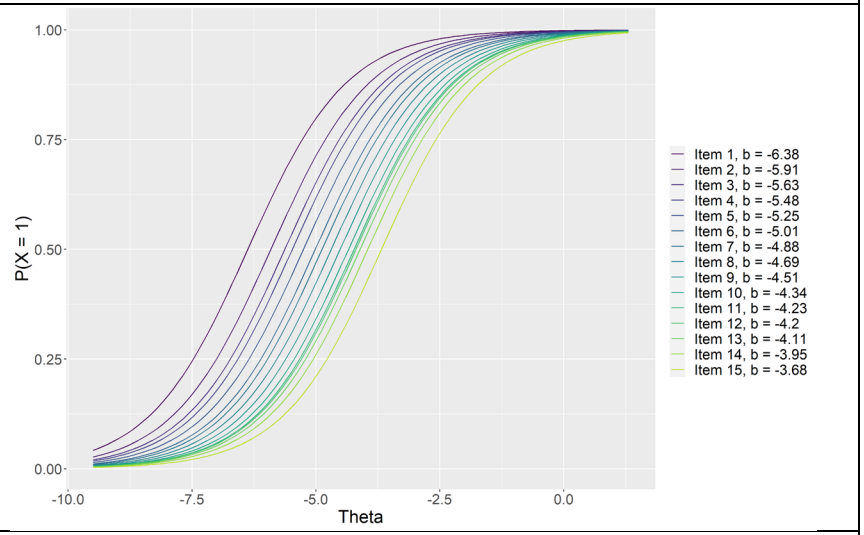


Figure 8. Diagnostic Assessment Item Characteristic Curves, Grade 2

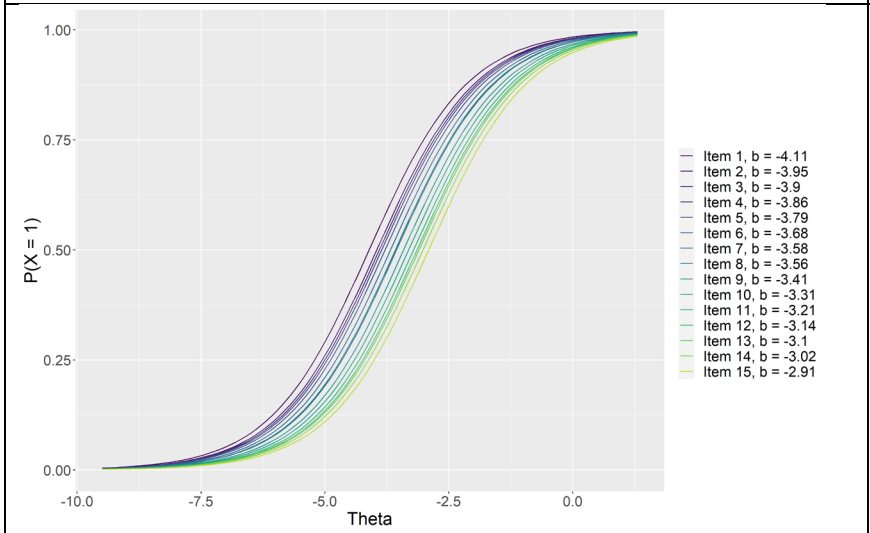
Router



Low



Moderate



High

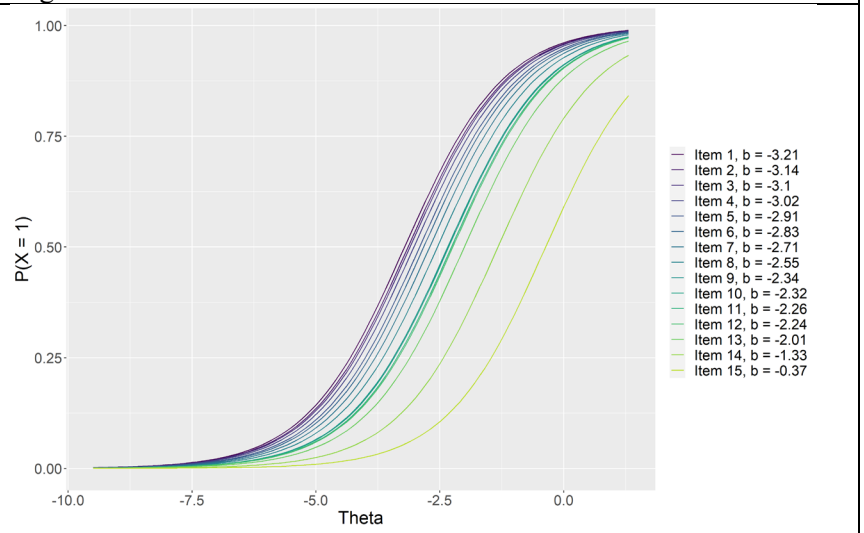
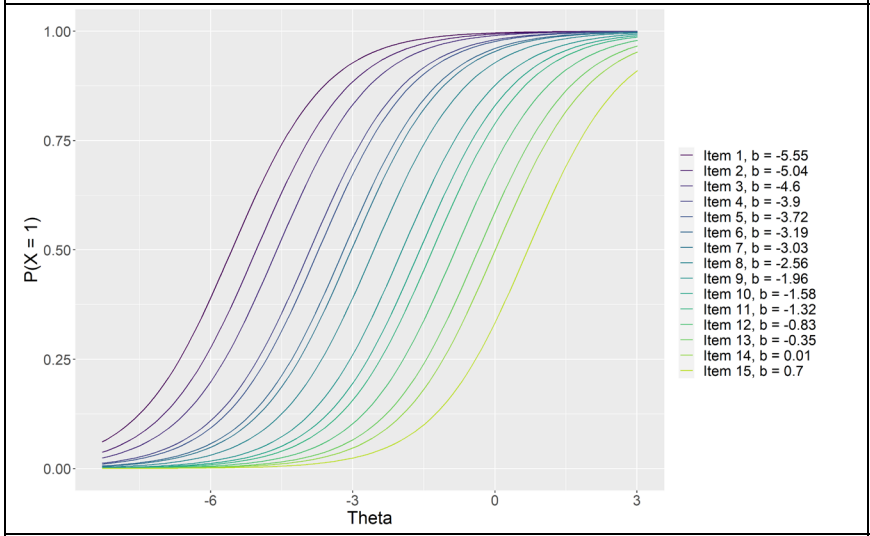
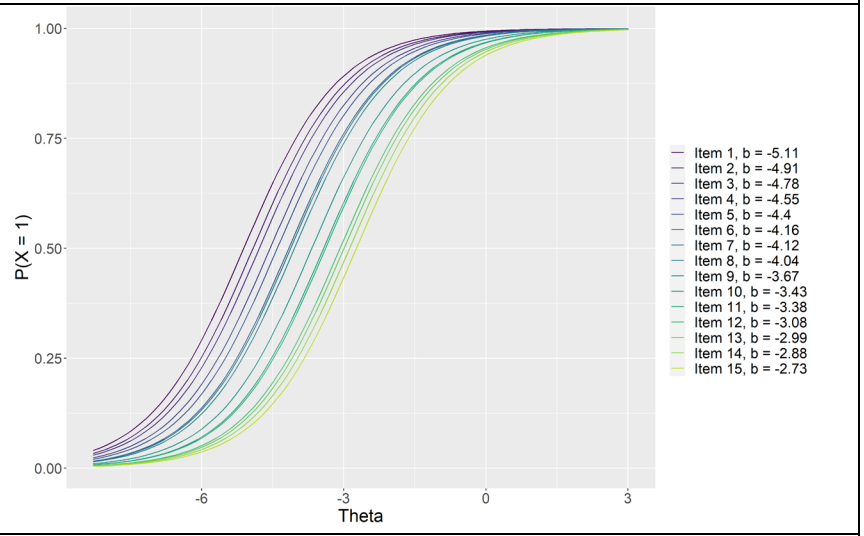


Figure 9. Diagnostic Assessment Item Characteristic Curves, Grade 3

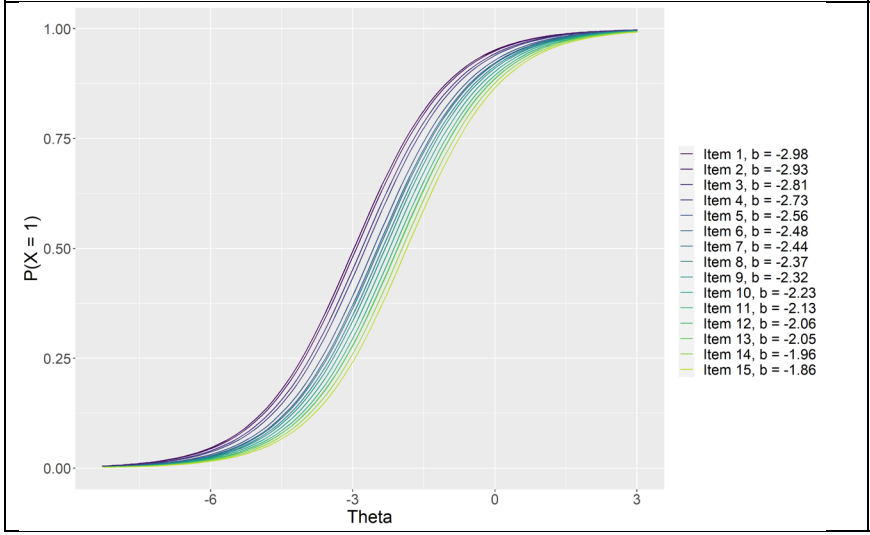
Router



Low



Moderate



High

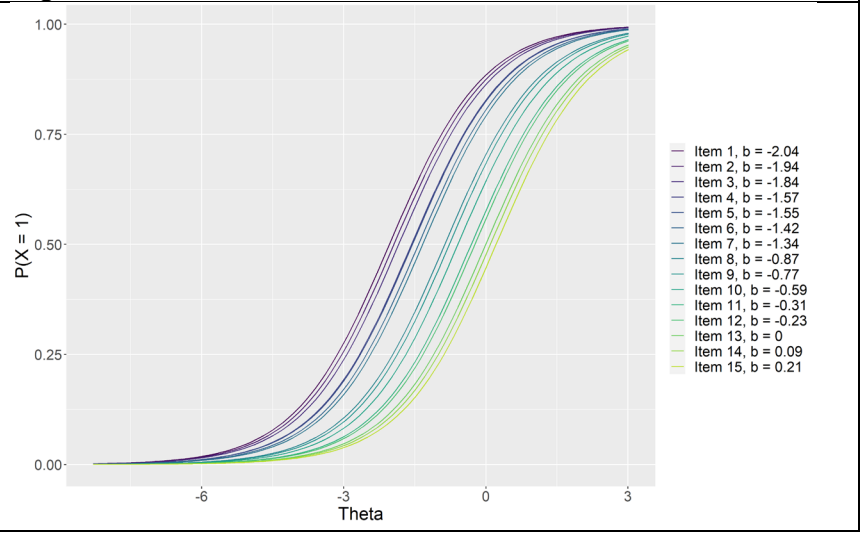
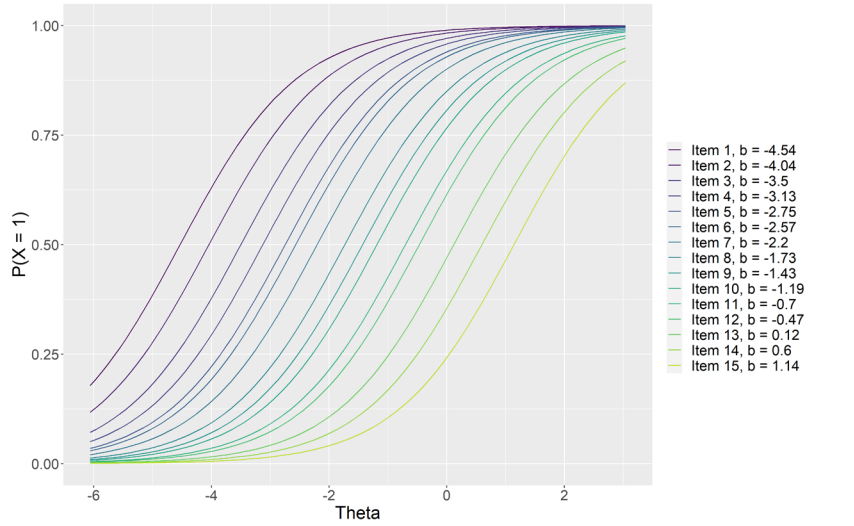
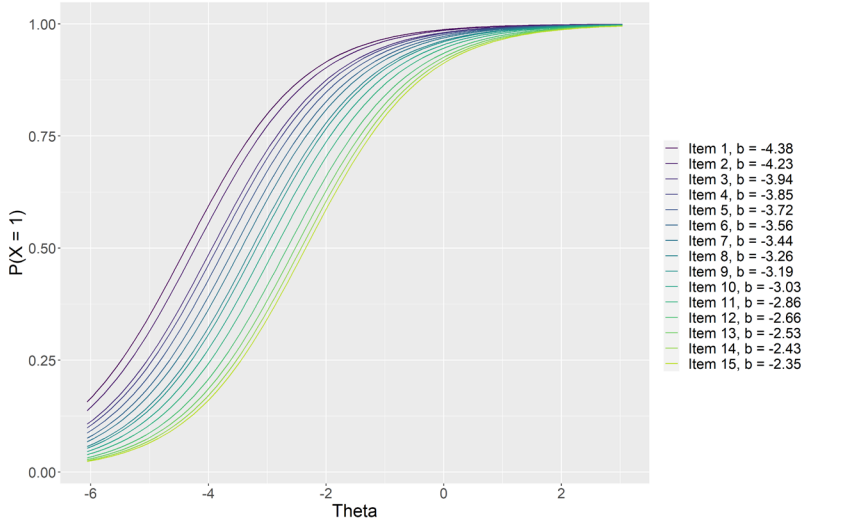


Figure 10. Diagnostic Assessment Item Characteristic Curves, Grade 4

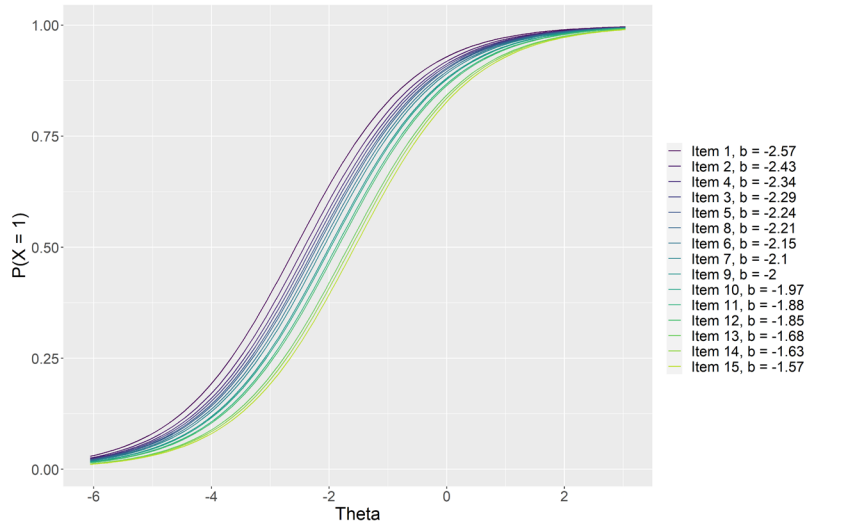
Router



Low



Moderate



High

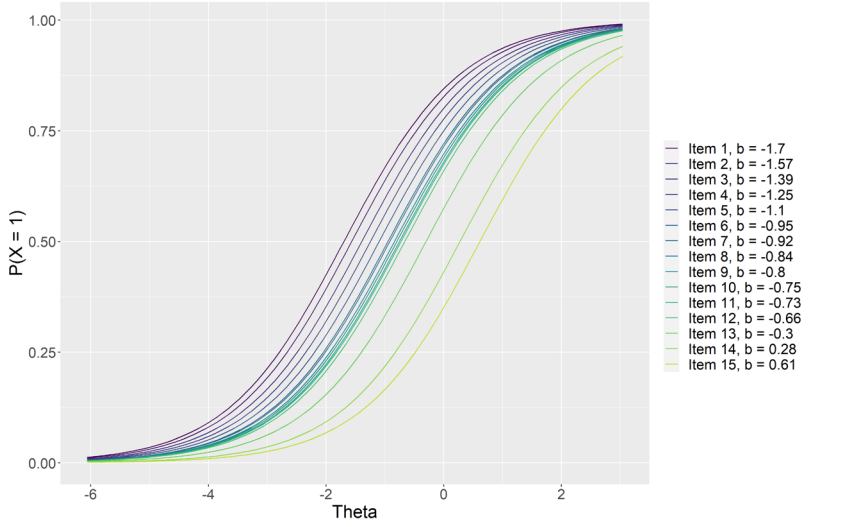


Figure 11. Diagnostic Assessment Item Characteristic Curves, Grade 5

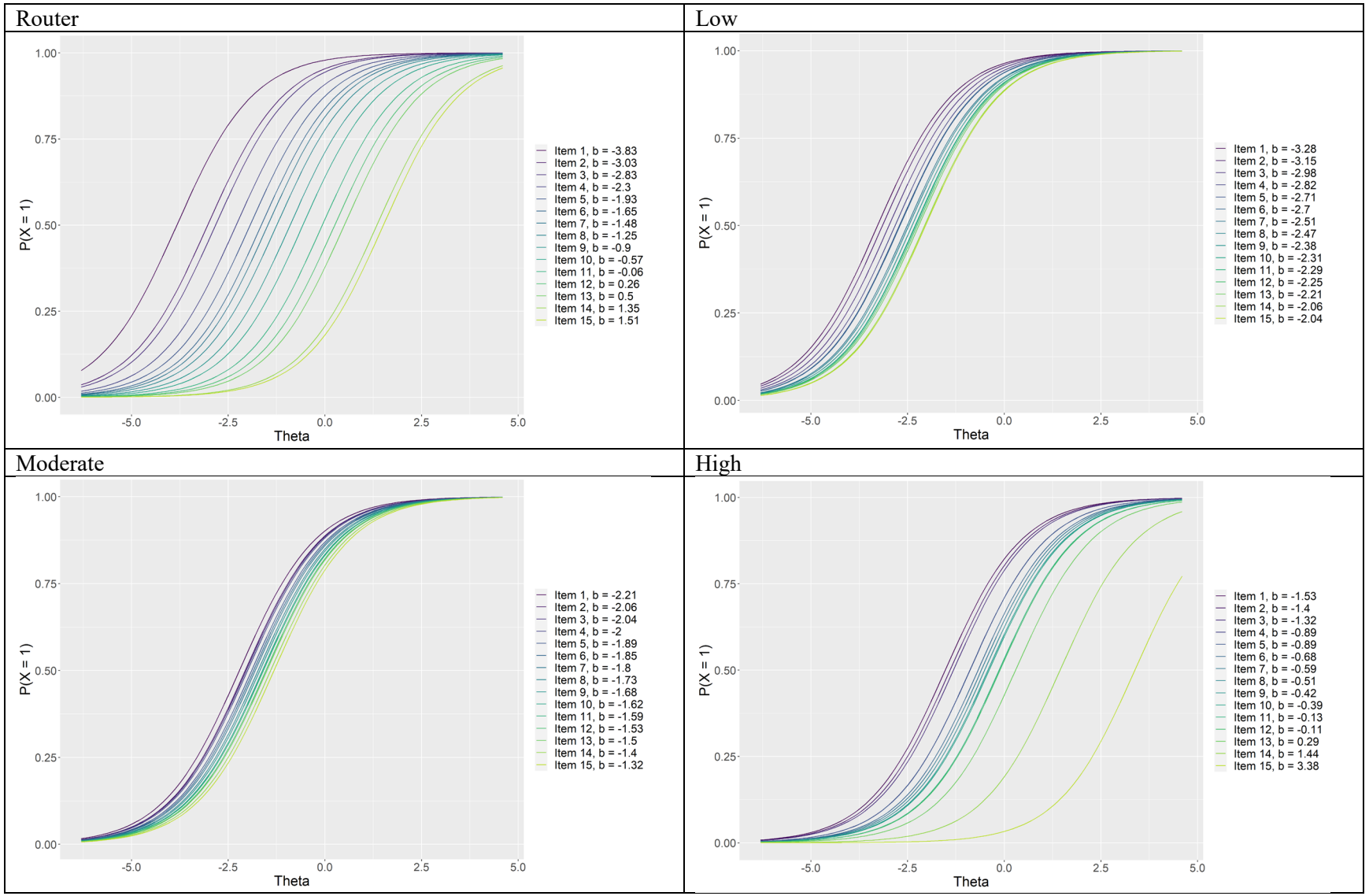
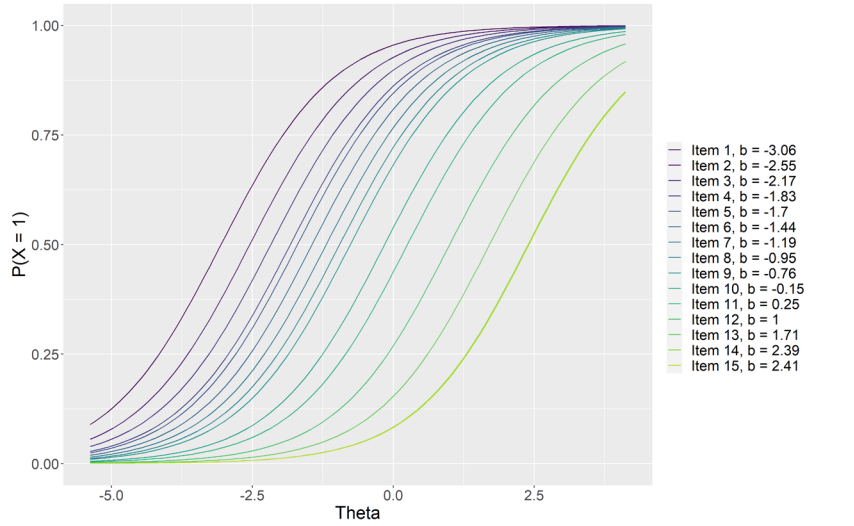
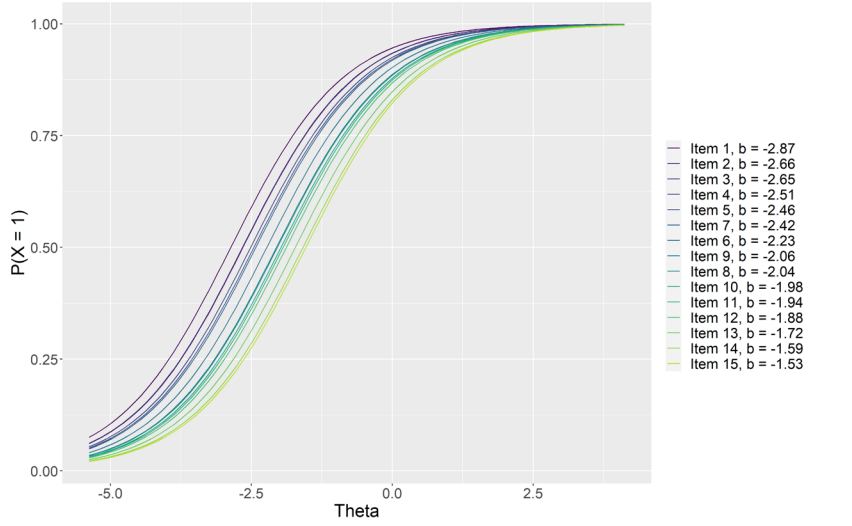


Figure 12. Diagnostic Assessment Item Characteristic Curves, Grade 6

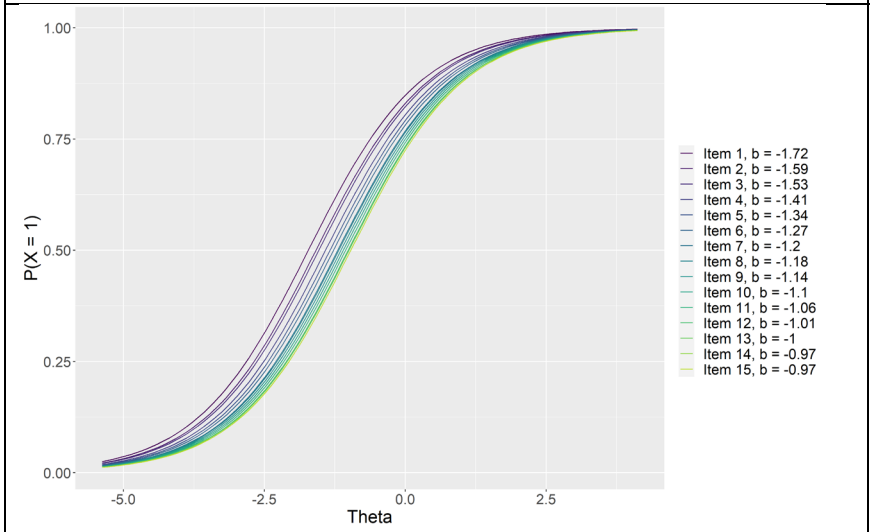
Router



Low



Moderate



High

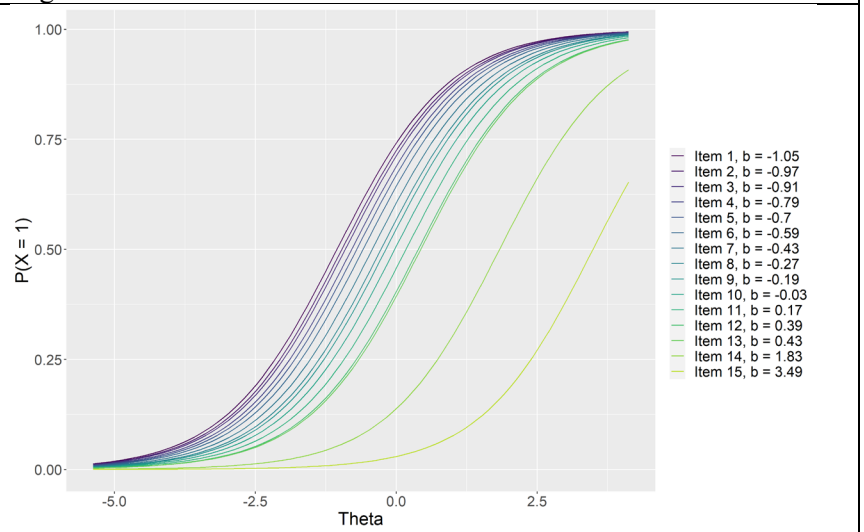


Figure 13. Diagnostic Assessment Item Characteristic Curves, Grade 7

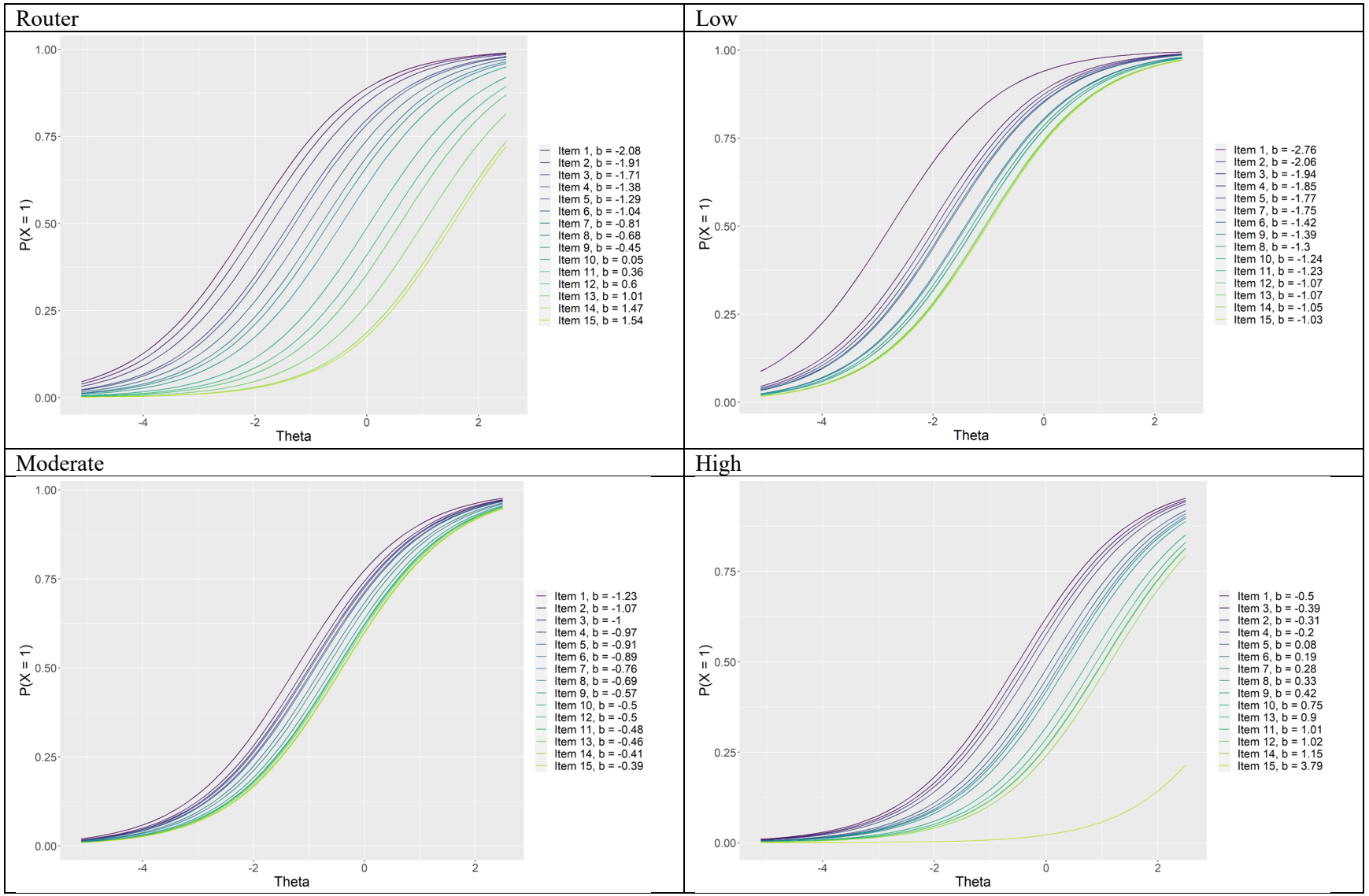


Figure 14. Diagnostic Assessment Item Characteristic Curves, Grade 8

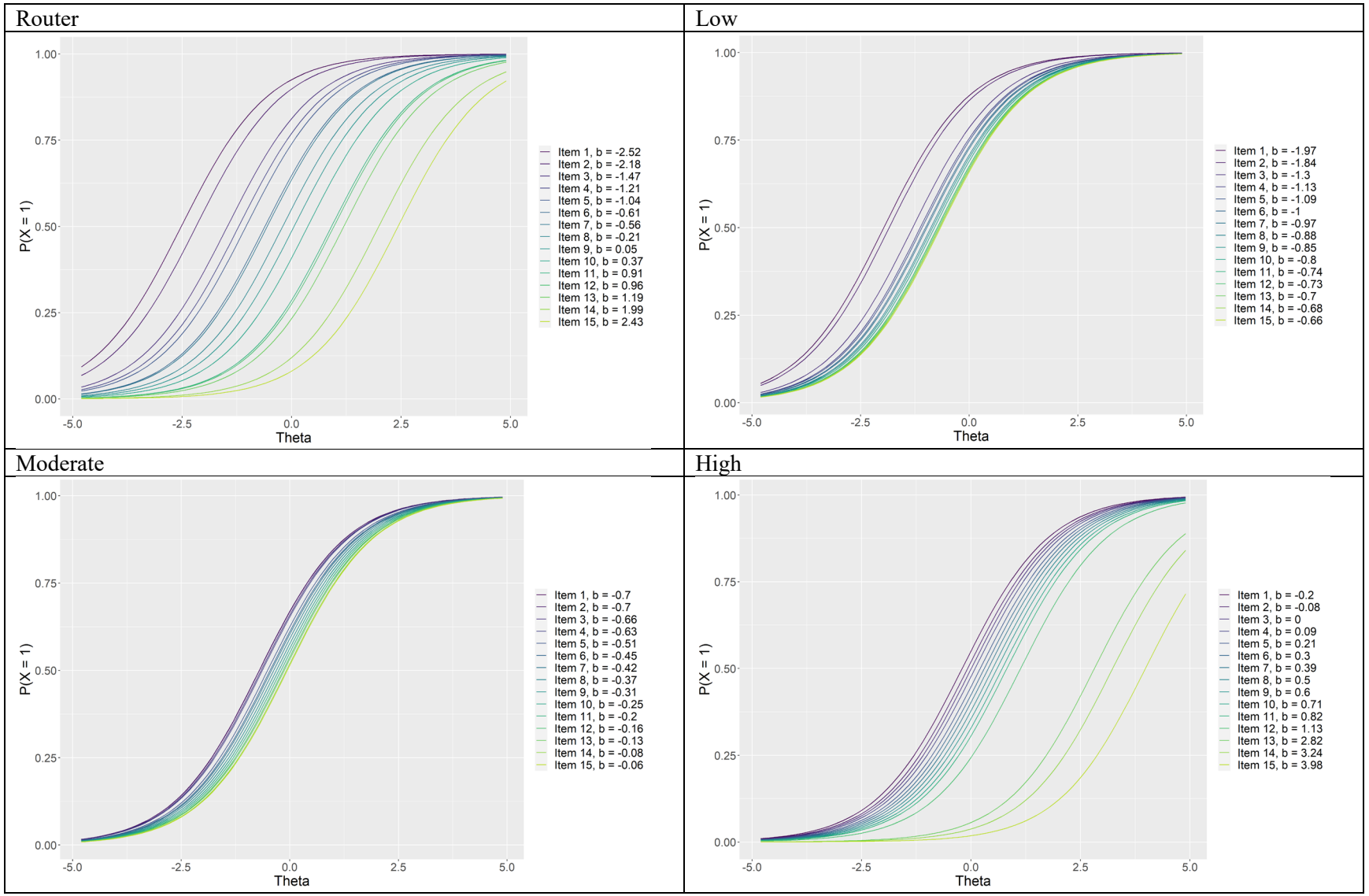


Figure 15. Frequency Distribution, Grade K

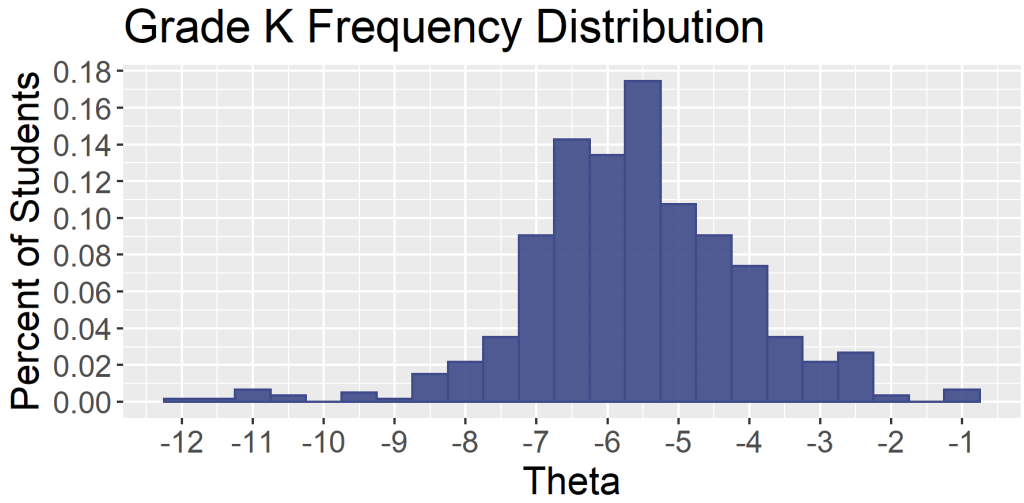


Figure 16. Diagnostic Test Characteristic Curves, Grade K

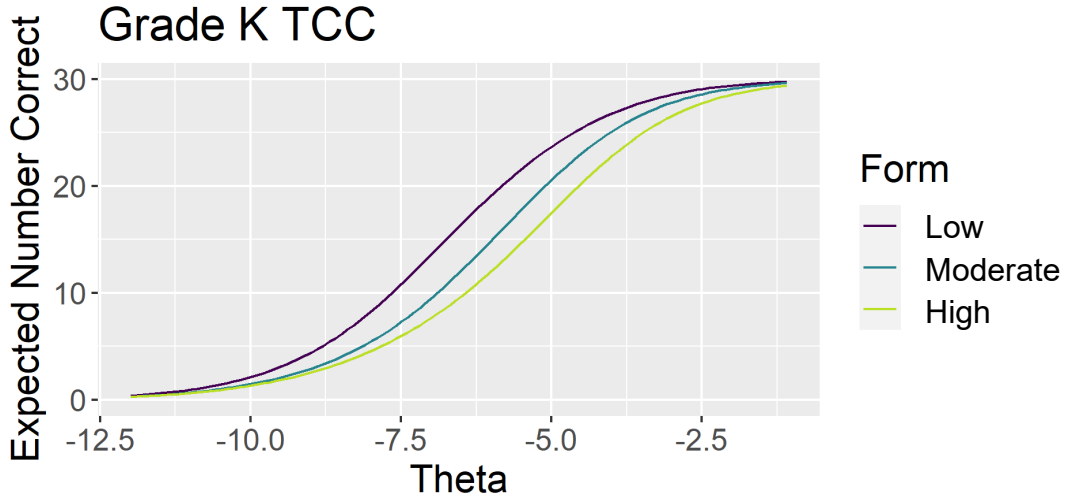


Figure 17. Diagnostic Test Information Function Curves and SEM, Grade K

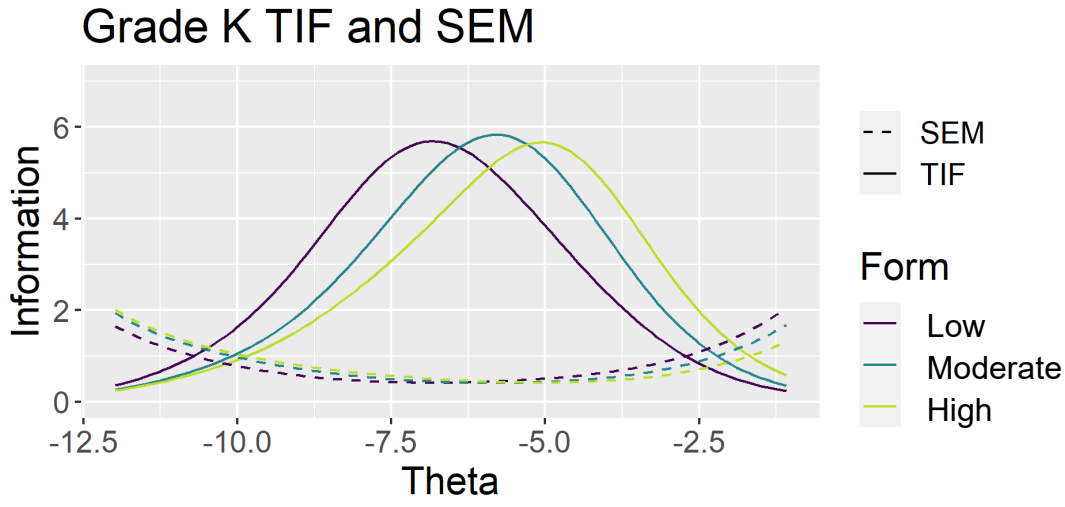


Figure 18. Frequency Distribution, Grade 1

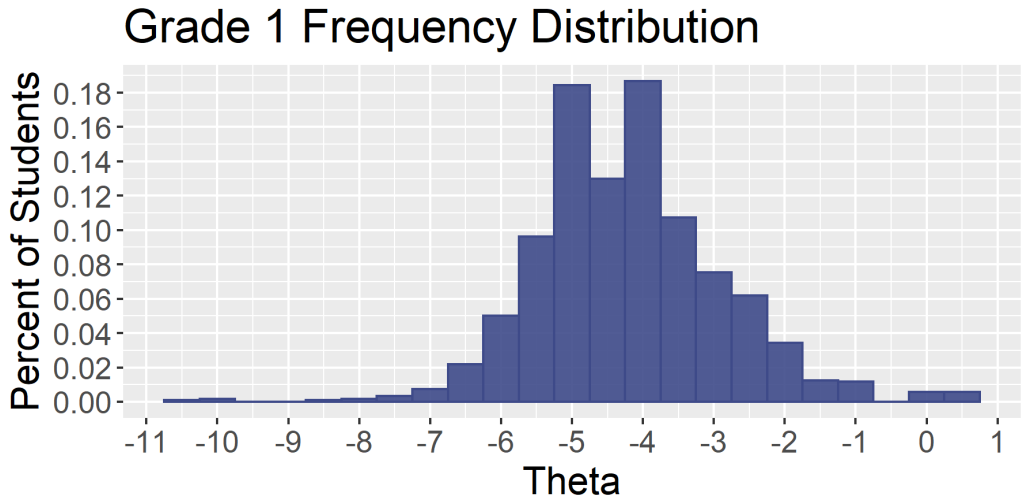


Figure 19. Diagnostic Test Characteristic Curves, Grade 1

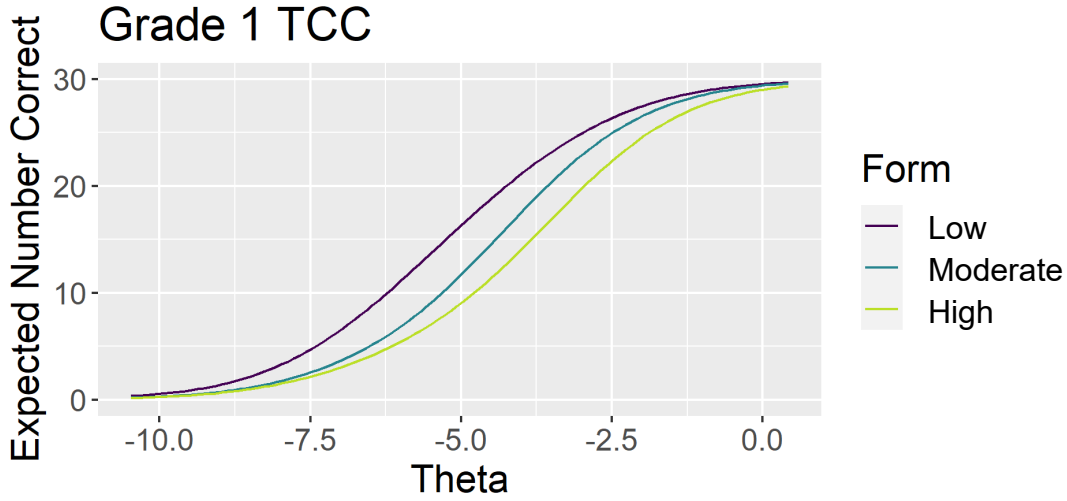


Figure 20. Diagnostic Test Information Function Curves and SEM, Grade 1

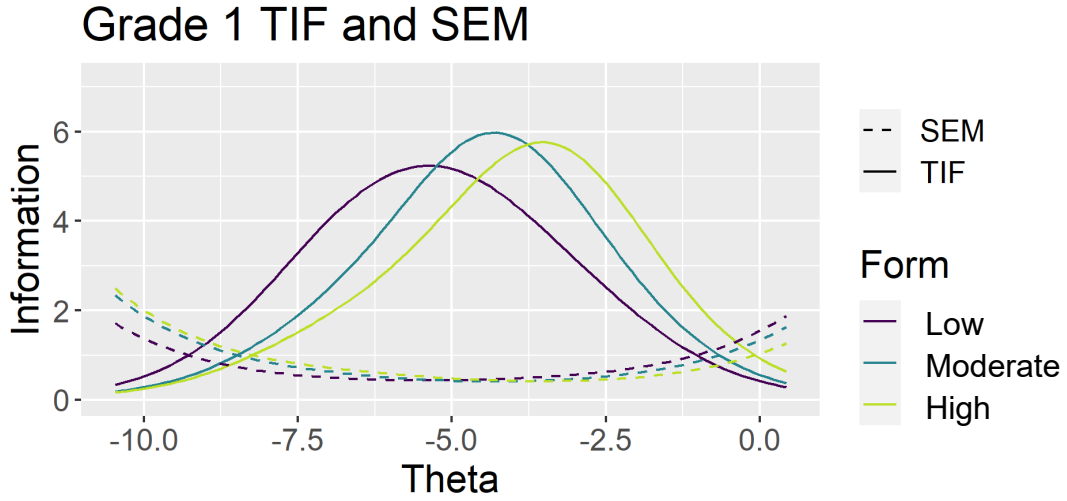


Figure 21. Frequency Distribution, Grade 2

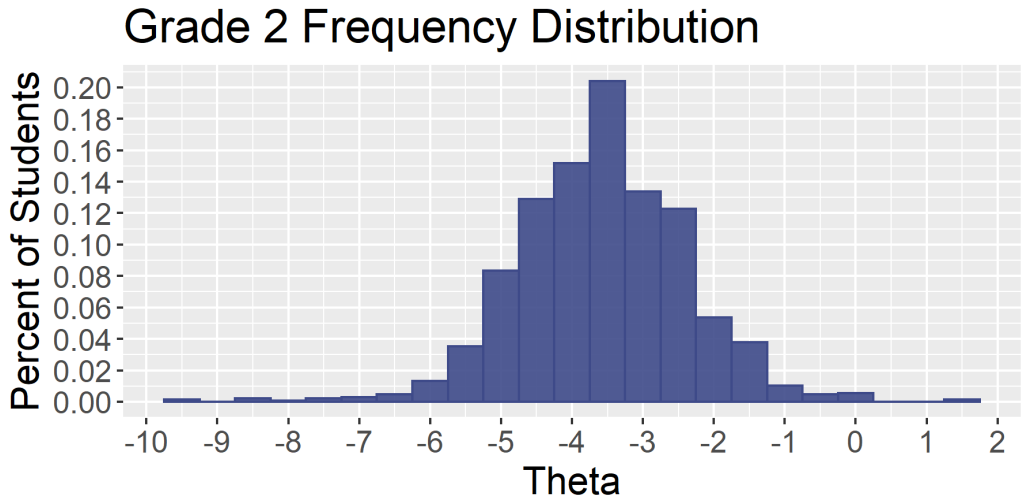


Figure 22. Diagnostic Test Characteristic Curves, Grade 2

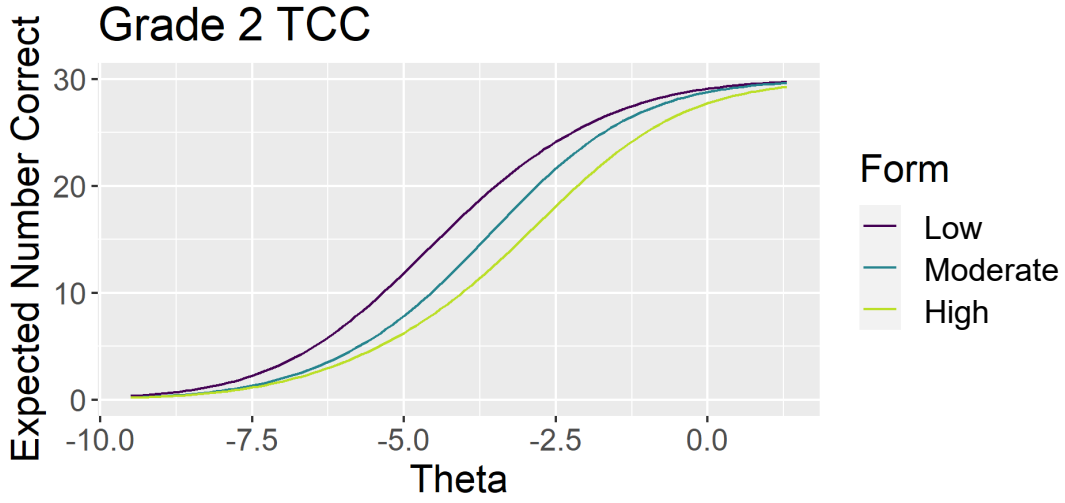


Figure 23. Diagnostic Test Information Function Curves and SEM, Grade 2

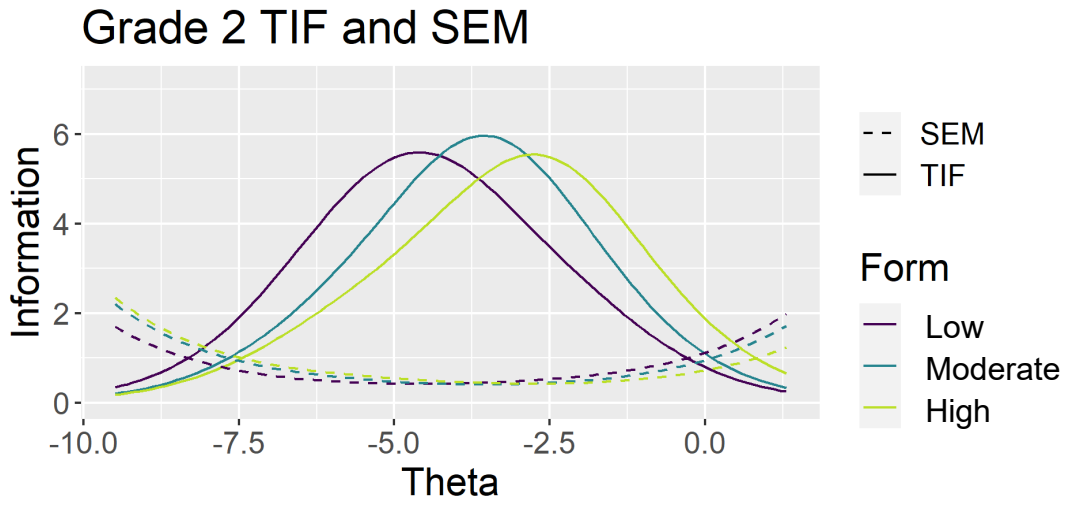


Figure 24. Frequency Distribution, Grade 3

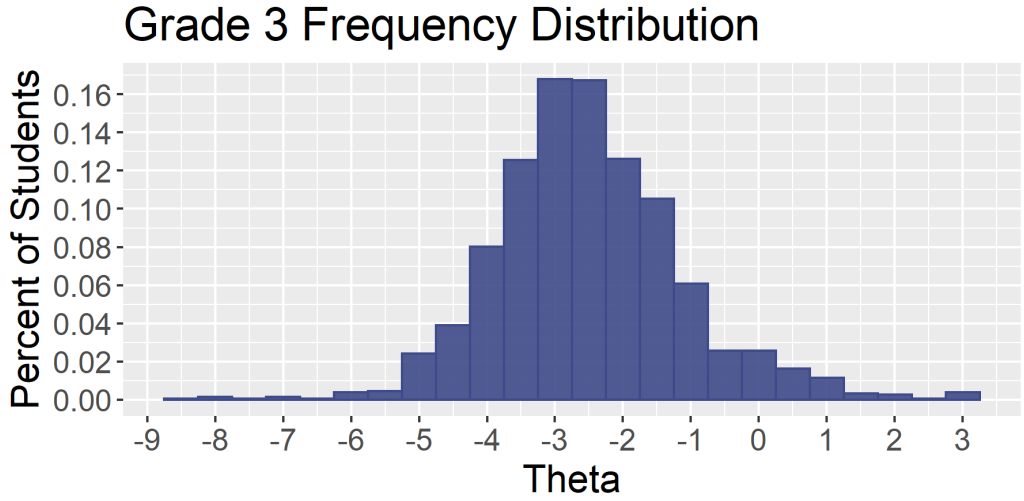


Figure 25. Diagnostic Test Characteristic Curves, Grade 3

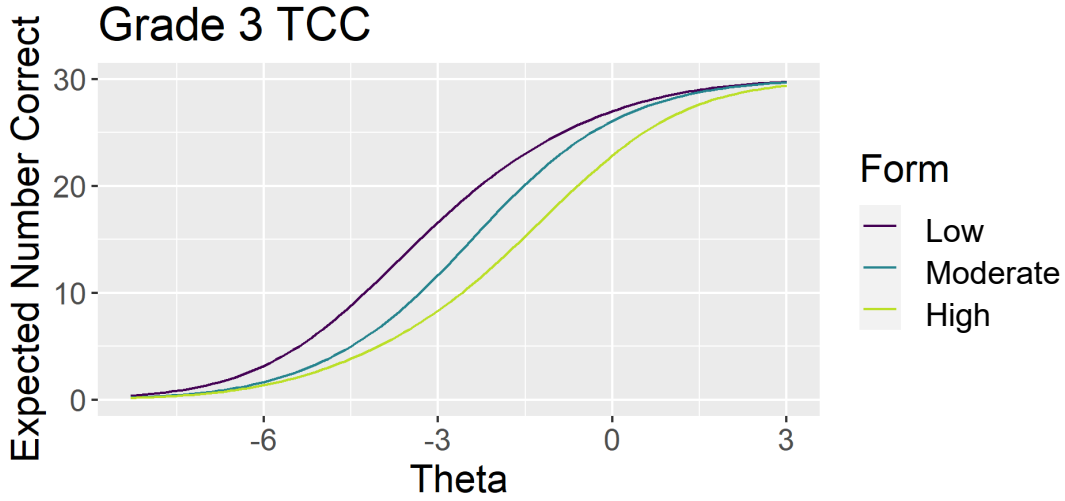


Figure 26. Diagnostic Test Information Function Curves and SEM, Grade 3

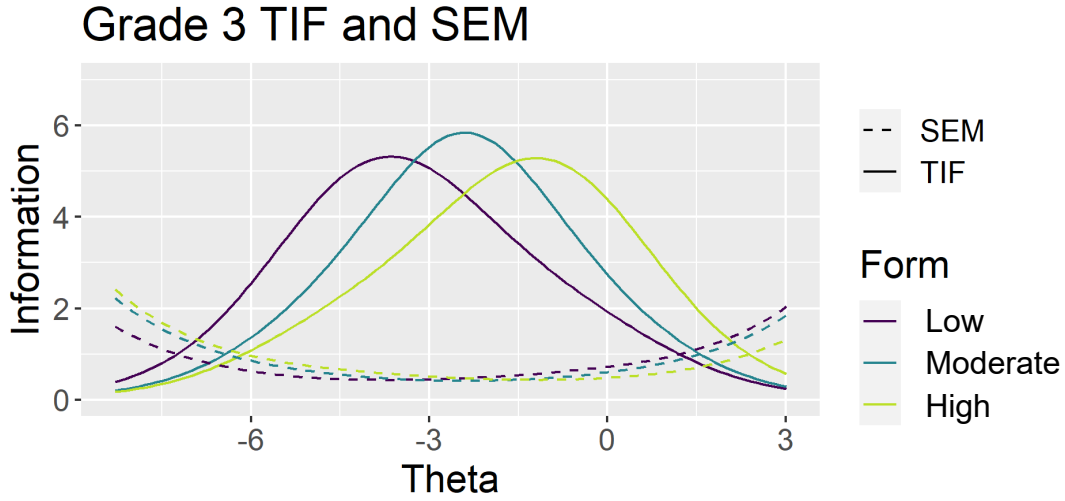


Figure 27. Frequency Distribution, Grade 4

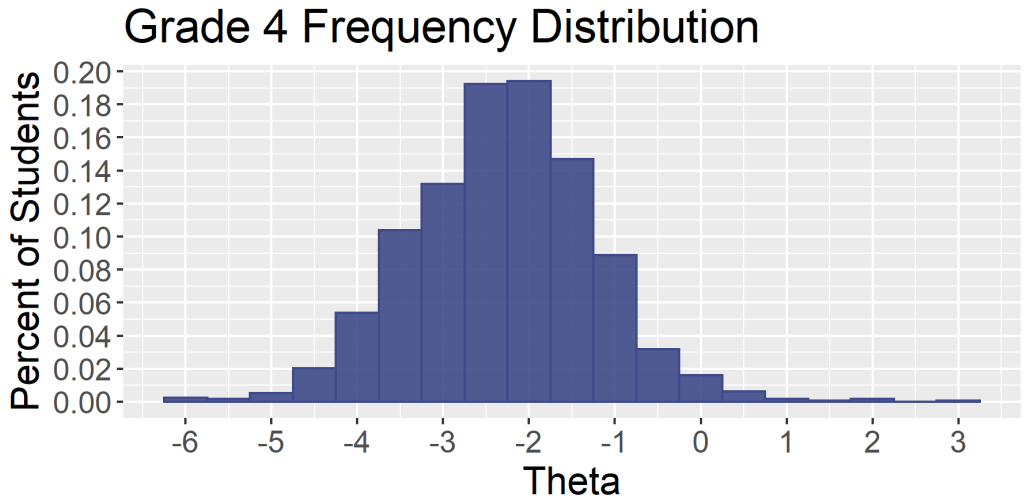


Figure 28. Diagnostic Test Characteristic Curves, Grade 4

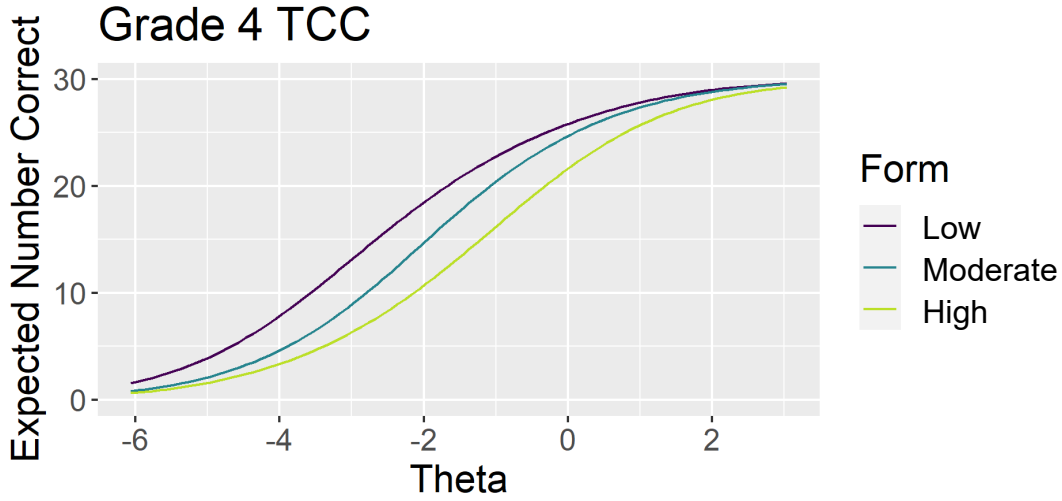


Figure 29. Diagnostic Test Information Function Curves and SEM, Grade 4

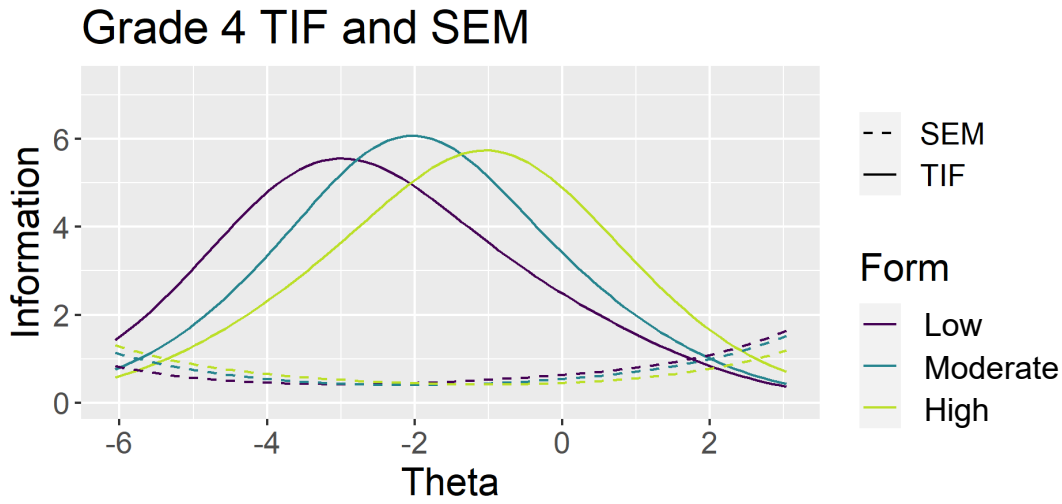


Figure 30. Frequency Distribution, Grade 5

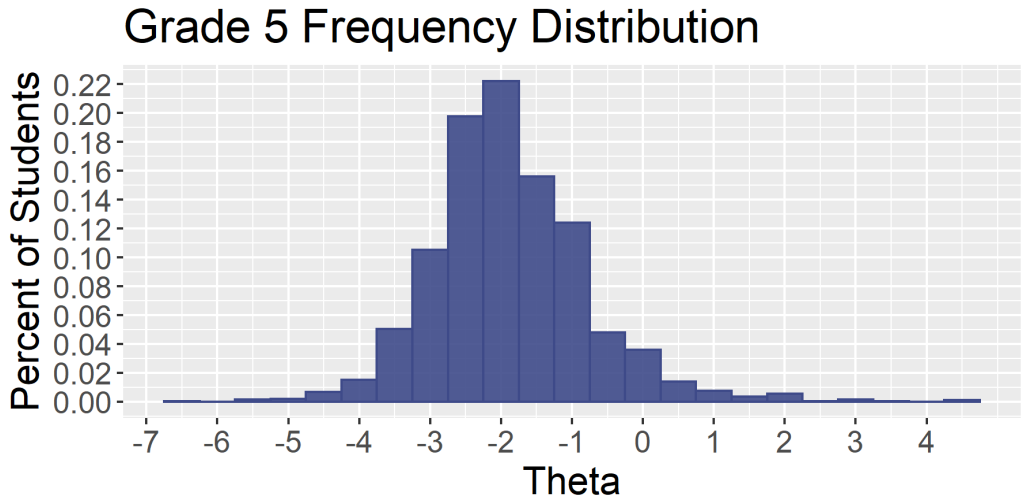


Figure 31. Diagnostic Test Characteristic Curves, Grade 5

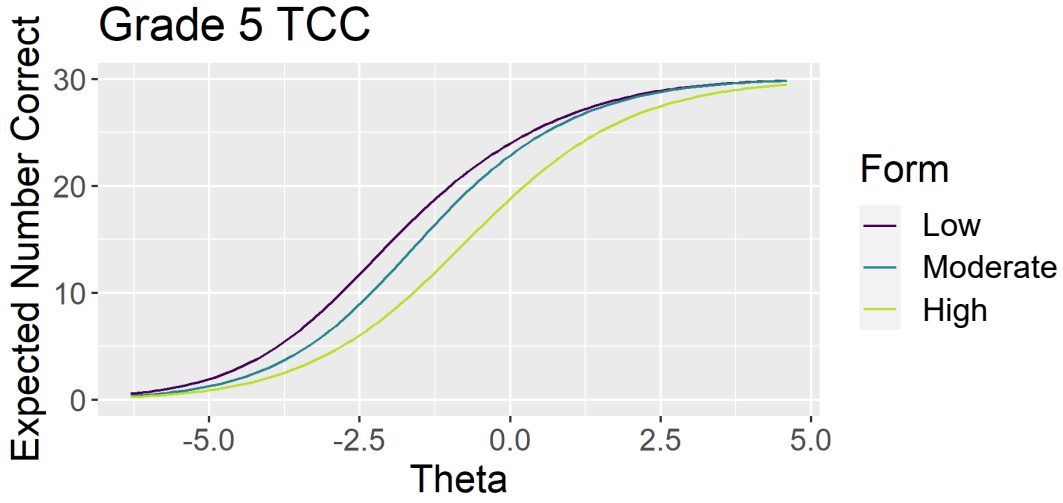


Figure 32. Diagnostic Test Information Function Curves and SEM, Grade 5

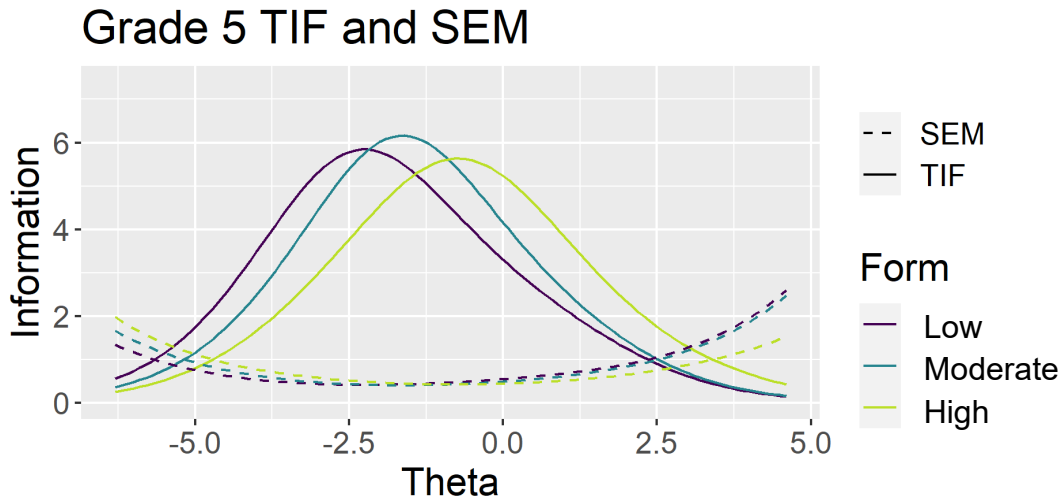


Figure 33. Frequency Distribution, Grade 6

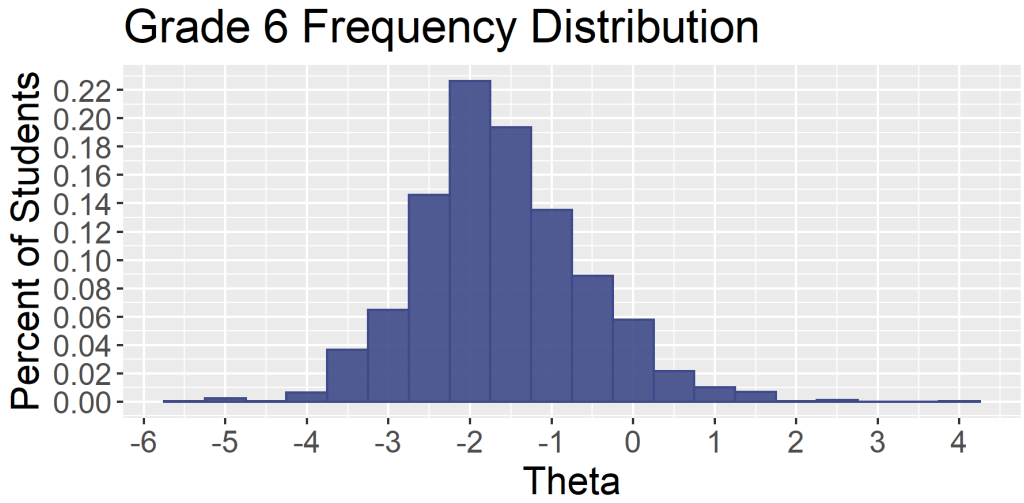


Figure 34. Diagnostic Test Characteristic Curves, Grade 6

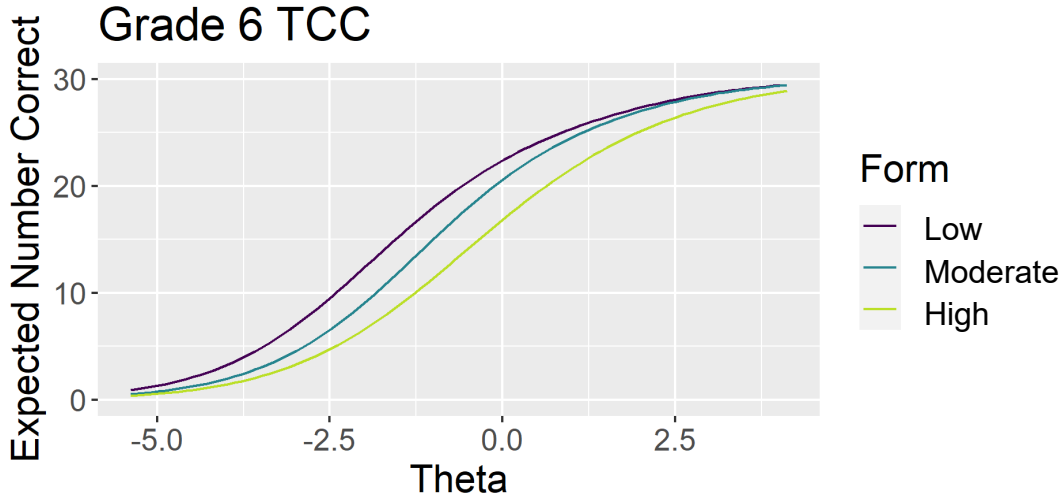


Figure 35. Diagnostic Test Information Function Curves and SEM, Grade 6

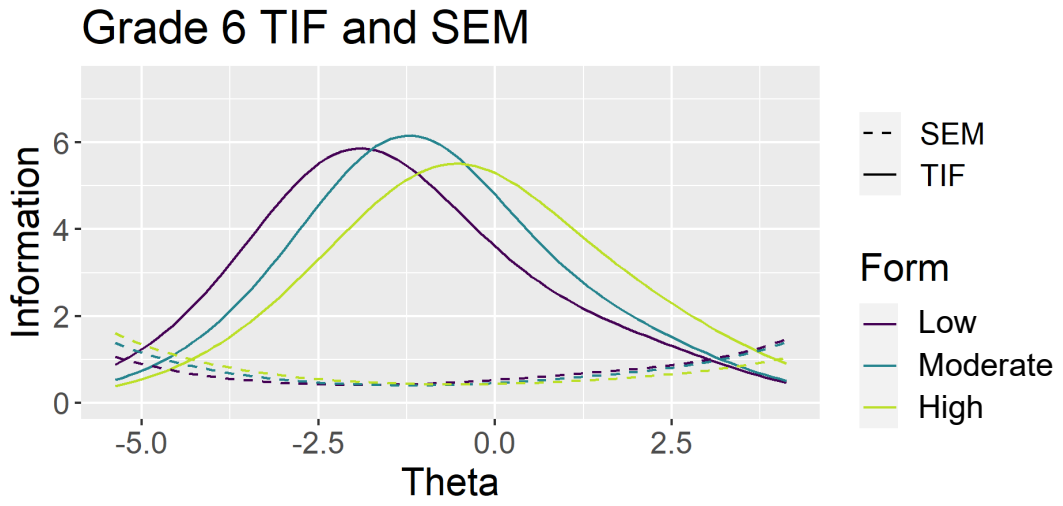


Figure 36. Frequency Distribution, Grade 7

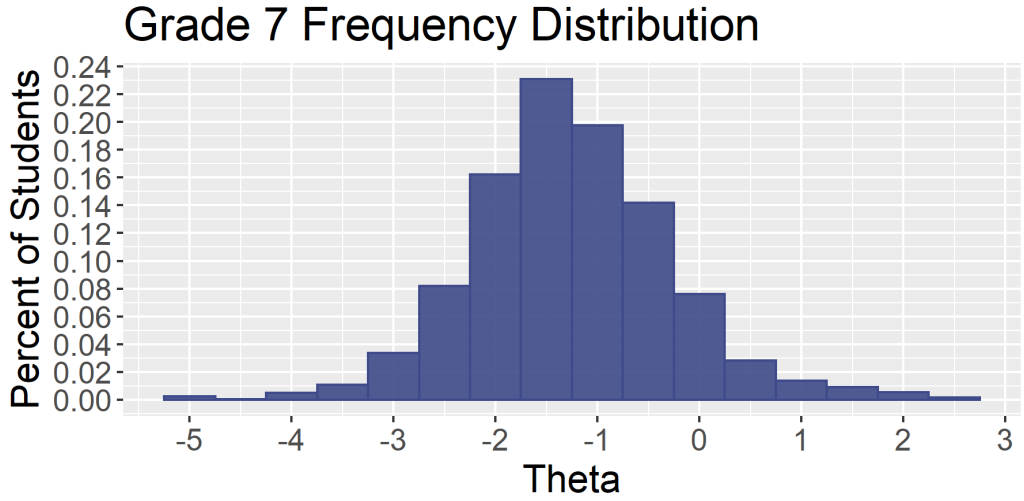


Figure 37. Diagnostic Test Characteristic Curves, Grade 7

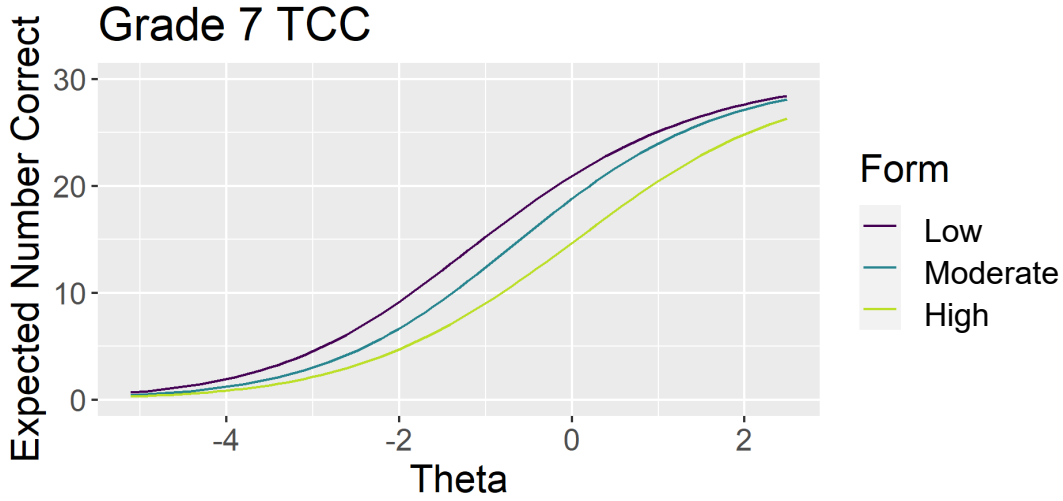


Figure 38. Diagnostic Test Information Function Curves and SEM, Grade 7

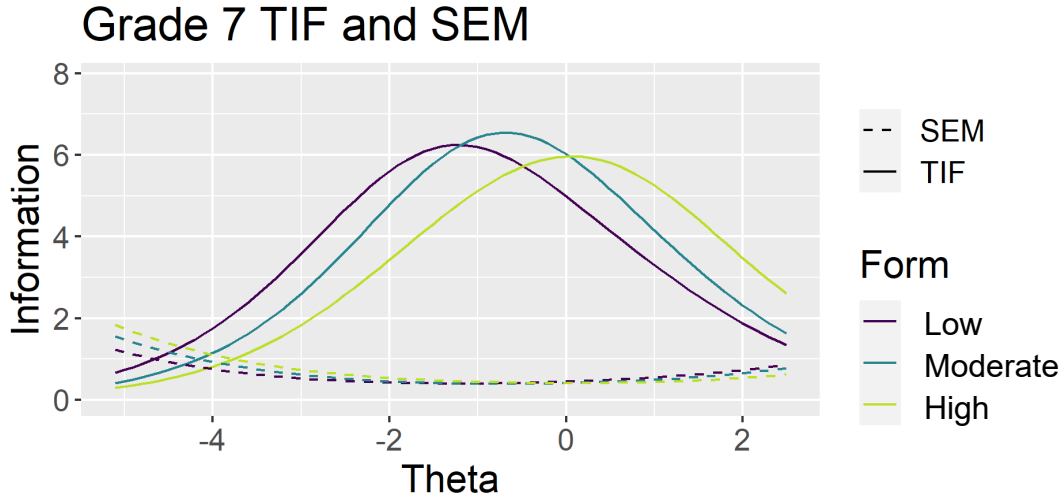


Figure 39. Frequency Distribution, Grade 8

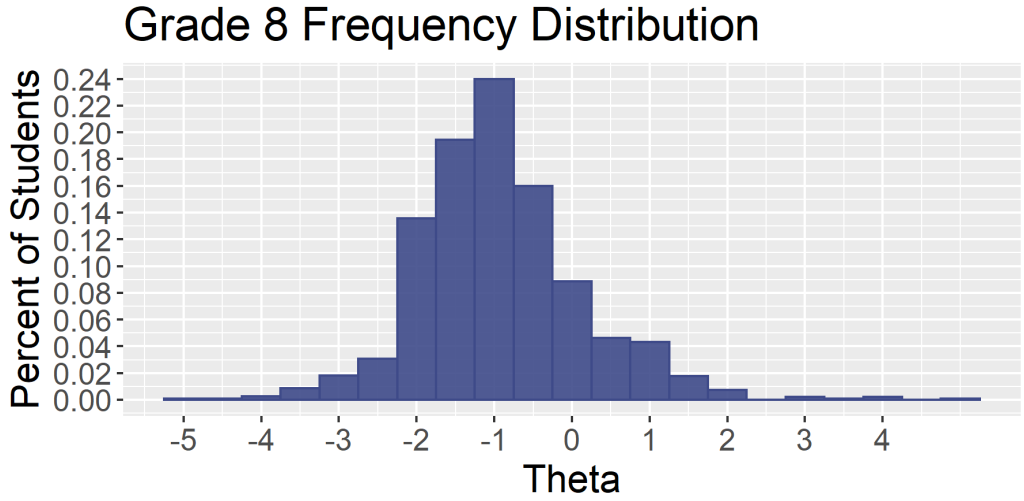


Figure 40. Diagnostic Test Characteristic Curves, Grade 8

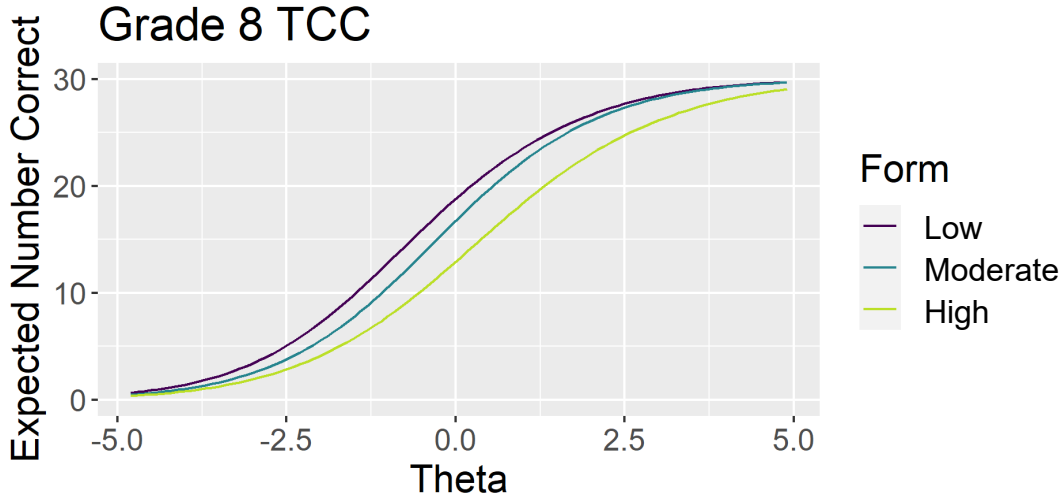
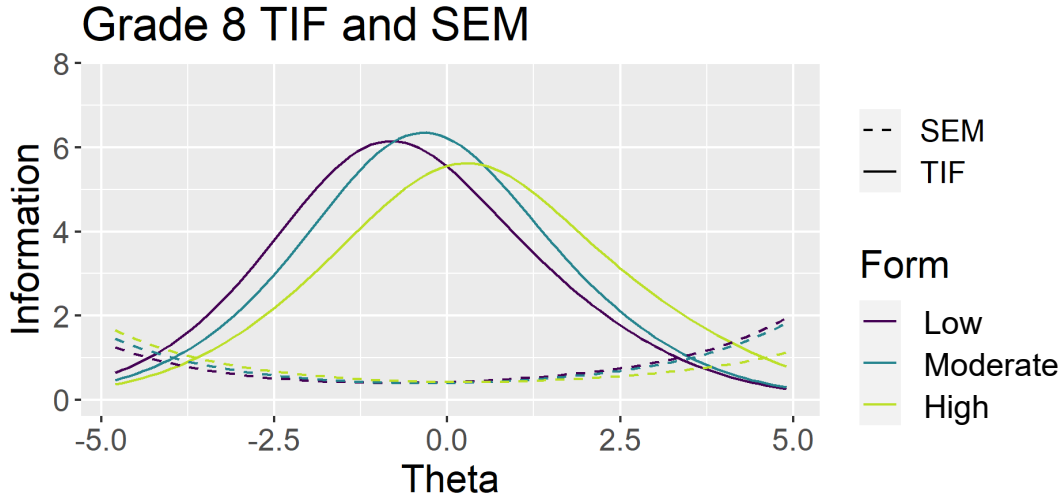
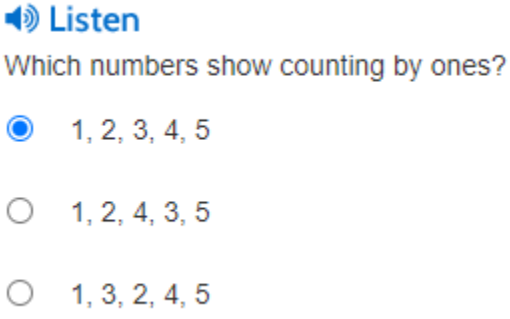


Figure 41. Diagnostic Test Information Function Curves and SEM, Grade 8



Appendix D: Example Item Template

Item Number	SMD_G0_101
Grouped (x of y) or Standalone Item	STANDALONE
Item Type	MULTIPLE CHOICE (Single Response)
Standard	K.CC.A.1
Difficulty	L
DOK	1
Calculator	NO
Art Instruction	
Source	
Art Spec Filename	
Alt Tag	
Stem	Which numbers show counting by ones?
Answer Options	
Option 1	1,2,3,4,5
Option 2	1,2,4,3,5
Option 3	1,3,2,4,5
Option 4	n/a
Correct Answer	Option 1
Rationales	
Option 1	correct
Option 2	reversed 3 and 4 in the counting sequence
Option 3	reversed 2 and 3 in the counting sequence
Option 4	n/a
Score Point	1



Appendix E: Item-PLD Alignment Subject Matter Expert Training Slides

Savvas Math K-8 Diagnostic Assessments
Embedded Standard Setting

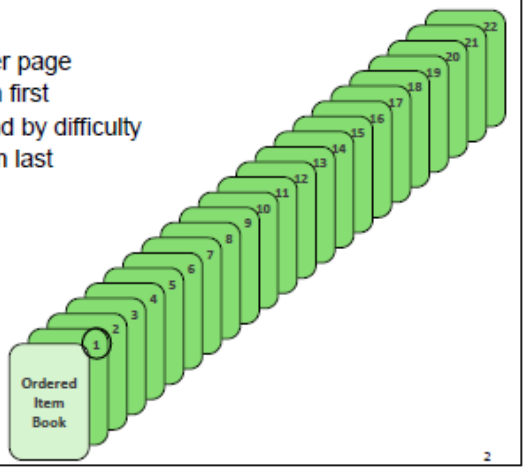
Subject Matter Expert
Item-Performance Level Alignment Training

April 5, 2021

1

Ordered Item Book

- One item per page
- Easiest item first
- Items ascend by difficulty
- Hardest item last

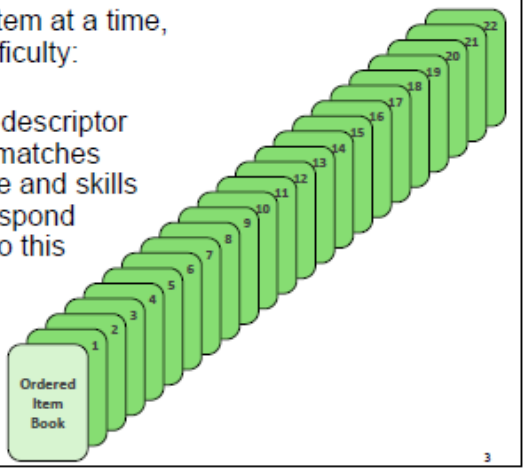


2

Alignment Criterion

Review one item at a time, in order of difficulty:

Which level's descriptor most closely matches the knowledge and skills required to respond successfully to this item?

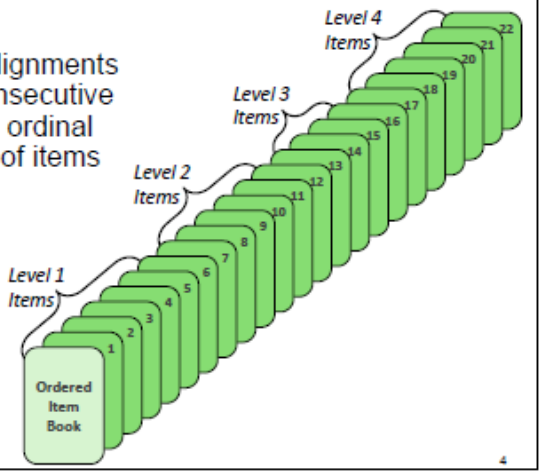


3

3

Coherent Alignments

Coherent alignments result in consecutive disjoint and ordinal sequences of items

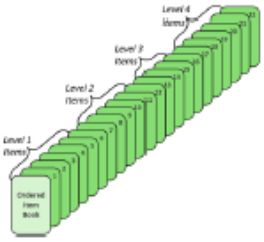


4

4

Expected Ordering

Coherent alignments result in consecutive disjoint and ordinal sequences of items



Order of Difficulty (out of 59 items)	Standard	Aligned Level (1, 2, 3, or 4)	Comments
1		1	
2		1	
3		1	
4		1	
5		1	
6		1	
7		2	
8		2	
9		2	
10		2	
11		2	
12		2	
13		3	
14		3	
15		3	
16		3	
17		4	
18		4	
19		4	
20		4	
21		4	
22		4	

5

Reality:
Operational ordering may have inconsistencies

Operationally, alignments may result in nonconsecutive sequences.

We seek to minimize this to a modest number of inconsistencies.

Order of Difficulty (out of 59 items)	Standard	Aligned Level (1, 2, 3, or 4)	Comments
1		1	
2		1	
3		1	
4		1	
5		2	
6		1	
7		2	
8		2	
9		2	
10		2	
11		2	
12		2	
13		3	
14		3	
15		3	
16		3	
17		4	
18		3	
19		4	
20		4	
21		4	
22		4	

6

Resolving Dual Alignments

Order of Difficulty (out of 59 items)	Standard	Aligned Level (1, 2, 3, or 4)
1		1
2		1
3		1
4		1
5		1
6		1 & 2
7		2
8		2
9		2
10		2
11		2
12		2
13		3
14		3
15		3
16		2 & 3
17		4
18		4
19		4
20		4
21		4
22		4

Order of Difficulty (out of 59 items)	Standard	Aligned Level (1, 2, 3, or 4)
1		1
2		1
3		1
4		1
5		1
[illegible]		1
7		2
8		2
9		2
10		2
11		2
12		2
13		3
14		3
15		3
[illegible]		3
17		4
18		4
19		4
20		4
21		4
22		4

7

4