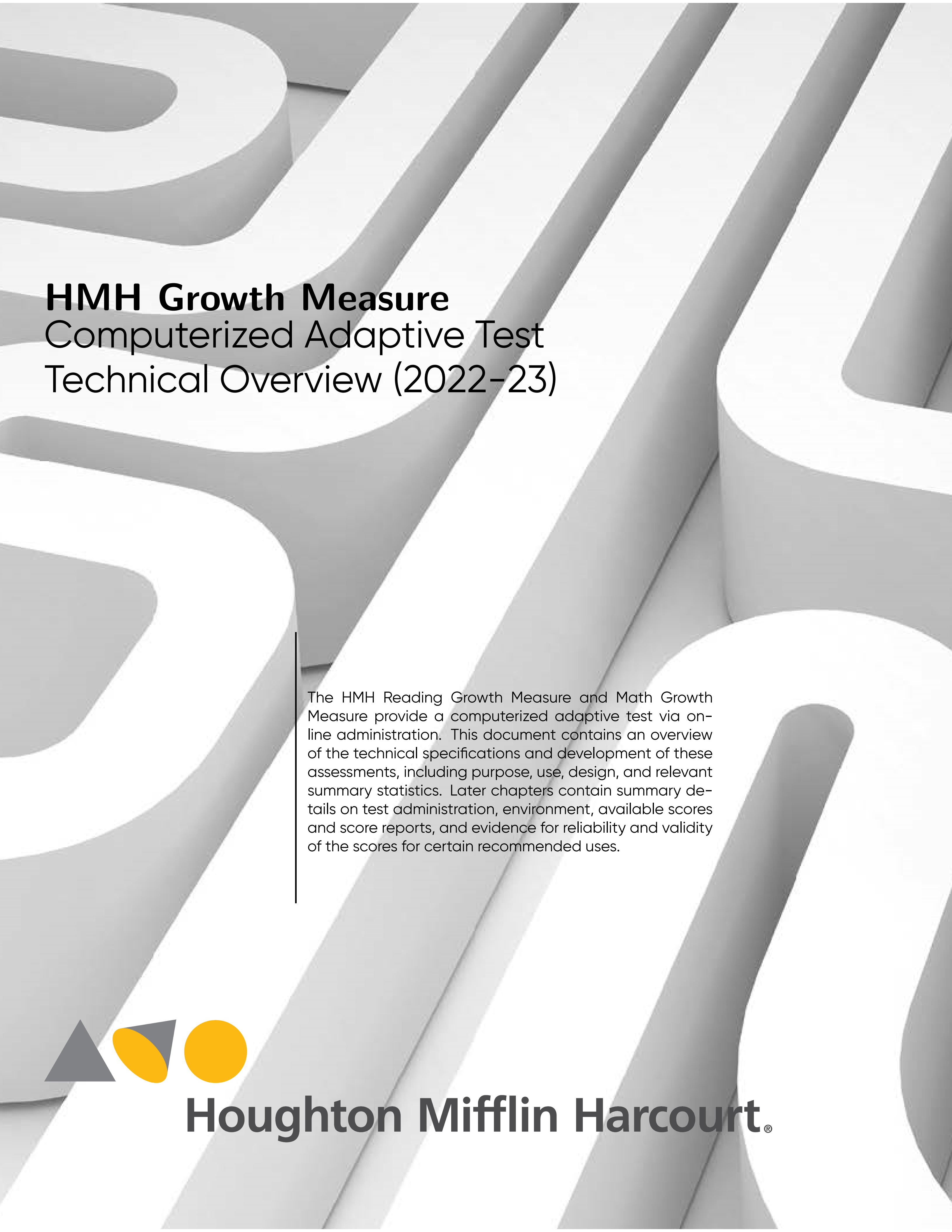


HMH Growth Measure
Computerized Adaptive Test Technical Overview (2022-23)

The HMH Reading Growth Measure and Math Growth Measure provide a computerized adaptive test via online administration. This document contains an overview of the technical specifications and development of these assessments, including purpose, use, design, and relevant summary statistics. Later chapters contain summary details on test administration, environment, available scores and score reports, and evidence for reliability and validity of the scores for certain recommended uses.

Houghton Mifflin Harcourt®

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. OVERVIEW AND PURPOSE

1.1. Overview

The Houghton Mifflin Harcourt (HMH) Growth Measure (GM) computerized adaptive tests for English/Language Arts (ELA) and Mathematics (Math) are standards-based interim assessment tools developed for Kindergarten through Grade 11. Specifically, Math assessments cover Kindergarten through Grade 8 and course-based assessments for High School (i.e., Algebra 1, Geometry, and Algebra 2). ELA assessments cover Grades 2 through 11. The GM assessments provide computerized adaptive online administration for ongoing information about teaching and learning in relation to a wide range of instructional standards (including the Common Core State Standards), and a range of grade-level expectations via a series of grade-specific tests. The GM is typically administered three times per year (e.g., fall, winter, and spring) and includes valid, reliable items that provide deeper student connections to content. In addition, the GM can be aligned to summative state assessments in ELA and Math to provide predictive information for student achievement.

1.2. Purpose and Use

The GM serves as a decision-making tool at the classroom, school, and district levels. This solution is achieved by providing an instructionally supportive student-assessment system. The implementation model is purposefully designed to improve educational outcomes for students by providing school leaders a tool that delivers data to drive instructional next steps. These instructional shifts can be further enhanced through professional development. The professional development should include an overview of test results and data analysis following each GM administration.

The purpose of the GM for ELA and Math is to provide ongoing information for practitioners to

- evaluate student learning in relation to grade-level expectations
- identify specific student strengths and opportunities for growth as they relate to the content in the assessed standards
- identify student instructional needs to inform placement in intervention programs to better enable student proficiency
- explain eligibility for intervention programs
- provide student feedback for reflection to enable motivation and deepen learning
- evaluate individual student growth over time

Classrooms, schools, and districts achieve positive outcomes when teachers are engaged in understanding the purpose, test design, and standards assessed and when teachers and district leaders analyze assessment results at the school, grade, subject, and classroom levels to inform improvements in curriculum, instruction, and assessments.

Connected Learning Model

To help schools overcome potential impact due to interruption of schooling caused by COVID-19 and the subsequent return to in-person schooling, HMH has proposed that a Connected Learning Model be implemented. This model relies, in part, on assessment data to drive instructional decision making and accelerate learning. Assessment is the bridge between teaching and learning that provides insight into the impact of instruction and guidance on determining instructional next steps. The GM provides teachers with an enhanced set of research-based and adaptive benchmark assessments for Math and ELA. Teachers can use GM results to plan a connected approach to instruction that integrates effective strategies based on the science of learning. These strategies are already included in HMH products such as *Into Math*, *Into Reading*, *Into Literature*, and *Waggle*.

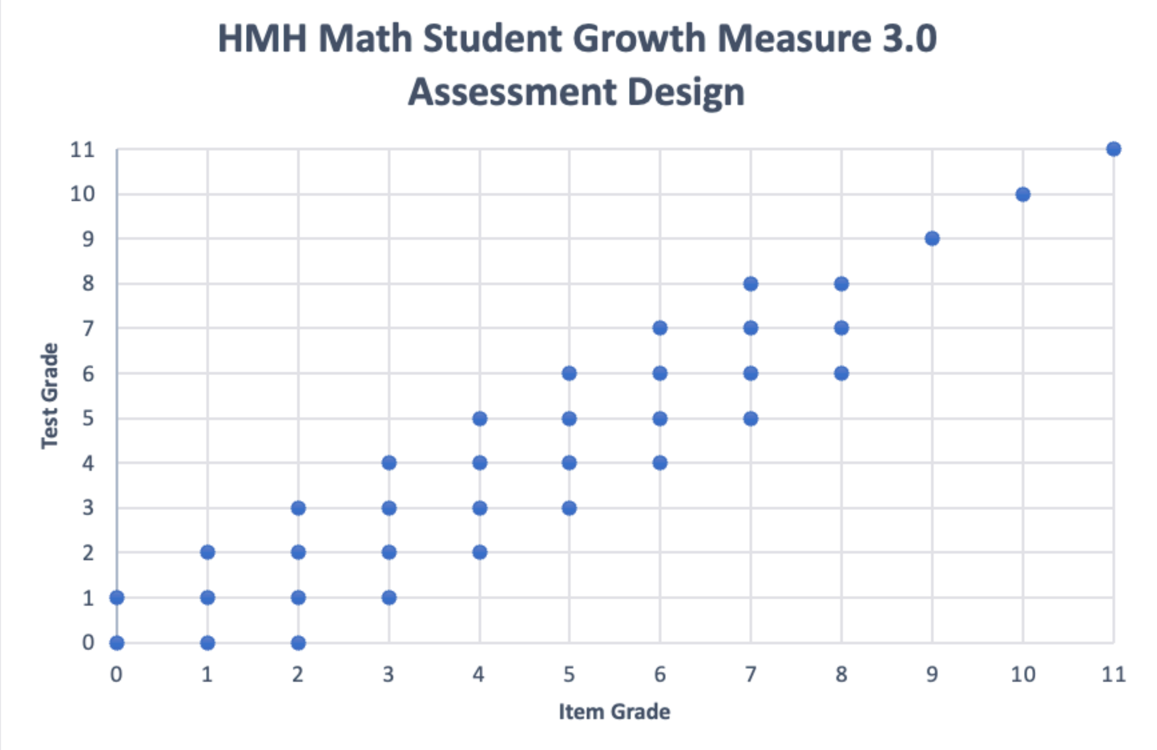
To make up for lost schooling, and to adjust learning experiences for re-entry to in-person schooling, teachers will strive to accelerate and deepen learning across a range of student abilities. To do this, they will want to focus on the most impactful evidence-based strategies for instruction and practice, including approaches found to have the largest impact on student achievement. *HMH Into Programs* provide the high-quality curriculum needed to support this acceleration of learning for all students, and the included GM is the first step in understanding what students' learning needs may be.

When given at the beginning of the year, the GM will help teachers understand where it will be most effective to focus their differentiation strategies. When given in the middle of the year, the GM will provide insights on what strategies are working to achieve their desired growth outcomes and where students may need continued or additional support or intervention. The end-of-year assessment measures the overall growth a student experienced and helps inform the learning plans they will need in the following year to maintain or accelerate growth.

1.3. Adaptive Nature of the GM

The HMH GM is a system that administers an interim-based summative computerized adaptive test using a modified version of on-the-fly multistage methodology. For example, the GM Math tests target a student's ability to learn mathematical concepts and measures the student's proficiency level relative to a grade-level span of content standards. The HMH GM is based on a Test Grade design where the item difficulties for each test grade are calibrated using on-grade examinees' student-level response data. The Test Grade is defined as a grade span of two to four item grades. For example, the Test Grade 5 comprises Grade 3 and Grade 4 items (two grades below), Grade 5 items (on grade), and Grade 6 items (one above grade). The GM also includes audio for the Math Grade K–5 assessments. Figure 1.1 illustrates the GM Math assessment design for Grades K–11 with respect to the span of item grades.

FIGURE 1.1: Item Grades Covered with Each GM Math Test Grade



Both student ability and item difficulty are measured on the logit scale. At the end of the test, an expected a priori (EAP) methodology is used to provide an estimated ability and transform this ability estimate to a reported scaled score. The psychometric rigor of the GM results from its tailored adaptive test design, which includes multiple layers of adaptability. Adaptive tests provide many advantages compared to traditional one-size-fits-all tests. For example, adaptive tests select items that match the student's ability level. In this way, students can avoid answering too many items that are too easy or too hard for them. Students' test experience is improved with less boredom or frustration and more stimulation by being constantly challenged at the right level. At the same time, test length, test time, and administration cost significantly reduce without sacrificing the accuracy of measurement. Another highly desirable feature of the HMH GM assessments is its adaptive blueprint design, which allows students to be tested beyond their grade level. This feature not only makes the assessment more relevant, appealing, and informative for a broader range of students, especially the struggling and the top achievers, but also reduces the floor and ceiling effects of the measurement instrument and enhances the measurement accuracy over a broad spectrum of the measured ability level. The GM also employs a set of Finetuner stages later in the test sequence in conjunction with a set of Locator stages that are in the early test sequence. The Locator stages are designed to determine whether test performance is On Level, Below Level, or Above Level with respect to grade-level standards. Once this location is estimated, the Finetuner stages select items that are aligned with the student's location on the grade-level equivalency (GLE) score scale. Overall, the adaptive stage-level nature of the GM algorithm, the adaptive blueprint set with the Locator stages, and the highly targeted nature of the Finetuner stages that administer items to the reported GLE score serve as these layers of adaptability.

1.4. Technical Overview

This technical overview provides basic information about the technical characteristics, development, and operation of the GM. The validity of intended uses of the scores and reliability of the assessments are reported explicitly in this document. While score reliability is relatively straightforward, the steps in creating the program and putting it into operation are all components of validity, which is also discussed. The validity of score use and interpretation for any assessment stems from the statement of the test's purpose and the intended use of the scores, the steps taken in designing the test, the processes of developing the content of the test, the process of consulting with stakeholders, the process of communicating about the test to users, the processes of scoring and reporting, and the process of data analysis. The careful documentation of each of these steps is a necessary piece of a comprehensive, defensible validity argument for the intended uses of the assessment scores. In reading this technical overview, it is critical to remember that the testing program does not exist in a vacuum; it is not just a test. It is one part of a complex network intended to help schools focus their energies on improvement in student learning. The GM should be part of an integrated program of testing, curricular, and instructional support. It can be evaluated properly only within its full context.

2. DEVELOPMENT PROCESS

This chapter provides an overview of the development of the GM computerized adaptive tests for ELA and Math, including (1) item and passage development, (2) test specifications, (3) item review, and (4) field testing of new items. According to the Standards for Educational and Psychological Testing (AERA et al., 2014), "important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure" (p. 14). The descriptions of the test-development procedures included in this chapter provide evidence that supports both the content and construct validity of the assessments.

2.1. Item and Passage Development

Item and Passage Writing

The individuals who created the test passages and items for the GM were members of the HMH Assessment Solutions and Design (ASD) team and item/passage writers with a minimum of a bachelor's degree in their area of expertise along with teaching experience. All item/passage writers were trained before writing began and received information on the following topics:

- Common Core State Standards (CCSS) and HMH Learning Spine
- vertical alignment of curriculum considerations
- depth of knowledge (DOK)
- passage-based item sets that give proper consideration to quantitative measures, qualitative measures, and reader and task considerations
- item-writing best practices
- passage and item bias, fairness, and sensitivity
- specific examples of appropriate and inappropriate items and passages
- strict adherence to the principles of universal design
- appropriate tiered vocabulary

Universal Design

HMH ASD team members are trained to write items that adhere to the principles of universal design, making the items accessible to the widest range of students. For example, items and passages were written using clear and concise language, and all art, graphs, and tables were labeled and not crowded with extraneous information. According to the National Center for Educational Outcomes Synthesis Report (Thompson et al., 2002), universally designed assessments have seven elements:

- inclusive assessment population
- precisely defined constructs
- accessible, unbiased items
- amenability to accommodations
- simple, clear, and intuitive instructions and procedures
- maximum readability and comprehensibility
- maximum legibility

All items for the GM were developed with these elements in mind.

Item and Passage Specifications

The item and passage specifications were developed by the HMH ASD team prior to developing passage and item sets for field testing. Once reviewed, revised, and approved, the item and passage specifications were implemented by writers and reviewers. Numerous documents make up the item and passage specifications, including the following:

- editorial and art style guides
- item writer–training presentation for selected-response items and technology-enhanced items (TEIs)
- item writer specifications for selected-response items and TEIs
- passage writer–training presentation
- CCSS passage and item specifications
- item writer guide to DOK
- item writer guide to the content management system
- item development checklist

The editorial style guide is a tool used primarily by copyeditors and item writers to ensure that the spelling, grammar, and formatting of test items are consistent. This extensive document utilizes *The Chicago Manual of Style* for most of its guidelines and specifies where the items may deviate.

The art style guide is used by both graphic artists and item writers to ensure that any graphics developed for the test items meet certain dimensions and other formatting requirements. These guidelines are important for both consistency and universal design.

TABLE 2.1: Reporting Domains for Math and ELA GM

							Test Grade						
	Domain	*K*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Mathematics	(Total)	X	X	X	X	X	X	X	X	X	X	X	X
	Fractions, Ratios, and Proportions				X	X	X	X	X				
	Operations and Algebra	X	X	X	X	X	X	X	X	X	X		X
	Geometry	X	X	X	X	X	X	X	X	X		X	
	Measurement, Data, and Probability	X	X	X	X	X	X	X	X	X	X	X	X
	Numbers	X	X	X	X	X	X	X	X	X	X		X
	Functions									X	X		X
English Language Arts	(Total)			X	X	X	X	X	X	X	X	X	X
	Reading Comprehension			X	X	X	X	X	X	X	X	X	X
	Language			X	X	X	X	X	X	X	X	X	X

2.2. GM Test Specifications

Standard 1.11 of the *Standards for Educational and Psychological Tests* specifically addresses the appropriateness of test content and its relationship to a solid validity argument. In addition, Standard 4.12 defines "test specifications" and provides examples of the type of information that should be included in a specification document. The test specifications describe the content and format of the test and delineate the ideal number of items and points assessed for each competency. This section overviews the development and use of the test-specification documents for the development of GM items.

The HMH ASD team developed draft test specifications for each grade. The specifications were finalized before the development of test items. The test-specification document served as the foundation for all item development. The material in the test specifications was designed for use by the HMH ASD team to construct items pools containing items that

- are aligned to the CCSS and HMH Learning Spine
- are aligned to Norman Webb's DOK cognitive levels
- are selected response and technology enhanced
- are standalone or passage based

Also, all item pools were designed to have

- balanced gender and ethnicity representation
- an appropriate key distribution
- graphic quantity and variety appropriate for the grade level
- a variety of contexts that are fair and that provide unbiased opportunities for students to demonstrate their understanding

Field test items are intentionally developed to target item-bank deficiencies with the intent of having a robust pool of items for the adaptive benchmark engine. In addition, checklists and quality-control procedures accompany development of the operational item pools.

2.3. Item-Review Process

After an item is written, an HMH ASD member reviews the item to ensure that the item

- is aligned to the standard, including DOK
- is grade appropriate
- meets universal design criteria
- does not contain bias and sensitivity issues
- meets the best practices standards described in the item-writing training

If the item meets the requirements or if the specialist can make slight modifications so that the item meets the requirements, it will be moved to the next step.

A senior test development specialist then reviews the item and confirms that all requirements have been met. If the item meets the requirements, it will be moved to the next step; otherwise, it will be returned to ASD for review and revision.

Following approval of ASD staff, the item is reviewed by editorial staff to check spelling, grammar, and adherence to the project style guide. Items that make it through this review process then undergo content and bias review.

The content and bias review process occurs once enough data on the item is collected through field testing. During field testing, the CAT engine gives up to five new items to each student's assessment. Since these items have not yet been validated, the student's score will not include responses based on those items. Field tested items undergo a review (descrived in the next section) for a variety of psychometric criteria such as validity, reliability, and bias statistics, so that they can potentially be included in the operational item pool.

2.4. Overview of Data Review and Flagging Criteria

The purpose of the data review is to evaluate statistics of flagged items in order to generate possible hypotheses explaining why these items behaved as they did on operational or field tests. Furthermore, an appropriate course of action (accept, reject, or revise and re-field tested) for the items is suggested. Note that item statistics were computed only when an item had at least 100 responses.

The GM items were flagged based on a variety of item statistics related to reliability, validity, and IRT model fitness, including

- n count: minimum number of responses received on each item
- SE: standard error (SE) of the estimate
- p value: an index that indicates the proportion of students in a specific group, such as students in a certain grade, who answered the test item correctly. P value should be used with caution in computerized adaptive testing (CAT), as items are administered to students based on ability estimates. Therefore, p value for each item in a CAT test would average out to be around 50%
- point measure: a correlation index between the calibrated Rasch measures and the responses. The index provides an indication of the item's discrimination of the low-ability group from the high-ability group (low values: low-ability students outperform high-ability students on item)
- discrimination parameter: indicates that high-ability students are selecting an incorrect distractor more often than they should be (for good measurement to happen). A good discrimination value is around 1.0 or above.
- item characteristic curve (ICC): these charts can be used for a deeper dive into how each answer option performed. When used in complement to a popular-key flag, in which a wrong key or multiple-key was found, ICC can provide insight based on student ability levels.

The data review is a collaborative effort led by the psychometric team and content specialist committee. The psychometric team analyzes item statistics and flags items based on their item statistic-flagging criteria. The content specialist committee reviews the flagged items, generates hypotheses explaining why the items were flagged statistically, and provides suggestions on items as follows: "accept item," "reject item," "revise and re–field test item." The criteria used for flagging items are presented in Table 2.2.

During the data review, both operational and field–test items are evaluated. Items that fall below criteria may be included in the operational item pools, subject to further revision by the test development expert committee.

TABLE 2.2: Criteria Used to Flag Items

Statistic	Item-Flagging Criterion	Indicates
P value	P value of keyed response < P value of distractor	Problematic answer key
P value	P value of distractor > 0.30	Possible multiple correct options
P value	above-grade item with P value < 0.2	Too difficult above-grade item
P value	below-grade item with P value > 0.9	Too easy below-grade item
point measure	< 0	Problematic answer key
point measure	< 0.15	Poor differentiation among students' ability levels
item discrimination	< 0.30	A distractor that is popular among high-ability students
item fit	< 0.7 or > 1.3	Possible issue with item model fit
total number of responses	< 100	Unreliable item statistics
standard error	> 0.3	Unreliable item statistics

Table 2.3 shows the number of unique operational items in each grade and content area as of June 2022. In total, there are 5,957 unique operational items across all ELA grades; for Math grades, that number is 6,049.

TABLE 2.3: Number of Unique Operational Items

Item Grade	ELA	Math
K	105	486
1	315	523
2	877	578
3	820	518
4	690	518
5	611	519
6	531	488
7	542	483
8	550	470
9	329	506
10	314	466
11	273	494

2.5. Data Review Results

To ensure appropriate maintenance and continuous improvement of the GM assessment items, the expert committee reviews both operational and field–tested items. Items that have been flagged for content reasons are rejected and subsequently removed from the item bank, making them invalid for future assessments. Problematic items may also be edited and reused, in which case they are put on hold for revision. Once an item is revised, it must be field tested again before returning to the operational item pool. That is, the item goes through the field testing process for administration and is revised in later data reviews.

Tables 2.5 and 2.4 summarize the Math and ELA data review results, respectively, for the beginning of the year of 2022.

To further investigate potential issues with item distractors or answer keys, an ICC chart is produced along with the item statistics for the data review. The ICC charts show how each option of a multiple-choice item performed. ICC charts are recommended for use with other statistics, such as point measure and discrimination parameter. Where a wrong-key or multiple-key situation was suggested, ICC can provide further insight based on student ability levels. Figure 2.1 is an ICC chart of a high-quality item, in which the correct response (option B) is more likely to be selected as the ability levels increase and the incorrect options are less likely to be selected as the ability levels increase.

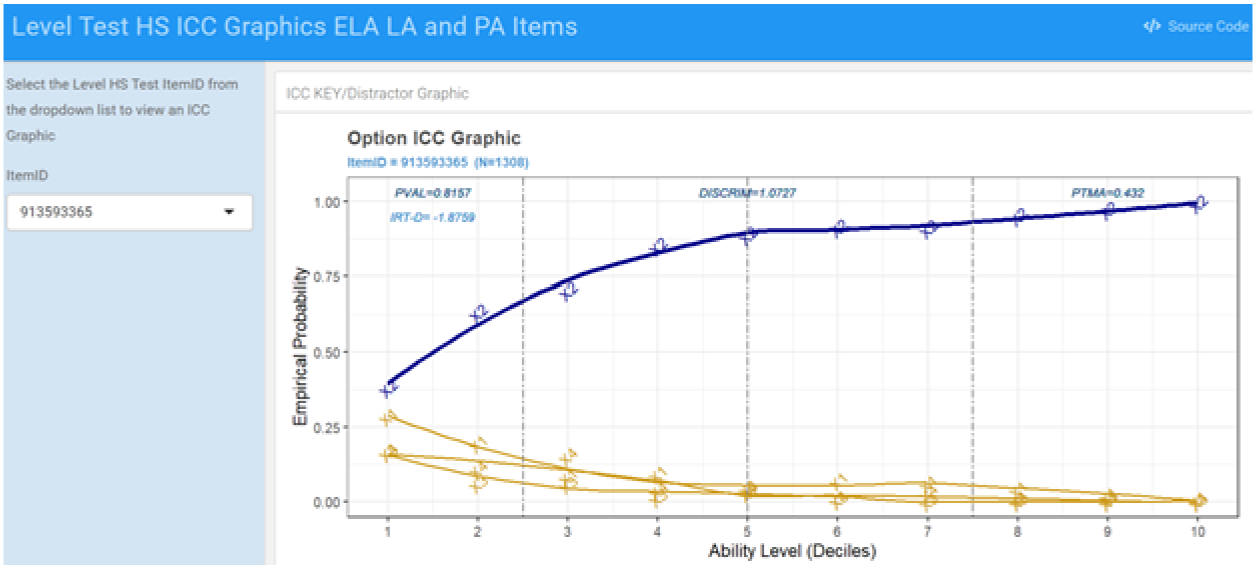
TABLE 2.4: Math Winter 2022 Data Review Results

Item Grade	Number of Items Reviewed	Number Accepted	Number Rejected	Number Edited	Proportion Accepted	Proportion Rejected	Proportion Edited
9	53	40	6	7	0.75	0.11	0.13
10	72	54	9	9	0.75	0.12	0.12
11	77	61	5	11	0.79	0.06	0.14

TABLE 2.5: ELA Winter 2022 Data Review Results

Item Grade	Number of Items Reviewed	Number Accepted	Number Edited	Number Rejected	Proportion Accepted	Proportion Rejected	Proportion Edited
K	36	28	0	8	0.78	0.22	0.00
1	63	55	1	7	0.87	0.11	0.02
2	12	9	0	3	0.75	0.25	0.00
3	14	11	2	1	0.79	0.07	0.14
4	14	13	0	1	0.93	0.07	0.00
5	4	4	0	0	1.00	0.00	0.00
6	7	6	0	1	0.86	0.14	0.00
7	8	7	1	0	0.88	0.00	0.12
8	3	2	1	0	0.67	0.00	0.33
9	6	1	1	4	0.17	0.67	0.17
10	4	2	1	1	0.50	0.25	0.25
11	2	2	0	0	1.00	0.00	0.00

FIGURE 2.1: ICC Chart of a High-Quality Item



Level Test HS ICC Graphics ELA LA and PA Items

Select the Level HS Test ItemID from the dropdown list to view an ICC Graphic

ItemID

913593365

ICC KEY/Distractor Graphic

Option ICC Graphic

ItemID = 913593365 (N=1308)

2.6. Summary

The HMH GM provides an indication of student progress toward achieving the knowledge and skills identified in the CCSS and the HMH Learning Spine. Just as the content framework guided the item and passage development and selection processes, content considerations played an equally important role in the development of items. A variety of classical item statistics and IRT fit statistics were used for item review. As items were selected for inclusion in the item pools, every effort was made to balance the content coverage and the overall difficulty of items.

3. SCALING AND EQUATING

3.1. Introduction

This chapter provides an overview of the scaling and equating procedures HMH implemented for the GM for ELA and Math. These analyses were performed to place all items to their corresponding GM Test Grade scales. Note that item response theory (IRT) was used for scoring the GM for reporting. First, the IRT model used for both scaling and equating is described below. This is followed by a description of the analyses used to evaluate the degree to which the assumptions of the Rasch model were met as well as the degree to which the test data fitted the model. Finally, a description of the cross-grade scaling procedures used to create the GM item pools for each Test Grade is presented.

3.2. Item Response Theory

Winsteps software (Linacre & Wright, 2000) and the TAM R package (Robitzsch et al., 2021) were used to accomplish the scaling and equating for the GM. Winsteps is designed to produce a single scale by jointly analyzing data from students' responses to selected-response items. Multiple-choice items were calibrated using the Rasch model (Hambleton & Swaminathan, 1985; Rasch, 1960; Wright & Stone, 1979). One feature of the Rasch model that distinguishes it from classical test theory (CTT) is the placement of estimates of a person's ability and item difficulty on the same scale. The Rasch model expresses the probability of a correct response to an item as a function of the ability of the person and the difficulty of the item.

Because the Rasch model is the basis of all scaling analyses associated with the GM, the utility of the results from the field test administrations depends on the degree to which the assumptions of the model are met as well as the degree to which the test data fit the model. The assumptions of the Rasch model include that (1) the data are unidimensional and (2) the data have the quality of local independence, or responses to one item do not depend on responses to another item. The following sections address these assumptions and include evaluations of the dimensionality and local independence of the data, as well as fit indices.

3.3. Assessing Unidimensionality and Local Independence of the Data

HMH researchers completed a principal components analysis (PCA) to assess the unidimensionality assumption of the Rasch model (Divgi, 1980). Essentially, the Rasch dimension was removed first, and the residual variance (the proportion of the variation in the data set that is unaccounted for by the Rasch model) was then analyzed. Consequently, for this model to hold, this analysis should not identify a second dimension that accounts for a practically significant amount of residual variance. Analyses for the GM for ELA and Math for each Test Grade generally showed the secondary dimension to represent an impact of less than 5% of the total variance, and the secondary dimension was, therefore, considered of little practical import.

Based on the same PCA as noted above, standardized residual correlations were produced to assess the local independence assumption of the Rasch model. The purpose of these analyses was to detect dependency between pairs of items. Results of these analyses generally supported the assumption of local independence; values for standardized residual correlations were generally low, indicating little dependency between pairs of items.

3.4. Assessing Data Fit to the Model

Two statistics were used to evaluate how well the data fitted the Rasch model: infit and outfit. High infit statistics indicate unexpected responses to items that are well targeted to the test taker's ability. For example, a test taker incorrectly answers a number of items that are well suited to their ability. Outfit is sensitive to outliers—in other words, to aberrant responses for items with difficulty far from a test taker's ability. For example, a test taker incorrectly answers items that should be easy, or correctly answers questions that should, in light of their performance on other test items, be difficult. High outfit values may indicate lucky guessing or careless mistakes. Relatively speaking, extremely high infit values are believed to be a greater threat to the measurement process than extreme outfit values.

Tables 3.1 and 3.2 provide item summary statistics for the overall Rasch analysis for ELA and Math, respectively, for Test Grade 5. These summary analyses include results prior to the data review process that flagged problematic items. Such items were then reviewed for their potential exclusion from the operational item pools. The tables include summary fit statistics for the GM and are typical of the results found for the other Test Grades.

TABLE 3.1: Rasch Summary Statistics for ELA Test Grade 5

Statistic	**Rasch Measure**	***P*-Value**	**INFIT**	**OUTFIT**	**PTBS***
# of Items	1,602	1,602	1,602	1,602	1,602
Mean	0	0.48	1	1.01	0.42
SD	0.81	0.15	0.12	0.2	0.03
Minimum	-2.21	0.07	0.73	0.5	0.27
10th Percentile	-1.09	0.31	0.86	0.79	0.39
25th Percentile	-0.51	0.37	0.9	0.86	0.41
50th Percentile	0.08	0.47	0.99	0.99	0.43
75th Percentile	0.55	0.58	1.07	1.11	0.44
90th Percentile	0.9	0.69	1.16	1.27	0.45
Maximum	3.36	0.86	1.38	2.09	0.49

TABLE 3.2: Rasch Summary Statistics for Math Test Grade 5

Statistic	**Rasch Measure**	***P*-Value**	**INFIT**	**OUTFIT**	**PTMA***
# of Items	1,585	1,585	1,585	1,585	1,585
Mean	0.285	0.44	1	1.05	0.347
SD	1.12	0.125	0.11	0.25	0.107
Minimum	-2.7	0.09	0.76	0.6	-0.13
10th Percentile	-1.14	0.29	0.9	0.87	0.2
25th Percentile	-0.47	0.36	0.93	0.92	0.28
50th Percentile	0.24	0.44	0.98	0.98	0.35
75th Percentile	0.99	0.5	1.05	1.1	0.42
90th Percentile	1.78	0.59	1.15	1.35	0.47
Maximum	4.3	0.91	1.49	3.3	0.65

3.5. Scaling Design

To permit the direct comparison of examinee scores within and across years on a Test Grade, the student ability estimates from the adaptive test are transformed mathematically to a more convenient metric for reporting purposes. For the HMH GM 4.0, the scaled metric ranges from 1 to 99 and is prefixed by the Test Grade. For example, the scaled scores on Test Grade 4 have a range from 401 to 499, and the Test Grade 8 scaled scores have a range from 801 to 899. Scores of students from Test Grade 2 to High School are divided into five performance levels, including *Far Below Level*, *Below Level*, *Approaching*, *On Level*, and *Above Level*. The scaled scores for *Far Below Level* are set to be between G01 and G20; *Below Level* are between G21 and G45; *Approaching* are between G46 and G60; *On Level* are between G61 and G80; *Above Level* are between G81 and G99, where *G* is the Test Grade prefix. At Test Grade 3, for example, students scoring below 320 are classified as performing *Far Below Level*, students with scaled scores from 361 to 380 are performing *On Level*, and a scaled score between 381 and 399 indicates *Above Level* performance. Scaled scores of students from Test Grade K will not be placed into *Far Below Level* or *Below Level*. Test Grade K performance levels are *Approaching*, *On Level*, and *Above Level*. Therefore, kindergartners with G01 to G60 are placed in the *Approaching* performance level. Similarly, first grade students will be placed into *Below Level*, *Approaching*, *On Level*, and *Above Level*, but not *Far Below Level*. Students from the first grade whose scaled scores range from G01 to G45 are placed in the *Below Level* performance category.

The general transformation formula used to obtain scaled scores for the GM Math/ELA is the following:

$$\text{Scale Score} = (\hat{\theta} - \theta_{std_4}) * Slope + Anchor + Grade * 100$$

where $\hat{\theta}$ is the student ability estimate, θ_{std_4} is the ability cut score that separates the *On Level* performance standard and the *Above Level* performance standard, *Anchor* is set to be 80 as the intercept for the scale (the scale cut score for the *On Level* performance standard), *Grade* is the grade of the administered test (Test Grade, and 0 is used for Test Grade K), and *Slope* is a (potentially) unique numerical constant that sets the spread of the scale.

For GM Math/ELA, the transformation formula uses two cut scores to construct each of the scaled scores (see Chapter 5). For Math and ELA, a set of approved psychometric targets were established through standard setting and literature review. The performance band cut scores were also built upon the GM scaled scores as they are intimately connected. Therefore, a meaningful update to the items per Test Grade would involve an updated scaling of the scaled scores. This could involve a future addition of technology-enhanced items (TEIs) and/or recalibration of item pools with operation data.

One goal for the scale transformation is to make the cut scaled scores for each performance level as consistent as possible across Test Grades. Using a linear transformation-like equation allows for two cut scaled scores to be fixed. For HMH GM 4.0, Test Grade K includes three performance levels. The cut scaled scores for the *Approaching* performance level and for the *On Level* performance level are set to G60 and G80, respectively, where *G* is the Test Grade prefix. Therefore, the *On Level* and *Approaching* performance levels show a difference of 20 scaled score points. For Test Grades 1 to 11, the cut scaled scores are G45 for the *Below Level* performance level, and G80 for the *On Level* performance category, which represents a difference of 35 scaled score points. The *spread* constant for each Test Grade per subject anchors both cut scaled scores to equal G45 and G80 for Test Grades 1 and above and to equal G60 and G80 for Test Grade K.

The formula used to find the Slope for Test Grade K is as follows:

$$\text{Slope} = \frac{20}{\theta_{std_4} - \theta_{std_3}}$$

The formula used to find the Slope for Test Grades 1 and above is as follows:

$$\text{Slope} = \frac{35}{\theta_{std_4} - \theta_{std_2}}$$

where θ_{std_2}, θ_{std_3}, and θ_{std_4} are the cut expected a posteriori (EAP) scores (students' ability scores) for the *Below Level*, *Approaching*, and *On Level* performance standards, respectively. These cut EAP scores are also considered as the psychometric targets. These psychometric targets were established and approved through research and standard setting procedure.

The potential *lowest observable scale score* (LOSS) is set to be no lower than G01, and the *highest observable scale score* (HOSS) is set to be no greater than G99 for each GM scaled score. The effective LOSS/HOSS may be higher/lower than G01 and G99, depending on the variance associated with the scaled scores per Test Grade. For example, on Test Grade 4, the empirical LOSS is greater than or equal to 401 and the empirical HOSS is less than or equal to 499. The LOSS and HOSS prevent extreme student scaled scores from being transformed outside the desired range of the scaled scores, which would suggest that the scaled score(s) contain only noise. Because GM Math EAP score estimates are constrained to be typically within the range –3 to 3 (for item pools that center on 0), the scaled scores of G01 or G99 may not be easily attainable and should be rather rare. Hence, the effective LOSS/HOSS may be lesser or greater than the theoretical G01 and G99 score units.

3.6. Scaling Method

For each content area, a separate calibration has been conducted for each Test Grade from Test Grades Kindergarten through 11 for Math and from Test Grades 2 to 8 for ELA. These separate calibrations for each Test Grade serve as the foundation for the GM test design, in which each Test Grade has its own unique score scales. The High School GM ELA item pools for Test Grades 9 to 11 are calibrated onto the same scale via the concurrent calibration of student responses, since these Test Grades share the same item pool. The implication that follows from this IRT scaling approach is that students' ability estimates are based on item difficulties as calibrated to their corresponding Test Grades. For example, the estimated latent ability of a Test Grade 5 student is based on how students taking the Grade 5 GM test typically respond to each item within the item pool.

As a result of these separate calibrations for GM Math from Test Grades Kindergarten to 11 and GM ELA from Test Grades 2 to 8, identical items occurring in different Test Grade pools would show discrepancies in item characteristics. In current GM practical operation, newly developed items will be implemented in tests with Test Grades being matched to the item grades of these field test (FT) items. Subsequently, adjacent Test Grades' item pools will utilize these FT items as off-grade items, as long as these FT items present qualified item characteristics in the process of item review. Each Each FT item will be placed into item pools that have its same scaling constant, until the FT item has been calibrated as an off-grade operational item in its corresponding item pool.

3.7. Summary

All GM items are calibrated based on the unidimensional Rasch model. The GM scaled scores for each Test Grade range from G01 to G99, which are classified into five performance levels. The quality of an adaptive assessment is largely dependent on the quality and characteristics of the item pools. For the GM, the item pools are subsets of the total item bank. These subsets of items (a four item-grade span where applicable) are what construct the item pool for each Test Grade. The GM item pools have been extensively expanded and improved, and their development continues concurrent with their implementation.

4. TEST ADMINISTRATION

4.1. Introduction

Computerized adaptive tests are a powerful tool that, when calibrated correctly, provide greater accuracy of assessment across a wider range of abilities than traditional fixed-form assessments. Computerized adaptive tests allow embedded test-design properties into the delivery of the assessment, which adds flexibility to develop a variety of rules to meet desired characteristics of the test administration. They also reduce the number of items required to get an accurate assessment, because of improved targeting of items, which reduces the number of items administered during each test.

4.2. Testing Windows

The HMH GM is constructed to be optimally administered up to three times per year (e.g., fall, winter, and spring). By default, the GM test windows are automatically enabled and set for administrators. However, there is flexibility around the precise number of times that a student can take the adaptive assessments, and this can vary across students within a classroom. For example, in a given class of students, if a student joins the class later in the school year, this student could be tested in the second window if they missed the first event. Since students will be able to take the GM up to three times per year, once a student views an item, the system has been designed such that the student cannot view that item for the current school year. That way, in a typical scenario, a student won't be able to respond to the same item in adjacent testing windows.

4.3. Student Rostering

Customers will be able to provide HMH with the student data files (e.g., location, user, and student roster) for the purposes of rostering. As test assignments are created, HMH will list these tests on the Student Dashboard page and the Assignments page in the ED platform. Students added to a roster file through the HMH user-facing administration screen must inherit the test assignment for other students in their location and grade.

4.4. Navigation

The GM assessments are generally self-paced tests, do not have a proctor led component, and are available only as an online test administration. The lower-level tests (Kindergarten and Grade 1) have an audio administration, while all other grades are self-paced with no audio. Students follow onscreen prompts during testing and are allowed to go on to the next item without providing a response. Because of the adaptive nature of the assessments, once a student views an item (whether they provide a response or not), they cannot go back and review their response or go back and change their answer. Students are able to take breaks and have their assessment paused as long as the testing session is open.

Adaptive Inputs

An adaptive assessment requires several inputs to the engine, or algorithm.

Calibrated and Tagged Item Bank

An item bank is the full set of assessment items that have been calibrated and tagged appropriately. The items need an IRT difficulty parameter, such as the Rasch item-difficulty value, and must be tagged appropriately for grade, content or knowledge domains, or other test-design properties. Note that an item bank is different from an item pool. The item bank is the entire collection of items, while an item pool is the set of items made available to a specific Test Grade. All items in the item pools are in the item bank; however, the reverse is not true. As the student progresses through the test, items are selected for the next stage based on the student's performance on previous stage(s).

Activity Definition

The activity definition (i.e., test definition) is the mechanism that controls how the adaptive engine will dynamically choose content and the resulting testlets and items as a user progresses through a test.

Student Initial Ability Estimate

An individual student taking an assessment may have a previous ability estimate from a prior adaptive assessment. This can be input into the system to start the examinee at the most appropriate level in terms of the assigned adaptive blueprint (Standard or Intervention). If no prior estimate is available, the system defaults to the Standard adaptive blueprint. The GM does not currently use any prior ability or starting value to administer items. Rather, a Routing Test (Stage 1) is used and items within a defined range of difficulty are randomly selected to provide a provisional initial ability estimate to route students to the appropriate level in Stage 2.

Adaptive Engine

The adaptive engine takes the inputs as adaptive configuration files and delivers appropriate items to the student via an algorithm that computes a provisional ability estimate based on the student's prior responses to all prior administered items, taking into consideration those item difficulties. For each iteration (stage), the adaptive engine gets a better estimate of the student's ability and selects the next stage's set of items, using tag-based rules so that the assessment meets the desired test blueprint design. The examinee moves through the adaptive stages (six for Math and seven for ELA) and according to the prior stage-level ability estimate is routed to one of three levels within the **Locator** (first half of the test) and one of six levels within the **Finetuner** (second half of the test). The Locator levels comprise Below-Grade (BG), On-Grade (OG), and Above-Grade (AG). The Finetuner levels comprise Below-Grade 3 (BG-3), Below-Grade 2 (BG-2), Below-Grade 1 (BG-1), On-Grade 1 (OG-1), On-Grade 2 (OG-2), and Above-Grade (AG).

FIGURE 4.1: Visual Depiction of the Adaptive Engine's Multistage Process

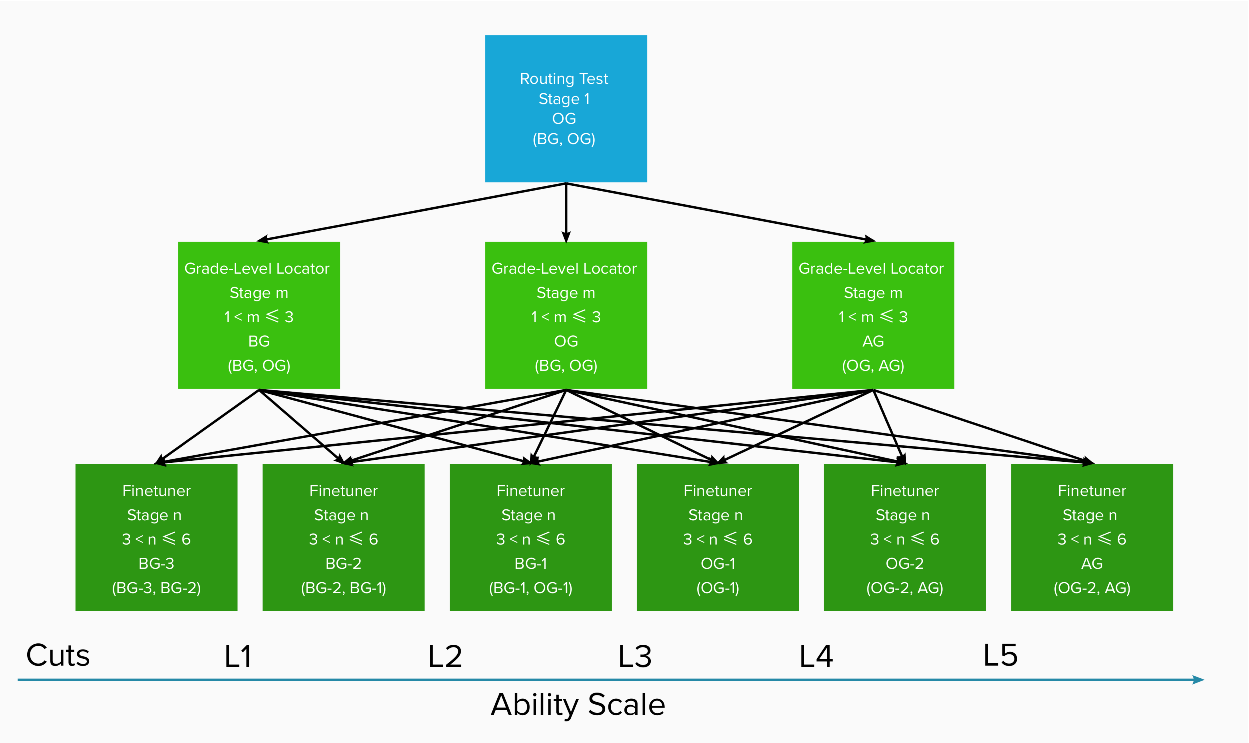
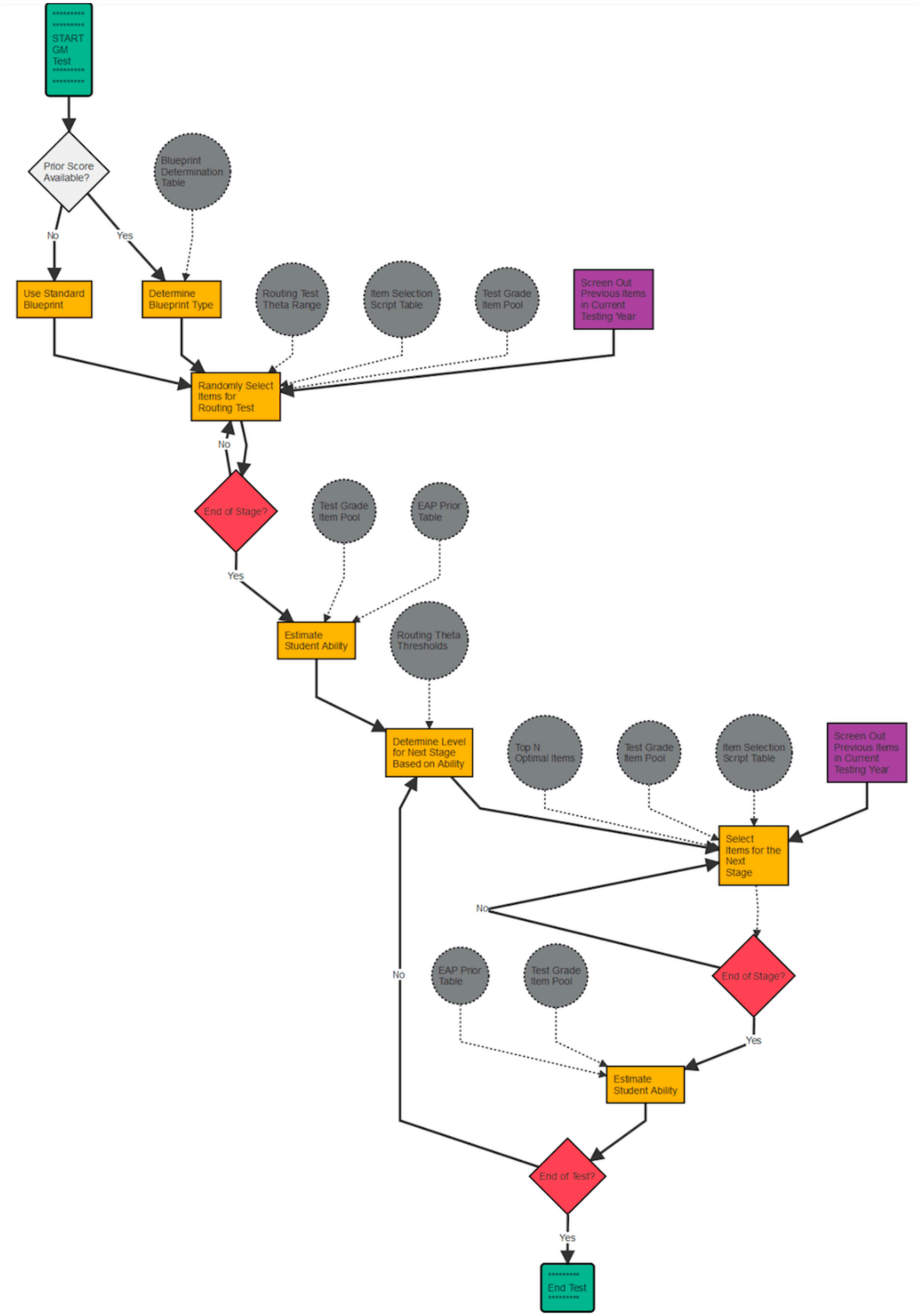


FIGURE 4.2: Diagram of the Multistage Adaptive Engine



Tagging

Tagging of items is at the core of the flexibility of the adaptive engine and is also core to the scoring and subscore reporting. The tagging system is very flexible and is made up of tag types and tags. A tag type is a container for tags and typically is a descriptive word for the tags. Typical tag types would be Grade, Subject, Domain, Item GLE, and DOK. Tags under each of these might include domain information, such as Key Ideas and Details, Craft/Structure, or Integration/Knowledge.

Each item in the item pool should be assigned tags from each of these tag types, which will allow the activity definition to select the tags at the appropriate level. It is important to note that care must be taken not to tag at too granular a level, as this will make the activity definitions and reporting more complex.

Activity Definitions

The activity definition allows the test developer control over all aspects of the test delivery, including the following:

- look and feel of the assessment
- available buttons and options for the environment
- content of the test
- adaptive algorithm settings
- termination criteria
- subscore definitions

The activity definition can either be predefined as an activity in the item pool or be generated on the fly by a host system at the point where the application programming interface (API) is being initialized. The most common use case is a combination of the two, which allows the core of activity definition to be configured and tested once, and then specific dynamic properties such as "Initial Ability Estimate" are updated at run time when the items' API is being initialized.

Content of the Test

The content of the test in an adaptive configuration is controlled by defining which tags the items should come from—and allowing the algorithm to pick the most appropriate items. The rules for selecting content can be done in broad strokes with the required tags setting or at a per-item level for a finer-grained control over the order of items. Defining the required tags settings allows the adaptive algorithm to make the choice as to which is the most appropriate item or item set to deliver to the test taker.

Adaptive Algorithm Settings

There are a number of adaptive algorithm settings that can be controlled for fine-tuning of the system. The options listed below have been rigorously tested and optimized for the delivery of GM content in an adaptive manner.

TABLE 4.1: List of Features Controlled by GM Adaptive Algorithm

Option	Description
Initial Ability	The initial ability measure for the student entering the adaptive assessment; used as the target difficulty for the first item and is considered that starting criteria
Item-Difficulty Tolerance	The size of the range of difficulties from which an item should be randomly selected
Item-Difficulty Offset	An offset to apply to each difficulty target, with positive values causing more-difficult items to be selected

Termination Criteria

For the GM, the termination criteria are set so that each student sees the same number of items. The test administration continues until it has fulfilled the termination criteria, ensuring the appropriate number of items covering the appropriate domains have been seen.

Subscores

Subscores enable the generation of scores based on different tags of the activity. A typical use case for this is to view the scores within specific content or knowledge domains assessed during a single activity. The following would be a common subscore usage.

TABLE 4.2: Example of Subscore Types

Subscore	Tag Type: Tag
Subscore 1	ELA Domain: Key Ideas and Details
Subscore 2	ELA Domain: Craft/Structure
Subscore 3	ELA Domain: Integration/Knowledge

4.5. Other Considerations

Following the principles of universal design, the GM was implemented with supporting resources to ensure that student participation is maximized. In addition, the Individuals with Disabilities Education Act (IDEA) and Every Student Succeeds Act (ESSA) provide guidelines for the inclusion of students with disabilities in educational assessment programs. Accommodations are changes in testing procedures that provide students with disabilities an equal opportunity to participate in testing situations and to demonstrate their knowledge and abilities. Test accommodations are grouped into the following categories: setting conditions, timing/scheduling conditions, presentation conditions, and response conditions.

Students with Individualized Education Programs

A student with disabilities, as classified under IDEA, has an individualized education program (IEP) that, in part, governs whether a particular assessment is appropriate for the student. The accommodation(s) that are provided for a student with a disability are decided by the student's IEP team and are documented in the IEP. English language learners (ELLs) should receive the same accommodations for the GM as they receive for other classroom assessments. This may include reading test items aloud verbatim in English or another language.

Accommodations

Research on how to assign accommodations (e.g., Kettler, 2012), on the effectiveness of accommodations (e.g., Gregg & Nelson, 2012), and on the equivalence of accommodations (e.g., Kim et al., 2009) has been conducted over many years and continues to be investigated. The accommodations available for the GM are separated into four categories: setting, timing, presentation, and response. Each accommodation is designed to address the specific matter that is interfering with a student's ability to answer a question based on his or her understanding of the content. Students should receive accommodations for the GM according to their individual plan (IEP or 504), just as they would for any other classroom assessment. Similarly, ELLs should receive the same accommodations for the GM as they receive for other classroom assessments.

Security and Uniformity of Test Administration

The GM is considered a secure test. It is important that assessment items be kept secure so this resource can continue to serve its purpose in a valid and reliable way. The intent of the security requirements is to preserve the integrity of the items, and therefore the reporting and the outcomes for students. Similar to any other type of assessment, results are invalid and not reliable if staff utilizes testing materials to prepare students for the assessment. Specific details regarding test security follow.

- At no time should tests or test items be used for instructional purposes.
- All teachers should be trained in administration procedures.
- Procedures and protocols should be in place to minimize the amount of time teachers have test materials before and after testing.
- Following testing, all secure testing materials (e.g., scrap paper) must be returned to the assessment coordinator for secure recycling. This may mean shredding or placing materials in a secure recycle bin.

4.6. Summary

The GM is an adaptive assessment. It is a powerful tool that provides greater accuracy of assessment across a wider range of abilities than traditional fixed-form assessments. The administration of the GM should be carefully communicated and executed through the use of detailed instructions. All standards related to test security, administration, and accommodations should be adhered to throughout the process.

5. STANDARD SETTING

One purpose of the GM assessment is to establish clear guidelines for educational decision making. By assigning meaning to test scores, standard setting allows stakeholders to make statements about the proficiency levels of individual students and groups of students. Many methodologies could be employed for this standard setting. The standard setting for the GM 4.0 is determined by a combination of psychometric judgement on GM test content, nationwide GM student data, and summative state assessment results, as well as policy input from field educators and the standard-setting committee. Our technical staff have used internal and external data to develop a proprietary method for setting standards.

For the GM BTS2023 release in July 2022 (GM 4.0), HMH enhanced the reporting of performance levels, increasing the number of levels from three to five, and also reset the GM standards. The number of cuts was increased from two (for three performance levels) to four (for five performance levels). Given that this assessment is used nationwide and that all states have their own expectations for performance, it was necessary to use a standard setting methodology that would offer some level of consistency in results with external criteria. The standard setting procedure for the five performance levels was determined using a modified briefing book methodology (Miles et al., 2010; O'Malley et al., 2012). This policy-based method is based on the briefing book methodology devised by Haertel (2002, 2008) and updated by Haertel, Beimers, and Miles (2012). It is informed by a briefing book, which is a compendium of information relevant to standard setting that is made available to the participants in the standard setting process. The briefing book process uses a variety of policy background, research data, test content information, and data about student performance in a comprehensive and focused fashion designed to structure participants' input and policymakers' deliberations.

The major focus of the standard setting was to answer the question of "*Given what states and districts are seeing in their own schools, what would be a reasonable percentage of students in each performance level?*" The assessment data used in the standard setting included test results from NWEA's MAP, Renaissance Star assessments, the National Assessment of Educational Progress, and a McKinsey study on the lingering effects of COVID-19. This collection of external data was then evaluated by a focus group of HMH experts (former educators, administrators, and policy makers) who made a determination of what would be appropriate percentages in each of the five new performance levels given what states and some larger key districts were seeing in their own assessment results.

5.1. Five Performance Levels

The classification of students' perfomance on the GM 4.0 has expanded from the three previous performance levels to five performance levels, which were established through a series of iterative analysis, evaluation of students being impacted, and discussions with the education researchers, educational practitioners, and psychometricians. Expanding to five performance levels allows more granular placement support, particularly at lower performance levels. Specifically, Test Grade K contains three performance levels, including *Approaching*, *On Level*, and *Above Level*. Test Grade 1 covers four performance levels, *Below Level*, *Approaching*, *On Level*, and *Above Level.* Test Grades 2 and above have five performance levels, including *Far Below Level*, *Below Level*, *Approaching*, *On Level*, and *Above Level*, which are illustrated as follows:

1. *Far Below Level*: GM scale scores less than or equal to G20 are classified as *Far Below Level*. The level denotes a lack of prerequisite knowledge and skills that are fundamental for on-grade work.
2. *Below Level*: GM scale scores between G21 and G45 are classified as *Below Level*. The level represents partial mastery of prerequisite knowledge and skills that are fundamental for on-grade work.
3. *Approaching*: GM scale scores between G46 and G60 are classified as *Approaching* level. This classification indicates mastery of prerequisite knowledge and skills that are fundamental for on-grade work.
4. *On Level*: GM scale scores between G61 and G80 are classified as *On Level*. This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over on-grade subject matter.

5. *Above Level*: GM scale scores between G81 and G99 are considered *Above Level*. Students achieving these scores demonstrate superior performance that are beyond on-grade work.

5.2. Percentile Rank Methodology

The setting of GM performance levels utilizes the approach of percentile ranks (PRs) to anchor students' scaled scores to performance levels for each Test Grade and subject. These percentile ranks for each performance level are defined as the psychometric targets in the GM 4.0. The effectiveness of these psychometric targets have been evaluated through a sequence of examinations by a standard-setting committee. Once the reasonable percentages were determined, these percentages were used to determine the cut scores on the underlying theta scale that were then translated into scaled score cuts.

Take ELA Test Grade 4, for example. The standard setting process determined percentages for each performance level as roughly around 25%, 20%, 20%, 25%, and 10%. As a result, the corresponding percentile ranks for each performance level of ELA Test Grade 4 were decided, as shown in the second column of Table 5.1. The established percentile ranks for each performance level will then dictate the theta score ranges for each performance level, based on the nationwide GM student data in school year 2021-2022. The theta score ranges for ELA Test Grade 4 are presented in the third column of Table 5.1. In Test Grade 4 ELA GM, given that the percentile rank for *Far Below Level* was $0 \leq PR \leq 25$ and the fact that the theta score corresponding to the 25th percentile rank was –1.3115, the theta score range associated with the *Far Below Level* at ELA Test Grade 4 was between negative infinity and –1.3115 (theta ≤ –1.3115). Similarly, the theta score corresponding to the 45th percentile rank was –0.1939, and the theta score range related to the *Below Level* category was –1.3115 < theta ≤ –0.1939. Theta score ranges were then linearly transformed into scaled scores with the equations presented in Chapter 6 of this technical overview. After the linear transformation, the scaled score ranges for each performance level at ELA Test Grade 4 were determined. These ranges are indicated in the fourth column of Table 5.1.

Note that the percentages of students for each performance level were individually determined for each Test Grade and subject in the standard setting process. The same methodology was applied to determine theta score ranges and scaled score ranges for each Test Grade and subject. The theta cut ranges and linear transformation formulas vary by subject and Test Grade. However, the GM scaled score ranges of these performance levels are consistent across all Test Grades and subjects. For example, G61 always means *On Level* and G81 always means *Above Level*. Through the standard setting process, appropriate performance level ability cuts on the logit scale (EAP ability scores) were transformed to GM scale scores so that the GM scale score cuts for each performance level have intrinsic meaning. This consistency for each performance level across Test Grades and subjects makes results easier for teachers and administrators to understand.

TABLE 5.1: Set Up Scale Scores for ELA Test Grade 4

Performance Level	Percentile Rank (PR)	Theta Score Range	Scaled Score (SS) Range
Far Below Level	0 ≤ PR ≤ 25	theta ≤ –1.3115	G01 ≤ SS ≤ G20
Below Level	25 < PR ≤ 45	–1.3115 < theta ≤ –0.1939	G21 ≤ SS ≤ G45
Approaching	45 < PR ≤ 65	–0.1939 < theta ≤ 0.4766	G46 ≤ SS ≤ G60
On Level	65 < PR ≤ 90	0.4766 < theta ≤ 1.3708	G61 ≤ SS ≤ G80
Above Level	90 < PR ≤ 100	theta > 1.3708	G81 ≤ SS ≤ G99

5.3. GM Psychometric Targets

The GM psychometric targets are effectively the percentage of students per performance level as well as the GM performance level scaled score cuts. The psychometric targets are evaluated yearly based on annual GM national student data. The GM 4.0 psychometric targets were developed based on the GM nationwide database with over one million students participating in school year 2021-2022. The involvement of the HMH Subject Matter Experts and Educational Researchers has further verified the psychometric targets for the GM 4.0. Additionally, the GM 4.0 psychometric targets were aligned with the expectations from the school districts, as well as the standard expectations of multiple state assessments, such as Florida Standards Assessment (FSA) and Colorado Measures of Academic Success (CMAS) that also offered five performance bands.

TABLE 5.2: Psychometric Targets per Performance Level by Test Grade

Test Grade	Scaled Score (SS)	Performance Level	Percentages of Students
K	SS ≤ G60	Approaching	65
	G61 ≤ SS ≤ G80	On Level	25
	SS ≥ G81	Above Level	10
1	SS ≤ G45	Below Level	30
	G46 ≤ SS ≤ G60	Approaching	35
	G61 ≤ SS ≤ G80	On Level	25
	SS ≥ G81	Above Level	10
2 to 3	SS ≤ G20	Far Below Level	20
	G21 ≤ SS ≤ G45	Below Level	20
	G46 ≤ SS ≤ G60	Approaching	25
	G61 ≤ SS ≤ G80	On Level	25
	SS ≥ G81	Above Level	10
4 to 11	SS ≤ G20	Far Below Level	25
	G21 ≤ SS ≤ G45	Below Level	20
	G46 ≤ SS ≤ G60	Approaching	20
	G61 ≤ SS ≤ G80	On Level	25
	SS ≥ G81	Above Level	10

Table 5.2 presents the detailed psychometric targets per Test Grade in both ELA and Math, including the thresholds of scaled scores per performance level and the percentages of students in specific performance levels. For Test Grade K, around 65% of national GM test takers would be classified as *Approaching* level (scaled scores between G46 and G60), 25% would be On Level (scaled scores between G61 and G80), and 10% would be *Above Level* (scaled scores between G81 and G99). For Test Grade 1, about 30% of national GM test takers were designated *Below Level*, a combined 60% of students were placed in *Approaching* and *On Level* (roughly 35% in *Approaching* and 25% in *On Level*), and 10% of students were *Above Level*. For Grades 2 and 3, a combined 40% of national GM test takers were placed in *Far Below Level* and *Below Level* (with roughly 20% of students in each level), and a combined 50% of students were in *Approaching* and *On Level* (with roughly 25% in each level), and 10% of students were in *Above Level*. For Grades 4 and

above, a combined 45% of national GM Test takers were in *Far Below Level* and *Below Level* (roughly 25% in *Far Below Level* and 20% in *Below Level*), and a combined 45% of students demonstrated *Approaching* and *On Level* performance (with roughly 20% in *Approaching* and 25% in *On Level*), and 10% of students were in *Above Level*.

5.4. Summary

Currently, there are no widely accepted standard-setting procedures for large-scale computerized adaptive testing (CAT) based on large item pools. The percentile rank methodology with a set of defined psychometric targets is specifically designed from the ground up to facilitate setting performance levels for CAT. These assessments usually contain large pools of items and require repeated testing time or multiple operational test forms. In addition, the procedures of standard setting for assessments typically involve multitudinous stakeholders with diverse provisions. The GM 4.0 standard setting has integrated as many needs as possible while keeping the whole process psychometrically sound.

6. SCORING AND REPORTING

6.1. Introduction

Assessment plays a key role in providing teachers and administrators with the data they need to carefully plan and deliver instruction. When benchmark assessment insights coexist with core program assessment data, teachers and students alike benefit from the ability to differentiate instruction and form small groups. The GM is included with HMH's connected programs. This means that a single GM ELA or Math assessment provides the data and content recommendations you need to optimize your time, in and out of the classroom. Developed by psychometricians, research scientists, and assessment specialists, the GM combines decades of research with real-world application to provide an intuitive, valid, and reliable assessment.

The purpose of scoring and reporting assessment data is to communicate test results to students, their parents or guardians, and their teachers. The GM reports provide useful information for determining the performance of students in a school or classroom. These reports help describe students' knowledge of a given set of expectations, allowing educators to determine specific instructional needs, measure student mastery toward any particular set of content standards, and evaluate educational programs. GM reports can also assist district administrators in their ability to see how their schools, classes, grades, and so on are performing within their district and provide information to help inform instructional next steps to those schools, grades, classes, and so on that aren't performing as they should be.

The following sections provide more information regarding the scores available from the GM. Table 6.1 gives a summary of student scores included in the reports.

TABLE 6.1: Summary of Student Scores

Student Scores
HMH Scaled Scores
Performance Levels
GLE
GLE Categories
Student Proficiency Indicator
Lexile/Quantile Intervals

6.2. Overview of Scores

A scaled score is a score that takes the student's primary ability estimate and adjusts it for differences in difficulty from one assessment to the next. The HMH Scaled Score is specific to each Test Grade and each subject. A grade-level scale has a range of G01 to G99, where *G* represents the Test Grade administered. G01 is the LOSS, and G99 is the HOSS. The LOSS and HOSS prevent extreme student scores from being transformed outside the desired range of the scale. For example, Test Grade 4 has a scaled score with a range of 401 to 499. The scale score metric for each test grade and subject is determined independently of those for other test grades and subjects. Therefore, direct comparisons should not be made across test grades or subjects.

When a GM Math test ends normally, a final expected a posteriori ability is estimated for the test taker. Using a linear transformation equation, which is available in each Test Grade (see below), this IRT expected a posteriori theta ability estimate (logit) is then converted to a corresponding reported scale score. In other words, there is a linear transformation for each Test Grade to map the IRT logit estimate to a reported scale score. This calculated ability estimate provides (via a linear transformation) both a scale score and all other reported scores that are derived from the scale score. If a student's estimated ability results in a scale score that is outside the range of allowed scores, then either the highest score (HOSS) or lowest score (LOSS) is applied.

The linear transformations for Test Grades K–11 Math to report a scale score based on a final expected a posteriori ability estimate are as follows:

$$\begin{aligned}
\text{GK Math SS} &= ((logit - 0.5371) * 22.3714) + 80 + 0 * 100 \\
\text{G01 Math SS} &= ((logit - 0.9312) * 18.2562) + 80 + 1 * 100 \\
\text{G02 Math SS} &= ((logit - 1.0283) * 20.8947) + 80 + 2 * 100 \\
\text{G03 Math SS} &= ((logit - 1.2498) * 21.0422) + 80 + 3 * 100 \\
\text{G04 Math SS} &= ((logit - 1.1826) * 24.4923) + 80 + 4 * 100 \\
\text{G05 Math SS} &= ((logit - 1.0758) * 27.2438) + 80 + 5 * 100 \\
\text{G06 Math SS} &= ((logit - 1.1619) * 26.4400) + 80 + 6 * 100 \\
\text{G07 Math SS} &= ((logit - 1.2445) * 24.6255) + 80 + 7 * 100 \\
\text{G08 Math SS} &= ((logit - 1.0326) * 28.7023) + 80 + 8 * 100 \\
\text{G09 Math SS} &= ((logit - 0.5703) * 31.7855) + 80 + 9 * 100 \\
\text{G10 Math SS} &= ((logit - 0.3757) * 36.3086) + 80 + 10 * 100 \\
\text{G11 Math SS} &= ((logit - 0.2741) * 26.1617) + 80 + 11 * 100
\end{aligned}$$

The linear transformations for Test Grades 2–11 ELA to report a scale score based on a logit (final) ability estimate are as follows:

$$\begin{aligned}
\text{G02 ELA SS} &= ((logit - 1.1057) * 17.2460) + 80 + 2 * 100 \\
\text{G03 ELA SS} &= ((logit - 1.2710) * 19.0070) + 80 + 3 * 100 \\
\text{G04 ELA SS} &= ((logit - 1.3484) * 22.3690) + 80 + 4 * 100 \\
\text{G05 ELA SS} &= ((logit - 1.3608) * 23.9477) + 80 + 5 * 100 \\
\text{G06 ELA SS} &= ((logit - 1.3010) * 24.7959) + 80 + 6 * 100 \\
\text{G07 ELA SS} &= ((logit - 1.2491) * 24.8576) + 80 + 7 * 100 \\
\text{G08 ELA SS} &= ((logit - 1.2771) * 24.7015) + 80 + 8 * 100 \\
\text{G09 ELA SS} &= ((logit - 1.0679) * 24.1165) + 80 + 9 * 100 \\
\text{G10 ELA SS} &= ((logit - 1.2764) * 23.3688) + 80 + 10 * 100 \\
\text{G11 ELA SS} &= ((logit - 1.3338) * 24.1512) + 80 + 11 * 100
\end{aligned}$$

The following are the defined LOSS and HOSS for Test Grades K–11 Math and Test Grades 2–11 ELA GM assessment scaled scores.

TABLE 6.2: GM Scale Score LOSS and HOSS for Math and ELA

Test Grade	Math LOSS	Math HOSS	ELA LOSS	ELA HOSS
K	1	99	–	–
1	101	199	–	–
2	201	299	201	299
3	301	399	301	399
4	401	499	401	499
5	501	599	501	599
6	601	699	601	699
7	701	799	701	799
8	801	899	801	899
9	901	999	901	999
10	1001	1099	1001	1099
11	1101	1199	1101	1199

GM Scaled Score

In the HMH GM, scale scores are statistical conversions of model-based ability estimates that maintain a consistent metric across test administrations (beginning of year [BOY], middle of year [MOY], end of year [EOY]) and permit direct comparison across all scale scores within a Test Grade and subject. Because scale scores adjust for different item difficulties administered across test events (a given with adaptive testing), they can be used to determine whether a student met the performance expectations in a manner that is fair across test events and administrations. School districts can also use scale scores (e.g., by averaging them) to compare the knowledge and skills of groups of students within a Test Grade and subject across years. These aggregate score comparisons can be used in assessing the impact of changes or differences in instruction and/or curriculum.

The scale scores for a given subject and Test Grade range from G01 to G99, where *G* represents the Test Grade administered. Table 6.3 presents the scale score range for each Test Grade. In the GM 4.0, scale scores G20, G45, G60, and G80 typically represent the upper-bound scale scores for *Far Below Level*, *Below Level*, *Approaching*, and *On Level*, respectively. Scale scores between G81 and G99 indicate the score range of the *Above Level* performance band. For a detailed illustration of each performance level with associated scaled scores, readers are referred to Chapter 5. Note that the scale score metric for each Test Grade and subject was determined independently, since the scale scores are related to the item pool at each specific Test Grade and subject. Therefore, direct comparisons should not be made across Test Grades or subjects. With the increase of one Test Grade, 100 score points are increased in the scaled score. Each Test Grade has a LOSS and a HOSS, which are theoretically G01 and G99, but typically a natural effective truncation occurs on each end of the scale for nearly all test scores.

TABLE 6.3: GM Scale Score Range

Test Grade	Scale Score Range
K	1–99
1	101–199
2	201–299
3	301–399
4	401–499
5	501–599
6	601–699
7	701–799
8	801–899
9	901–999
10	1001–1099
11	1101–1199

Performance Levels

To help parents/caregivers and schools interpret scale scores, performance levels are reported. The GM 4.0 reports five performance levels based on students' scaled scores. Readers are referred to Chapter 5, in which each performance level has been illustrated in detail. Table 6.4 lists the scale scores that are included in each performance level and Test Grade. Counts and percentages of students attaining each performance level can be aggregated at the classroom, school, or district level. The same performance levels provided for the GM scaled score are also given for each reported domain for ELA and Math.

TABLE 6.4: GM Performance Levels in Math and ELA

Test Grade	Performance Level	Scaled Score Band
K	Approaching	G01 to G60
	On Level	G61 to G80
	Above Level	G81 to G99
1	Below Level	G01 to G45
	Approaching	G46 to G60
	On Level	G61 to G80
	Above Level	G81 to G99
2 to 11	Far Below Level	G01 to G20
	Below Level	G21 to G45
	Approaching	G46 to G60
	On Level	G61 to G80
	Above Level	G81 to G99

GLE

The GLE score is a placement indicator developed by HMH, intending to provide a quasi-functional grade-level score based on the linkage to the Lexile/Quantile frameworks and the standard setting approach for scaled scores and performance levels. In the GM 4.0 (which was released in July 2022), the Lexile/Quantile frameworks provide the range of integer GLE scores (e.g., GLE 4, 5, and 6, etc.) for each Test Grade, and the standard setting procedure gives the psychometric targets (percentages of students) for each GLE category per Test Grade and subject. Referring to the standard setting procedure ensures that the GLE scores derived from the scaled scores and expected a posteriori scores are consistent with the placement of students based on their performance levels. In other words, the grade level as indicated by a student's GLE score is expected to align with the grade level information derived from this student's performance level. However, the GLE scores are more granular than the performance levels. One performance level may correspond to multiple GLE scores.

Table 6.5 presents the psychometric targets for each GLE category in Test Grade 4, as well as the operational definition of the GLE categories. For example, when the integer value of a student's GLE score minus this student's Test Grade was less than –2, this student's GLE category was *>2 Grade Levels Below* (More than 2 Grade Levels Below). Overall, about 25% of students in Test Grade 4 were placed in the GLE category of *>2 Grade Levels Below*. The GLE category definition in other Test Grades are the same as in Test Grade 4. The psychometric targets for the GLE categories may be slightly different across Test Grades. Referred to Chapter 5 for all the psychometric targets.

TABLE 6.5: GM GLE Psychometric Targets in Test Grade 4

GLE Minus Test Grade	GLE Category	Psychometric Targets
Less than –2	>2 Grade Levels Below	25%
–2	2 Grade Levels Below	20%
–1	1 Grade Level Below	20%
0	On Grade Level	25%
Equal to or greater than 1	Above Grade Level	10%

Student Proficiency Indicator

Overview

The student proficiency indicator (SPI) is an indicator for growth measure developed by HMH. The SPI is a growth model that provides a criterion-referenced gain measure designed to determine if a targeted gain score is maintained between administrations (e.g., fall to spring). The SPI has meaningful connections to grade-level content expectations and provides a growth target of 100 for all Test Grades and subjects. This greatly simplifies the interpretation of growth for teachers, administrators, parents or guardians, and students. The criterion-referenced target of 100 for each SPI score assures that a student has displayed enough growth to support their progression toward learning more-advanced concepts and skills within a school year.

The target value for every student is an SPI of 100 throughout a school year. Targeted growth is a range of +/– 5 around the 100 SPI value, low growth (*Did Not Meet Targeted Growth*) is indicated by SPI values below 95, and high growth (*Exceeded Targeted Growth*) is indicated by SPI values above 105. The SPI is a criterion-referenced growth model that allows educators to categorize a student's growth from one assessment to the next within a school year. The SPI provides insight into whether the student is growing according to yearly goals, without having to compare a student to their peers (i.e., reliance on growth norms). To help make peer-growth comparisons that are relevant, the new BTS2023 GM assessments will report Expected Growth Targets, which show the average national change in scaled scores in the student's current performance level. It can be used to evaluate whether the student is lagging or leading their peers in their performance between administrations.

The SPI uses a criterion-referenced growth indication by comparing a student's growth to that of a criterion-referenced target employed to define the following:

- catch-up growth
- targeted growth
- maintenance growth

The SPI is designed to gauge student growth against content standards and, hence, can better inform teachers, students, and parents; and it operates separately for each Test Grade, which means it is psychometrically more valid. Moreover, the SPI addresses another critical issue that other growth measures commonly overlook: the phenomenon of regression to the mean in imperfect measures. Lord (1956, 1962) pointed out over half a century ago that, when a measurement instrument is not perfectly reliable—that is, when it contains a certain extent of measurement error because of item sampling, environmental variability, and/or temporary changes in the person's state—individuals with extreme scores in the Time 1 test tend to regress to the mean of the population in the Time 2 test as an artifact of the measurement error. In other words, a high achiever who achieves a top score in the Time 1 test tends to score lower in the Time 2 test simply as a statistical artifact. Hence, overlooking this issue and reporting growth based on the observed score difference between Time 1 and Time 2 can unfairly penalize those high achievers with a wrong label of not achieving growth or falling behind.

The SPI tackles this critical problem. Formulas derived by Lord (1956, 1962) were used to estimate the true score gain as opposed to the observed score gain. The growth targets are further adjusted for the extreme-ends students to assess them against a fairer standard that incorporates the notions of accelerated and maintenance growth. For further discussion of the SPI, see Chapter 8, which contains the statistical summary of evidence for the usefulness of this growth model.

Lexile/Quantile Intervals

HMH provides Lexile and Quantile intervals predicted from each student's GM scale scores. The predicted Lexile interval of the student's reading ability provides the upper and lower range that helps match the student with literature appropriate for their reading skills. The predicted Quantile interval of the student's mathematical ability provides the upper and lower range that helps place the student's test performance in the appropriate location within the Quantile Framework of mathematical skill development. Both the Lexile and Quantile intervals have a range of 150 score values. The predictive model from GM scale scores to Lexile and Quantile was established through an initial study in 2016 and a follow-up study in 2022.

Background

In 2016, HMH conducted a study in partnership with MetaMetrics to establish an initial correspondence between GM scores and Lexile/Quantile measures. The result of this study was a score association presenting a range of the Lexile/Quantile scores for each Continuum total scale score. These linkages provided the initial research-based connections from GM scale scores to both the Lexile and Quantile frameworks and were developed by the HMH Learning Sciences Division. The Lexile psychometric linking methodology employed in the 2016 study was a single-group design where 14 embedded Lexile-based items, which all have theoretical Lexile measures, were administered as part of the GM ELA assessment. These items were scored to provide a direct estimate of a Lexile measure. At the same time, these items along with the other 16 GM items (passage-based and language items) were scored using the GM IRT item difficulties. The study provided a direct link of a Lexile measure to a GM scale score. Nonparametric regression methods were used to estimate a smoothed relationship between these two scores that served to construct the GM-to-Lexile linking table. The Quantile psychometric linking methodology employed in 2016 was a scaling analysis on the MI 3.1 assessment data that had the Quantile scores serving as the reported score scale. As part of this study, the GM scaling methodology (see Section 3.5) was applied to each of the MI (Math Inventory) 3.1 Level Tests, and the resulting G01–99 scale score was linked to the corresponding estimated Quantile score. Thus, each MI 3.1 student score had both a reported Quantile score and an associated GM G01–99 scale score. Finally, nonparametric regression methods were used to estimate a smoothed relationship between these two scores that served to construct the GM-to-Quantile linking table.

In 2022–2023, HMH was able to revisit the 2016 study by conducting an updated study of the relationship between GM scores and Lexile measures, as reported by the Reading Inventory, using several years of data collected from both measures. As part of this follow-up study, large national samples of data were used to describe the distribution of scores on each scale. Then scores that aligned with the same percentile of each distribution were matched to one another, providing a correspondence between each scale.

Method

All data from the Fall Reading Growth Measure testing to date were used to summarize the distribution of GM scale scores. This includes results from the 2020–2021, 2021–2022, and 2022–2023 school years. The data were combined to provide the most representative view of the GM scaled score distributions possible. These students came from schools and districts across the United States. A summary of the participation in the Fall Reading Growth Measure testing to date is provided in Table 6.6.

TABLE 6.6: Growth Measure Participation Summary

Grade	Districts	Schools	Students
2	1,161	2,686	119,647
3	1,328	3,210	185,985
4	1,386	3,361	199,861
5	1,369	3,289	216,279
6	1,229	2,355	241,201
7	1,069	1,953	245,557
8	1,018	1,963	247,328
9	707	1,230	196,624
10	714	1,202	175,539
11	706	1,204	216,355

A distribution summary of the GM ELA scale scores in this data set is provided in Table 6.7.

TABLE 6.7: Growth Measure Distribution Summary

Grade	*N*	Min. Scale Score	Max. Scale Score	Average Scale Score	*SD* Scale Score
2	119,647	201	299	252.577	20.378
3	185,985	301	399	349.232	21.570
4	199,861	401	499	446.509	23.222
5	216,279	501	599	546.686	22.829
6	241,201	601	699	647.041	23.210
7	245,557	701	799	748.016	22.970
8	247,328	801	899	847.564	23.511
9	196,624	901	999	949.637	22.918
10	175,539	1,001	1,099	1,050.338	22.791
11	216,355	1,101	1,199	1,149.037	23.541

Data from five years (2012–2013, 2013–2014, 2014–2015, 2015–2017, 2017–2018) of the Reading Inventory were used to summarize the distribution of Lexiles reported from the Reading Inventory. These data were combined to provide the most representative view of the Lexile distributions. The students in this sample came from schools and districts across the United States. A summary of the distribution of Lexiles in each grade is provided in Table 6.8.

TABLE 6.8: Lexile Distribution Summary

Grade	*N*	Min.	Max.	Average	*SD*
K	1,922	–661	1,395	–297.9253938	329.8223
1	25,802	–660	1,802	–156.0702458	379.0680
2	140,166	–657	1,891	175.6516572	366.3068
3	385,407	–541	1,999	375.5173134	353.8305
4	443,545	–541	2,003	542.9359369	335.4527
5	466,683	–541	1,999	677.2201198	321.3867
6	715,119	–541	2,007	746.5803502	328.8038
7	713,034	–541	2,011	797.2579215	355.3217
8	663,302	–541	2,011	877.0364569	346.9386
9	506,223	–541	2,011	861.5837815	367.9952
10	313,246	–541	2,011	928.8243263	371.4927
11	178,917	–541	2,011	976.152727	375.4676
12	123,070	–540	2,010	998.9870998	386.9464

Finally, Reading Inventory Lexile results are used as part of the *Read 180* program. Both eligibility and placement decisions are based on Lexiles. The majority of the historical data on the use of Lexiles for *Read 180* eligibility and placement is based on Lexile as reported by Reading Inventory. However Lexile is now also available from the GM. To ensure the validity of these decisions, the impact of implementing the Lexile as reported by GM as compared to Reading Inventory was investigated.

First, eligibility for *Read 180* is determined by establishing whether the student achieved proficiency on the Reading Inventory during the preceding spring. Students who were proficient given their Lexile from Reading Inventory and the Reading Inventory proficiency cuts would not eligible to start *Read 180* the following fall. Students who were not yet proficient based on their Lexile from Reading Inventory would be eligible for *Read 180* the following fall. Therefore, it is important to describe the impact of implementing the GM-based Lexile ranges on *Read 180* eligibility decisions. The data used for the eligibility analysis consist of national data from the EOY testing window over the same five years used to establish the percentiles for the Reading Inventory as described above. The sample participation is included in Table 6.9.

TABLE 6.9: Reading Inventory EOY Participation for Subsequent Read 180 Eligibility Grade

Test Grade	Testing Window	*Read 180* Eligibility Grade	*N*
2	EOY	3	334,294
3	EOY	4	566,307
4	EOY	5	603,030
5	EOY	6	689,866
6	EOY	7	954,980
7	EOY	8	916,417
8	EOY	9	837,426
9	EOY	10	586,924
10	EOY	11	327,036

Second, an automatic procedure is applied such that *Read 180* students are placed in one of six increasingly difficult levels of *Read 180* programming, based on their Lexile. The distribution of placements from users in two test years (2017–2018 and 2020–2021) is shown in Tables 6.10 and 6.11. It is important to describe the impact of using the GM-based Lexiles on Read 180 placement decisions. The samples for each of these summaries are based on 27,203 students in Grades 3–12 who used Read 180 during the 2017–2018 academic year and 82,942 students in Grades 3–12 who used Read 180 during the 2020–2021 academic year. These distributions (in Grades 3–11) were compared to that of placements resulting from the linked scores using the updated GM-to-Lexile correspondence as the final step of the study.

TABLE 6.10: *Read 180* Placement across Six Comprehension Levels (2017–2018)

Grade	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
3	77%	10%	8%	5%	0%	0%
4	81%	10%	6%	2%	0%	0%
5	63%	19%	12%	5%	1%	0%
6	46%	19%	20%	12%	3%	0%
7	44%	14%	19%	16%	6%	1%
8	37%	13%	17%	19%	11%	2%
9	32%	12%	16%	20%	15%	4%
10	32%	10%	14%	21%	17%	6%
11	38%	11%	11%	16%	17%	7%

TABLE 6.11: *Read 180* Placement across Six Comprehension Levels (2020–2021)

Grade	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
3	87%	7%	3%	1%	1%	0%
4	78%	13%	6%	2%	1%	0%
5	59%	20%	14%	5%	2%	1%
6	39%	19%	22%	14%	5%	1%
7	37%	14%	20%	18%	9%	2%
8	33%	13%	18%	19%	13%	4%
9	29%	11%	16%	19%	16%	8%
10	31%	10%	14%	19%	17%	9%
11	39%	10%	12%	17%	12%	9%

Results

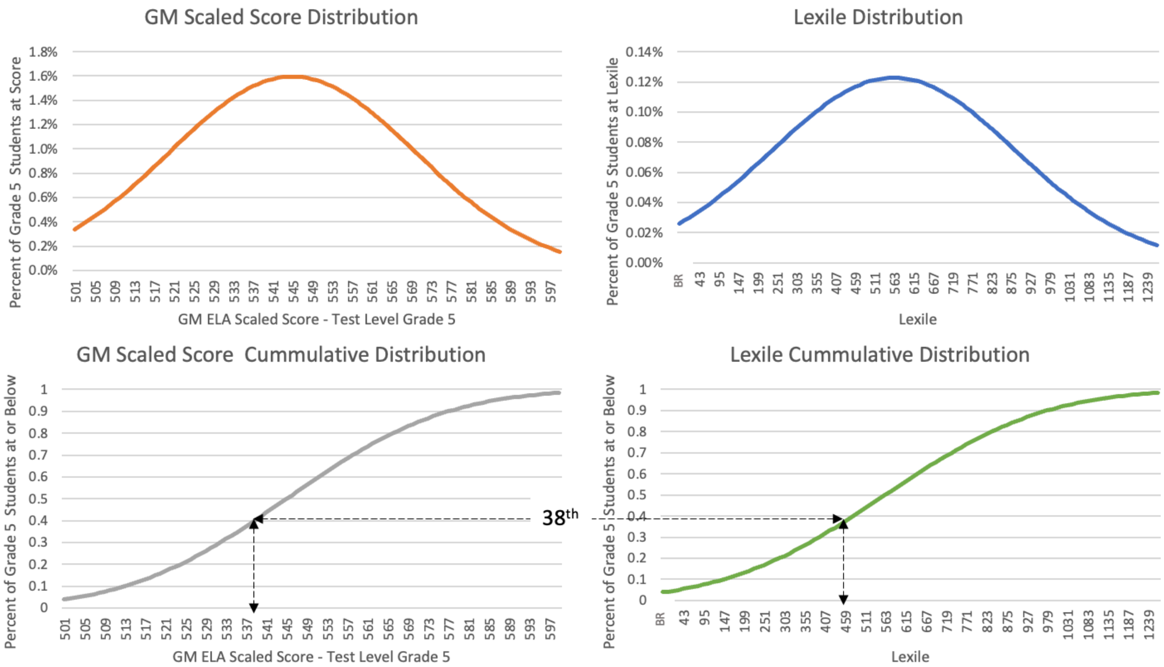
Using the GM and the five-year Reading Inventory Lexile distribution percentiles as a matching variable, a linking table was established between GM scale scores and Reading Inventory Lexile. An excerpt from the table describing this correspondence is shown in Table 6.12.

TABLE 6.12: Excerpt from the GM ELA Scale Score to Lexile Linking Table

Grade	GM Scale Score	Percent of Students at or below Score (Min.)	Percent of Students at or below Score (Min.)	Lexile Midpoint	Lexile Range Min.	Lexile Range Max.
4	430	22.73	23.83	350	275	425
4	431	23.84	24.96	360	285	435
4	432	24.97	26.27	375	300	450
4	433	26.28	27.56	390	315	465
4	434	27.57	28.84	400	325	475

The table describes a non-linear relationship between GM ELA scale scores and Lexiles that is based on the distributions of both of these scales as illustrated in figure 6.1.

FIGURE 6.1: Illustration of Linking Procedure



The distributional linking table described in this section is designed to ensure that similar decisions would be made about students using either scale, as shown in Tables 6.13, 6.14, and 6.15. It should be expected that similar proportions of students would be eligible for *Read 180* and that similar proportions of students would be placed in each level of *Read 180* using the original scores from Reading Inventory or linked scores from the GM. To investigate this hypothesis, the linking table described above was also applied to the eligibility sample and to two test sets of *Read 180* users to assess the implications of implementing the new linking table. Results show that differences in the distributions of eligibility and placement that result from using Reading Inventory Lexiles or linked GM scale scores are minimal (between 0% and 2%).

TABLE 6.13: Differences between *Read 180* Linked Level Eligibility from the GM and Reading Inventory

Eligibility Grade	Eligibility Category	Min. Lexile	Max. Lexile	Percent (Reading Inventory)	Percent (GM - Linked)	Difference
3	Not Eligible	420	1997	47%	48%	1%
3	Eligible	–541	419	53%	52%	–1%
4	Not Eligible	520	1894	49%	49%	0%
4	Eligible	–655	519	51%	51%	0%
5	Not Eligible	740	2003	42%	42%	0%
5	Eligible	–539	739	58%	58%	0%
6	Not Eligible	830	2003	43%	43%	0%
6	Eligible	–541	829	57%	57%	0%
7	Not Eligible	925	2006	34%	34%	0%
7	Eligible	–541	924	66%	66%	0%
8	Not Eligible	970	2003	36%	36%	0%
8	Eligible	–542	969	64%	64%	0%
9	Not Eligible	1010	2011	40%	39%	–1%
9	Eligible	–541	1009	60%	61%	1%
10	Not Eligible	1050	2008	33%	34%	1%
10	Eligible	–541	1049	67%	66%	–1%
11	Not Eligible	1080	2009	38%	39%	1%
11	Eligible	–540	1079	62%	61%	–1%

TABLE 6.14: Differences between *Read 180* Linked Level Placement and Original Level Placement for the 2017–2018 Sample

Grade	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
3	–2%	1%	1%	0%	0%	0%
4	0%	0%	0%	0%	0%	0%
5	1%	0%	–1%	0%	0%	0%
6	–1%	0%	0%	0%	0%	0%
7	0%	1%	0%	–1%	0%	0%
8	1%	0%	–1%	1%	0%	0%
9	0%	1%	–1%	–1%	1%	0%
10	0%	–1%	1%	0%	0%	0%
11	1%	0%	0%	0%	–1%	1%

TABLE 6.15: Differences between *Read 180* Linked Level Placement and Original Level Placement for the 2020–2021 Sample

Grade	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
3	0%	0%	0%	–1%	0%	1%
4	–1%	0%	0%	0%	0%	0%
5	1%	–2%	1%	0%	0%	0%
6	0%	–1%	2%	–1%	0%	0%
7	1%	–2%	2%	–1%	0%	0%
8	0%	0%	1%	1%	–1%	0%
9	1%	–1%	0%	–1%	0%	0%
10	0%	–1%	1%	–1%	1%	0%
11	0%	0%	–1%	2%	0%	–1%

6.3. Reports

As previously described, the GM is a research-based and adaptive benchmark assessment for ELA and Math. It uses assessment performance to connect to HMH's curriculum and instruction resources—including *Into Math*, *Into Reading*, *Waggle*, and *Into Literature*. The GM delivers valid and reliable achievement scores, including an overall scaled score, overall performance categories, domain performance category, SPI, and GLE, GLE Category, Lexile Interval and Lexile Midpoint (ELA) or Quantile Interval and Quantile Midpoint (Math).

The GM provides a variety of reports. The reports available include the following:

- Teacher Growth Reports—Class and Student Views

- Teacher Growth Measure Student Details
- Teacher Standards Report
- Administrator/District Leader Growth Reports

7. RELIABILITY

7.1. Introduction

Reliability refers to the consistency of student test scores. *Measurement error* refers to the random variability in test scores. Both are indicators of the degree of precision in scores of a test. In general, measurement error and reliability are inversely related. When measurement error is large, reliability is small. Increasing reliability by minimizing measurement error is an important goal in the construction of any test. Errors in measurement can result from any of a multitude of factors, including environmental factors (e.g., testing conditions) and examinee factors (e.g., fatigue, stress). Feldt & Brennan (1989) note that "quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis" (p. 105). The expected score reliability produced from the adaptive assessment was evaluated by examining real data and conducting simulations based on the operational item pools that compose the GM.

7.2. Reliability: Classical Test Theory Perspective

Classical test theory (CTT) provides a means for the quantification of measurement error due to examinee-related factors. The CTT approach builds on the notion of an ideal, error-free test score. The test score is defined as a composite of a true, ideal score and its associated error. Test reliability is a measure of such error, which can be estimated via the correlation of scores on equivalent test forms (equivalence reliability) or from test-retest data (stability reliability). It can also be estimated from a single test administration (internal-consistency reliability) using any one of a variety of techniques (Brown, 1910; Cronbach, 1951; Kuder & Richardson, 1937).

Generally, CTT reliability is the ratio of the variance of the (unknown) true score and the variance of the observed test score:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$$

IRT, on the other hand, has an analog index to CTT reliability. One such index, marginal reliability, was suggested by Green et al. (1984). This index follows the fundamental definition of reliability as the proportion of variance in the observed score owing to true score variance. It is estimated from the parameters associated with the variance of the ability scores and the average of the error variance, and it ranges from zero (not reliable) to one (highly reliable). Table 7.1 provides the marginal reliabilities for the GM by content area and grade based on the current set of item pools and the abilities of the observed population of examinees for the 2021/2022 school year. As shown in Table 7.1, the marginal reliabilities are high and the item pools will produce reliable estimates of student achievement on the GM.

TABLE 7.1: Reporting Domains for Math and ELA GM

Marginal Reliability		
Grade	*ELA*	*Math*
K	-	0.87
1	-	0.87
2	0.86	0.87
3	0.86	0.87
4	0.86	0.87
5	0.86	0.87
6	0.86	0.86
7	0.86	0.86
8	0.86	0.86
9	0.86	0.87
10	0.86	0.87
11	0.86	0.87

7.3. Reliability in an Adaptive Assessment

In an adaptive assessment, as the student progresses through the test, their items are selected based on performance in previous items. The goal of the adaptive algorithm is to construct a test unique to each examinee and targeted to the examinee's ability level as measured throughout the assessment. The observed test score under adaptive conditions is generally measured more precisely than in a fixed-form environment, which is one of the advantages of an adaptive assessment. Although a different set of test items is administered to each student, scores from different students are comparable because each test is measuring the same content.

The purpose of this section is to provide evidence of reliability associated with the adaptive assessment engine used for the GM. These analyses allow HMH to evaluate the implementation fidelity and quality of the adaptive item-selection algorithm to assess the degree to which individualized test forms match the intended test specifications and measurement precision.

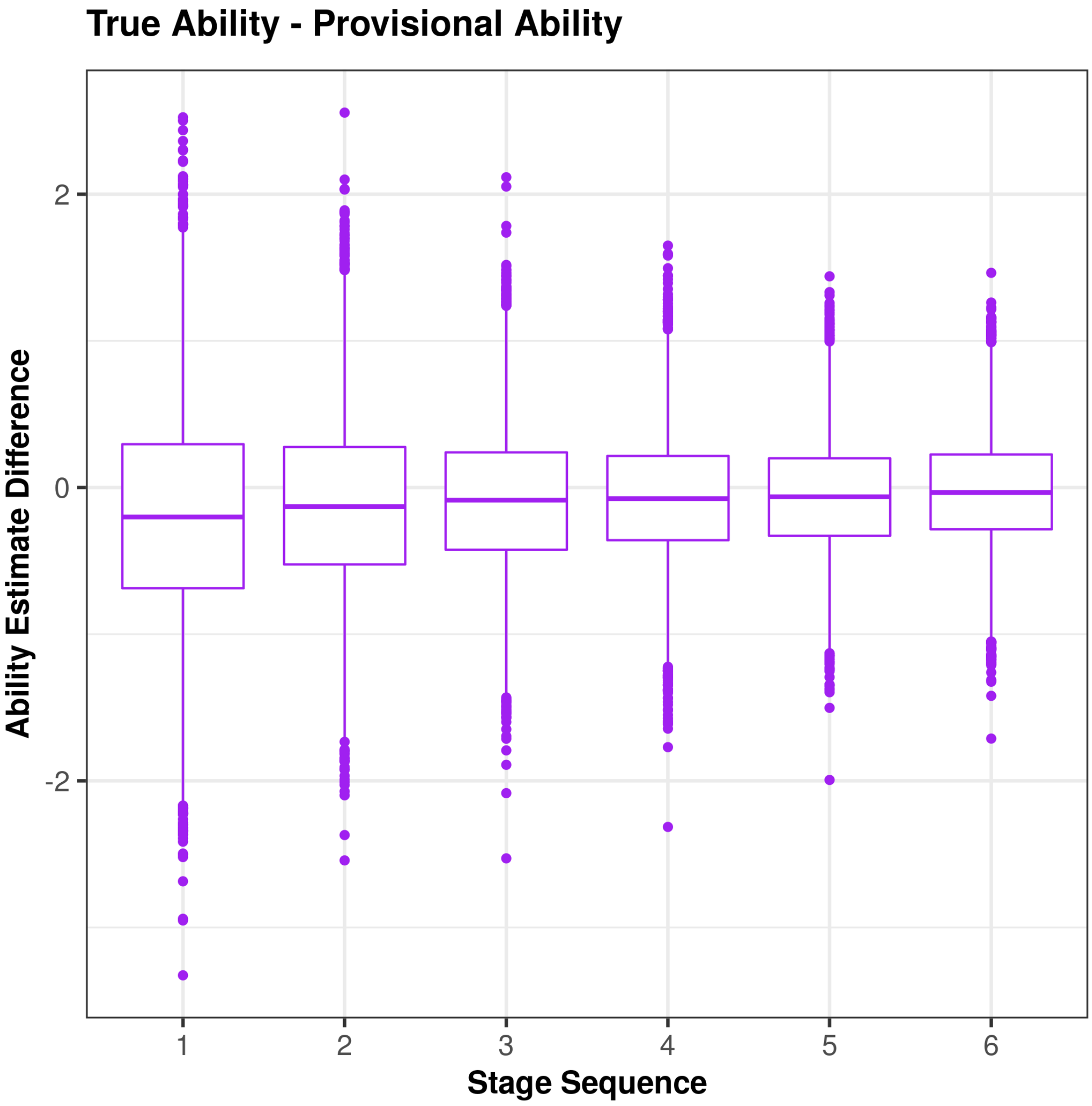
The conditional standard error of measurement (CSEM) is reported in Table 7.2 at the 5th, 25th, 50th, 75th, and 95th percentiles of student ability by grade and content area. All values are around the generally accepted value of 0.30, which demonstrates good measurement precision across all ability levels.

Furthermore, Figure 7.1 provides a graphic based on the Math Grade 6 adaptive test simulation that illustrates how the precision in the provisional (stage-level) ability estimates increases as an examinee progresses through the test.

TABLE 7.2: GM Statistical Summaries of CSEMs by Grade and Content Area

Content	Grade	Overall Average CSEM	Avg. CSEM at 5th Percentile	Avg. CSEM at 25th Percentile	Avg. CSEM at 50th Percentile	Avg. CSEM at 75th Percentile	Avg. CSEM at 95th Percentile
Math	K	0.383	0.374	0.371	0.376	0.385	0.412
	1	0.388	0.371	0.365	0.374	0.393	0.455
	2	0.384	0.376	0.371	0.376	0.388	0.418
	3	0.379	0.375	0.365	0.367	0.380	0.410
	4	0.380	0.378	0.369	0.373	0.380	0.407
	5	0.377	0.386	0.366	0.367	0.375	0.398
	6	0.378	0.391	0.369	0.368	0.375	0.395
	7	0.384	0.399	0.377	0.370	0.375	0.422
	8	0.380	0.405	0.375	0.369	0.371	0.389
	9	0.383	0.427	0.382	0.373	0.369	0.380
	10	0.381	0.420	0.383	0.370	0.367	0.373
	11	0.387	0.438	0.392	0.376	0.369	0.378
ELA	2	0.410	0.422	0.381	0.382	0.414	0.486
	3	0.398	0.399	0.371	0.383	0.403	0.450
	4	0.391	0.394	0.369	0.373	0.395	0.446
	5	0.400	0.413	0.381	0.379	0.397	0.451
	6	0.390	0.400	0.367	0.371	0.394	0.439
	7	0.390	0.400	0.369	0.373	0.391	0.435
	8	0.395	0.406	0.373	0.375	0.395	0.450
	9	0.395	0.412	0.376	0.377	0.393	0.441
	10	0.399	0.405	0.373	0.378	0.400	0.469
	11	0.399	0.402	0.376	0.380	0.399	0.468

FIGURE 7.1: Math Grade 6 Precision in Stage-Level Ability Estimates



8. VALIDITY

According to the Standards for Educational and Psychological Testing (AERA et al., 2014), "ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system . . . and includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question" (p. 22). While this chapter summarizes evidence that supports claims about the validity of GM scores, many other parts of this technical report also provide appropriate evidence for validity. Elements of this evidence are cross–referenced below for added convenience. The procedural and empirical evidence available along with the rationale presented below provide support for the standards–based interpretations of the GM.

8.1. Evidence Based on Test Content

Linn et al. (2002) suggest that "two questions are central in the evaluation of content aspects of validity. Is the definition of the content domain to be assessed adequate and appropriate? Does the test provide an adequate representation of the content domain the test is intended to measure?" (p. 6). The following sections help answer these two very important questions. Chapter 2 of this technical report details the steps completed to generate the test items and pools. Each test item was developed and reviewed in a rigorous quality–control process that used former educators as subject matter experts in the areas of content and bias. The content review specifically focused on the alignment of items to the CCSS and on item–writing best practices. The bias review focused on properties of universal design, with an emphasis on ensuring fairness and sensitivity to bias issues and preventing differential item functioning. Additional details about the development of test items can be found in Chapter 2 as well. Finally, the use of subject matter experts during the data review resulted in the removal of items from the GM item pool that were not suitable for an operational item pool. The item pools for the GM were built only with items that passed each step of the entire quality–control process.

8.2. Evidence Based on Response Processes

The GM is composed of selected–response items. As students respond to selected–response items, they must choose the best answer from the three or four options listed. To help ensure that appropriate answer choices were generated, a distractor rationale was created for each incorrect answer choice. This rationale specifies the misconception or error in thought that would lead a student to choose the incorrect answer. Thus, during the item–writing process, incorrect answer choices are created as plausible and/or attractive choices to students who do not fully understand the competency and objective being assessed. The use of distractor rationales is part of a best–practice system for item development and reduces the possibility of construct irrelevance.

8.3. Evidence Based on Special Studies

To support the introduction of the next generation of the HMH Growth Measure assessments, the HMH Psychometric Research and Data Analytics team completed a series of empirical studies using student–level GM 4.0 score data collected in school year 2021/2022. The results show that the GM has sufficient validity in the following areas:

- accurately measuring student growth across all grades
- precisely assessing low–performing students' ability
- establishing reasonable performance categories
- interpreting growth and placement metrics
- evaluating SPIs versus SGPs

Empirical results show that the GM scaled score, SPI, and placement metrics, which include the GLE, GLE categories, and performance levels, provide a relatively easy communication and analysis narrative for the reported test scores. The newly developed GM (version 4.0) performs better than its predecessors, since it includes (1) larger item bank, (2) more–robust algorithm features, (3) improved item development and construction, and (4) a more diverse set of content standards. In the following sections, summary results of the GM 4.0 version of the SPI model metric are presented for the population tested.

Summary of Measuring Growth via SPI

Table 8.1 provides the within–year (from fall [BOY] to spring [EOY]) summary of the SPI using empirical GM 4.0 data for the 2021/2022 school year. It includes the median and standard deviation of SPI, the percentages of students falling into each SPI level, and the percentage of students having SPI $\geq$ 100 (at or above targeted growth). The SPI had median between 95 and 100 in both ELA and Math. In ELA Grades 5 and below, targeted growth had a similar or slightly higher percentage than the low growth. Targeted growth defined by the SPI typically had higher percentages in Math. Additionally, about half of students in ELA had SPI $\geq$ 100 in these grades. The proportion for Math was about 60%. Table 8.1 illustrates that the GM is measuring growth appropriately and with a reasonable SPI allocation of growth categories across grades and content areas.

TABLE 8.1: Summary of ELA SPI Values

		SPI			% SPI Level			
Subject	*Test Grade*	N	*Median*	SD	*Low*	*Targeted*	*High*	*Percent of SPIs ≥ 100*
ELA	2	15,915	102	11.6	25	39	37	60
	3	30,761	100	12.0	33	38	29	51
	4	34,059	97	12.7	44	34	22	41
	5	34,908	95	13.0	48	31	20	37
	6	46,368	93	13.4	55	28	17	31
	7	47,363	93	13.4	54	30	16	31
	8	46,535	93	13.2	55	29	16	31
	9	29,604	98	13.6	41	33	27	44
	10	23,246	94	13.5	52	31	18	33
	11	22,447	91	14.7	59	26	15	27

TABLE 8.2: Summary of Math SPI Values

		SPI			% SPI Level			
Subject	*Test Grade*	N	*Median*	SD	*Low*	*Targeted*	*High*	*Percent of SPIs ≥ 100*
Math	K	8,530	105	13.9	23	29	48	66
	1	14,859	103	11.6	23	37	40	62
	2	17,716	103	11.6	25	36	39	61
	3	18,526	104	12.1	22	34	45	66
	4	18,838	101	13.8	33	30	37	55
	5	17,472	100	14.6	36	28	36	51
	6	20,998	99	14.4	39	29	33	48
	7	18,248	98	13.7	40	32	28	45
	8	15,819	98	15.3	41	27	33	47
	9	3,744	103	15.5	31	26	43	59
	10	906	105	16.4	26	25	49	62
	11	926	106	14.8	23	25	52	68

Precisely Assessing Low-Performing Students’ Ability

Using over 170,000 ELA scores and nearly 80,000 Math scores from empirical GM 4.0 data from the 2021/2022 school year, Figures 8.1 and 8.2 and Table 8.2 demonstrate how the GM score scale was psychometrically built to produce reliable assessments for students throughout the entire ability continuum, with CSEM less than 0.7 in all cases. From Table 8.2, we see for the LOSS/HOSS score values a very small (trivial) percentage at the lower/higher end on the GM score scale. This implies that the adaptive GM assessment has a robust item pool that can adapt to a large range of student abilities. As an example, for Math Grade 5, Figure 8.2 shows that the standard error of measurement (or, equivalently, the precision of the estimated ability) is largely uniformly distributed across a wide range of the ability spectrum. This result is a central feature of an adaptive assessment that targets the difficulty of the items administered to each student throughout the assessment.

FIGURE 8.1: CSEM Plot for Grade 5 GM ELA

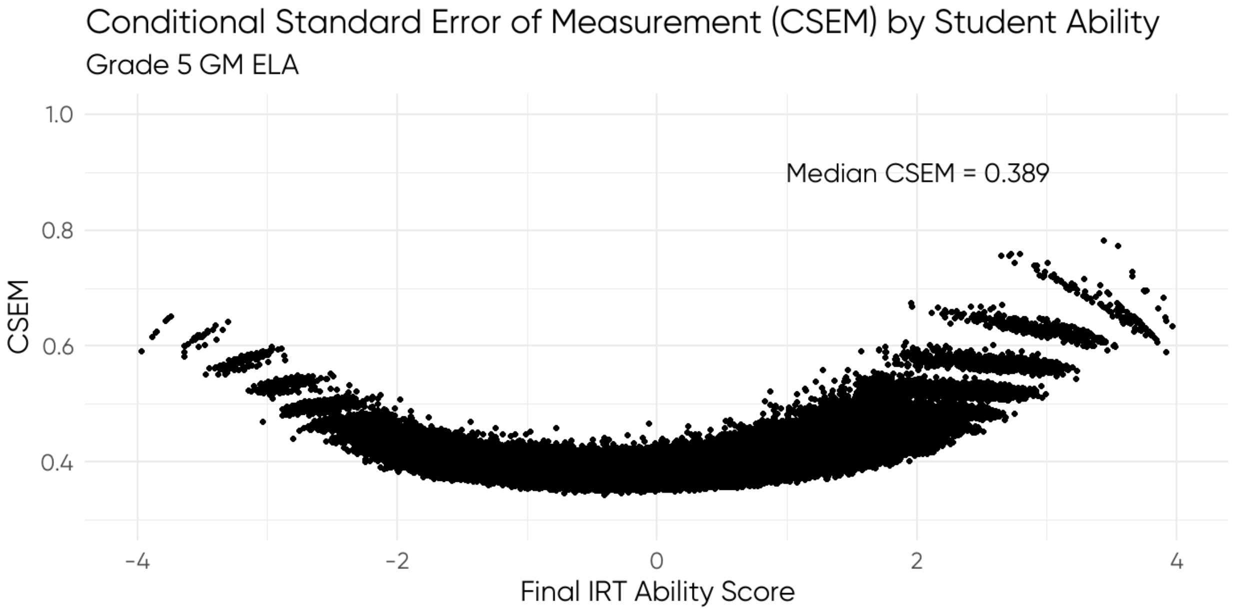


FIGURE 8.2: CSEM Plots for Grade 5 GM Math

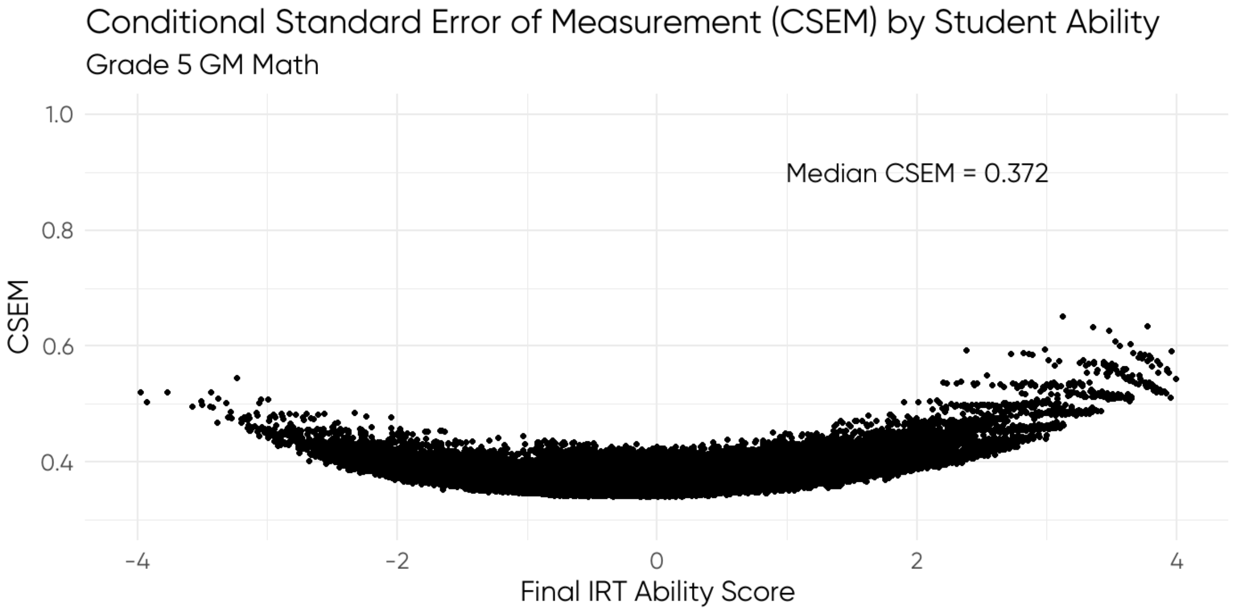


TABLE 8.3: Percentage of Students That Reach LOSS/HOSS on the GM Score Scale

	ELA			**Math**		
Test Grade	*Number of Scores*	*% LOSS*	*% HOSS*	*Number of Scores*	*% LOSS*	*% HOSS*
K	-	-	-	46,302	1	3
1	-	-	-	59,359	0	4
2	89,289	0	2	71,265	0	2
3	14,526	2	2	79,002	1	4
4	16,165	6	3	78,782	3	4
5	17,015	7	2	78,555	4	4
6	22,539	6	2	92,911	4	4
7	22,874	6	2	84,387	2	3
8	23,137	6	2	78,921	3	4
9	17,476	5	2	21,832	2	4
10	14,322	5	2	8,706	4	4
11	16,669	6	2	7,555	0	2

Establishing Reasonable Performance Categories

To provide instructors with more detailed feedback on students' performance, the number of performance categories on the GM 4.0 was increased from 3 to 5 levels, namely Far Below Level, Below Level, Approaching, On Level, and Above Level. Tables 8.4 and **??** show the percentages of students in each performance levels in the empirical GM 4.0 data for ELA and Math, respectively. If the test is appropriately measuring growth from one time window to the next, then these percentages would be expected to shift from a lower category to a higher category as instruction was received across the year (for example, a student starts Below Level at BOY and moves to Above Level by EOY).

Evaluating the GM 4.0 data reveals that the data follow the expected pattern. The EOY test results show that the movement between categories occurs as expected, with fewer students in Below Level and more students in On Level and Above Level. By spring, On Level was the largest category for all grades.

TABLE 8.4: Percentages of Students by Performance Levels for ELA)

Subject	Season	Test Grade	Far Below Level	Below Level	Approaching	On Level	Above Level
ELA	BOY	2	13	38	25	19	5
		3	16	33	25	20	5
		4	20	32	23	19	7
		5	18	33	23	19	7
		6	18	31	24	20	8
		7	18	30	23	21	7
		8	19	31	23	20	7
		9	17	32	23	20	8
		10	16	31	23	21	9
		11	17	30	23	21	8
	MOY	2	8	26	25	30	12
		3	10	23	24	29	13
		4	14	26	23	25	12
		5	13	27	24	24	12
		6	16	28	23	23	11
		7	16	28	22	23	11
		8	15	28	23	23	11
		9	16	29	23	22	10
		10	16	29	23	22	10
		11	16	28	23	22	11
	EOY	2	4	20	22	33	20
		3	8	19	21	32	20
		4	12	24	21	26	17
		5	12	25	23	25	16
		6	14	27	22	23	14
		7	14	27	21	24	14
		8	13	28	22	24	13
		9	12	24	23	25	15
		10	15	29	22	22	12
		11	18	31	22	19	10

TABLE 8.5: Percentages of Students by Performance Levels for Math

Subject	Season	Test Grade	Far Below Level	Below Level	Approaching	On Level	Above Level
Math	BOY	K	-	-	81	15	4
		1	-	42	32	22	4
		2	9	44	25	17	5
		3	9	42	28	16	5
		4	16	36	25	18	6
		5	16	36	24	19	5
		6	18	33	24	20	6
		7	14	36	25	17	8
		8	13	37	25	18	7
		9	13	34	27	19	7
		10	16	29	24	24	8
		11	9	43	23	20	6
	MOY	K	-	-	60	30	10
		1	-	22	29	36	13
		2	5	27	26	30	12
		3	5	28	26	29	11
		4	10	30	24	23	13
		5	11	28	24	24	13
		6	14	29	23	23	12
		7	11	31	23	24	10
		8	12	30	23	23	12
		9	14	33	23	20	10
		10	15	27	21	24	13
		11	4	33	22	28	14
	EOY	K	-	-	38	37	25
		1	-	13	20	39	28
		2	2	16	22	35	25
		3	3	17	19	35	26
		4	8	21	20	26	26
		5	9	23	20	24	24
		6	11	25	21	23	20
		7	9	28	22	26	16

TABLE 8.5: Percentages of Students by Performance Levels for Math

Subject	Season	Test Grade	Far Below Level	Below Level	Approaching	On Level	Above Level
		8	10	27	20	24	19
		9	5	21	20	28	26
		10	8	17	20	28	26
		11	2	18	20	33	27

TABLE 8.6: Percentages of Students by Performance Levels for Math (cont.)

Subject	Season	Test Grade	Far Below Level	Below Level	Approaching	On Level	Above Level
Math	BOY	K	-	-	81	15	4
		1	-	42	32	22	4
		2	9	44	25	17	5
		3	9	42	28	16	5
		4	16	36	25	18	6
		5	16	36	24	19	5
		6	18	33	24	20	6
		7	14	36	25	17	8
		8	13	37	25	18	7
		9	13	34	27	19	7
		10	16	29	24	24	8
		11	9	43	23	20	6
	MOY	K	-	-	60	30	10
		1	-	22	29	36	13
		2	5	27	26	30	12
		3	5	28	26	29	11
		4	10	30	24	23	13
		5	11	28	24	24	13
		6	14	29	23	23	12
		7	11	31	23	24	10
		8	12	30	23	23	12
		9	14	33	23	20	10
		10	15	27	21	24	13
		11	4	33	22	28	14
	EOY	K	-	-	38	37	25
		1	-	13	20	39	28
		2	2	16	22	35	25
		3	3	17	19	35	26
		4	8	21	20	26	26
		5	9	23	20	24	24
		6	11	25	21	23	20
		7	9	28	22	26	16

TABLE 8.6: Percentages of Students by Performance Levels for Math (cont.)

Subject	Season	Test Grade	Far Below Level	Below Level	Approaching	On Level	Above Level
		8	10	27	20	24	19
		9	5	21	20	28	26
		10	8	17	20	28	26
		11	2	18	20	33	27

Interpretation of Growth and Placement Metrics

The SPI is a criterion–referenced growth model intended to characterize the amount of growth a student has achieved throughout the school year by referencing each student's estimated true gain score to a targeted gain score between administrations (e.g., fall to spring) that has meaningful connections to grade–level range content expectations. Figures 8.3 and 8.4 show SPI distributions for an example grade in Elementary School (Grade 4), Middle School (Grade 7) and High School (Grade 10). Given that the established mean SPI is 100 across all grades, the SPI distributions for ELA Grades 4, 7, and 10 present a normal to slight positively skewed distribution with a median SPI of 99, 96, and 96, respectively. The growth levels are also as expected. For example, for ELA Grade 4 the percentage of low growth (SPI < 95) is 37%; the percentage of typical growth ($95 \leq \text{SPI} \leq 105$) is 34%; and the percentage of high growth (SPI > 105) is 30%.

The SPI distributions for Math follow a similar pattern as for ELA. For Math Grades 4 and 7, the SPI distribution is Normal to slight positively skewed, with the median SPI of 99 for both grades; Math Grade 10 has a slight negatively skewed distribution, with a median SPI of 103. Also similarly to ELA, the distribution of students among growth levels is as expected. For example, for Math Grade 4, the percentage of low growth (SPI < 95) is 37%; the percentage of typical growth ($95 \leq \text{SPI} \leq 105$) is 33%; and the percentage of high growth (SPI > 105) is 31%.

The SPI distribution across grades and content areas is consistent with the expectation of the student distribution in terms of growth. That is, the percentages of low- and targeted–growth students are both higher than the percentage of high–growth students. These numbers reflect that the SPI model used within the GM provided a valid indicator for measuring growth.

As a criterion–referenced metric, the SPI growth target of 100 is the same for all students across subjects and grades.[1] This advantage of the SPI in comparison with other growth models and/or gain scores can principally help school districts and parents interpret student growth more meaningfully.

[1] However, the SPI model rests on empirical data, so the scores that derive the targeted 100 SPI can potentially be updated to reflect better and/or more–current data.

FIGURE 8.3: Histogram of SPI Values for ELA

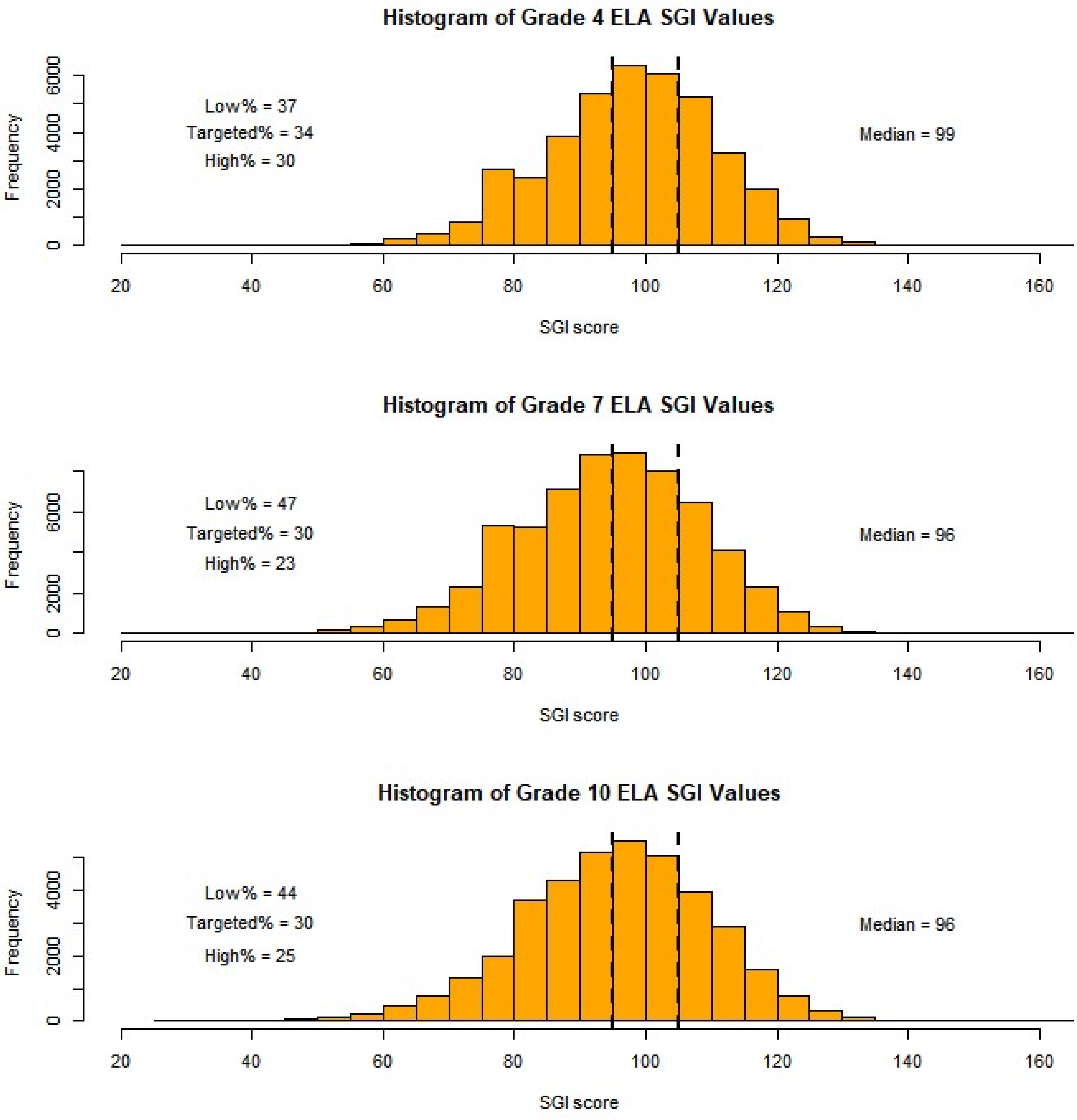
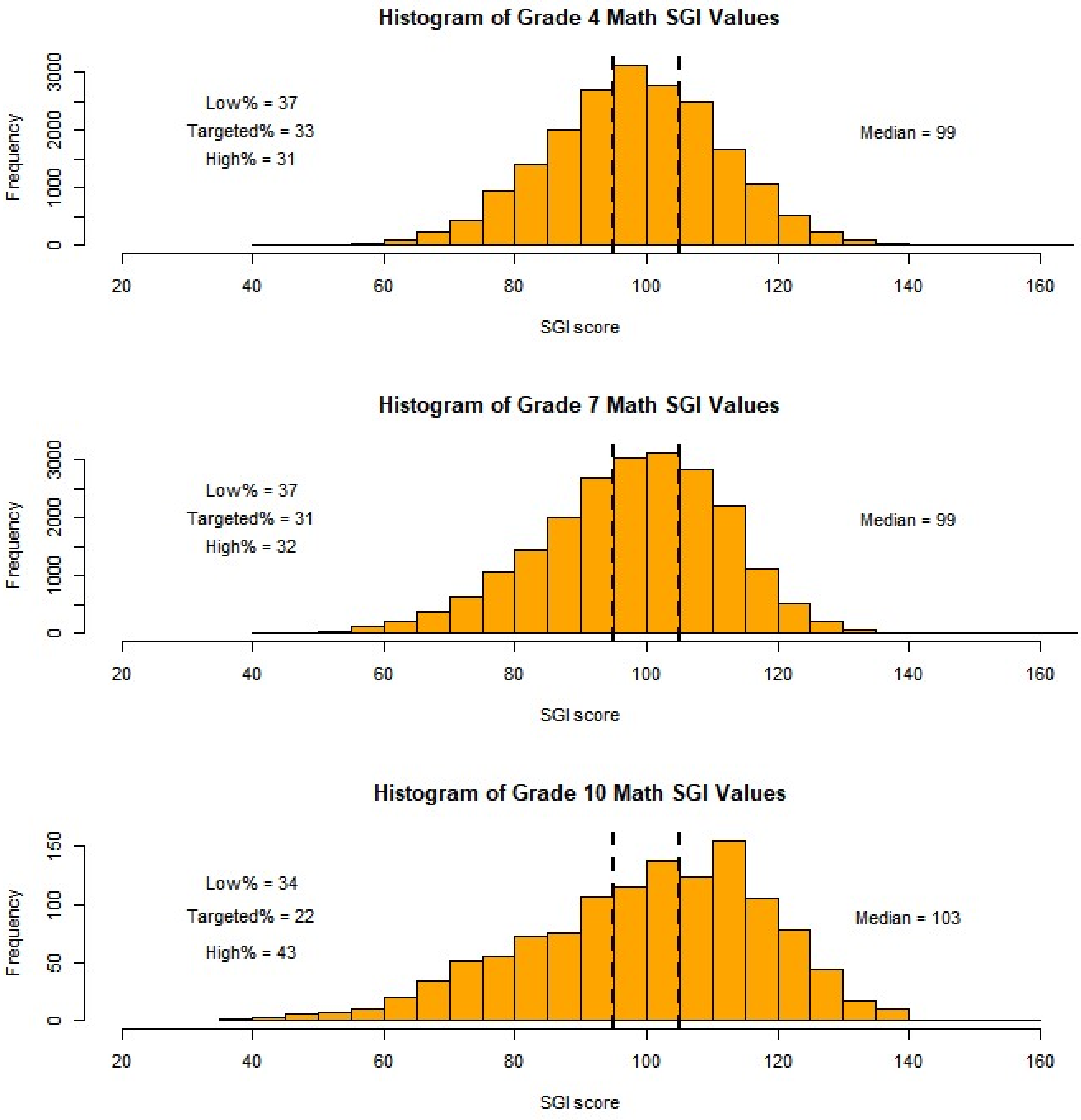


FIGURE 8.4: Histogram of SPI Values for Math



Histogram of Grade 4 Math SGI Values

Histogram of Grade 7 Math SGI Values

Histogram of Grade 10 Math SGI Values

SPI versus SGPs

The SPI was developed with the intention of having a similar conceptual framework as SGPs. The three main similarities between these growth scores are the following:

1. Both scores use a single "targeted" value to denote typical growth for a group of students. SGP typically has this value as 50, whereas the SPI has this value as 100.
2. Both models define growth using the Test 2 score in relation to information about the Test 1 score. SGP relies on a statistical approach and applies a norm–reference perspective (e.g., how the student did in comparison to their cohort), whereas the SPI relies on a measurement approach and applies a criterion–reference perspective (e.g., what the student's rate of growth is in relation to a grade–level range of content standards).
3. Both scores' utilization of (1) targeted values to denote typical growth and (2) distinct approaches to modeling growth leads to a growth measure that is largely devoid of correlation with initial status. This feature, unlike growth measures obtained via subtraction (e.g., Test 2 – Test 1 = gain score), allows student growth to be discussed in a simplified manner.

While there are these similarities, the SPI departs and improves upon the use of SGPs in the following ways: (1) SPI–derived values remain intact on the actual score scale from which the growth score was calculated. Thus, the interpretation of growth can be tied directly to performance progress on the content standards assessed within the item pool. (2) Moreover, since SPIs are criterion–referenced, interpretations do not depend on how other students perform but only on the student compared to the score scale and grade–level range content expectations.

To better evaluate the SPIs relative to SGPs, corresponding SGPs were calculated from the same Grade 4 Math GM 2.0 data used to calculate the SPIs. The notion of the SPI maintaining its connection to the score (unlike SGPs) is illustrated in Figure 8.5, which plots the relationship between the SGPs and SPIs. SGPs demonstrates a noticeable LOSS effect (SGP = 1) and HOSS effect (SGP = 99). The SPI is not affected by LOSS or HOSS, as the SPI scale has a larger range. For example, for students who received an SGP of 1, there is a range of SPI scores supported (70 to 92), and the same occurs for SGP 99 (SPI range is 117 to 135). Even with this truncation of SGPs, there is a high correlation between the SPI and SGP growth scores (r = 0.93), which lends strong concurrent validity to the SPI approach for calculating growth.

Because of the strong relationship seen with the SPI and SGP values, we can use the SPI to predict an SGP. For example, the linear prediction equation for the Grade 4 Math GM 2.0 data is

$$SGP = (SPI * 2.79) – 225$$

This means that an

- SPI of 105 (top of targeted range) predicts an SGP of 68
- SPI of 100 (targeted growth value) predicts an SGP of 54
- SPI of 95 (bottom of targeted range) predicts an SGP of 40

In practice, this implies SPIs above 105 ("high growth") are associated with SGPs greater than 68 and SPIs below 95 ("low growth") are associated with SGPs lower than 40. This prediction analysis shows that the SPI targeted growth ranges are valid, as they line up almost perfectly with commonly accepted SGP interpretations. Figure 8.6 clearly shows that if a student had an SGP of 70 or greater, then the average of those meeting an SPI of 100 or greater was essentially 100%. It also reveals that if a student had an SGP of 35 or lower, the average of those meeting an SPI of 100 was essentially 0%.

To further evaluate the SPIs, the scatterplot between the SPIs and initial GM scaled score from the same Grade 4 Math GM 2.0 data is shown in Figure 8.7. It illustrates the SPI is not seriously confounded by the initial test score. This means that all students can show growth without suggesting some students can make systemically more gains than others, which often happens with a more simplistic gain score (subtraction) methodology.

To sum up, the SPI scores

- are easily interpreted across grades and have meaning relative to grade–level content standards

FIGURE 8.5: Relationship between Grade 4 Math SPI and SGP Values

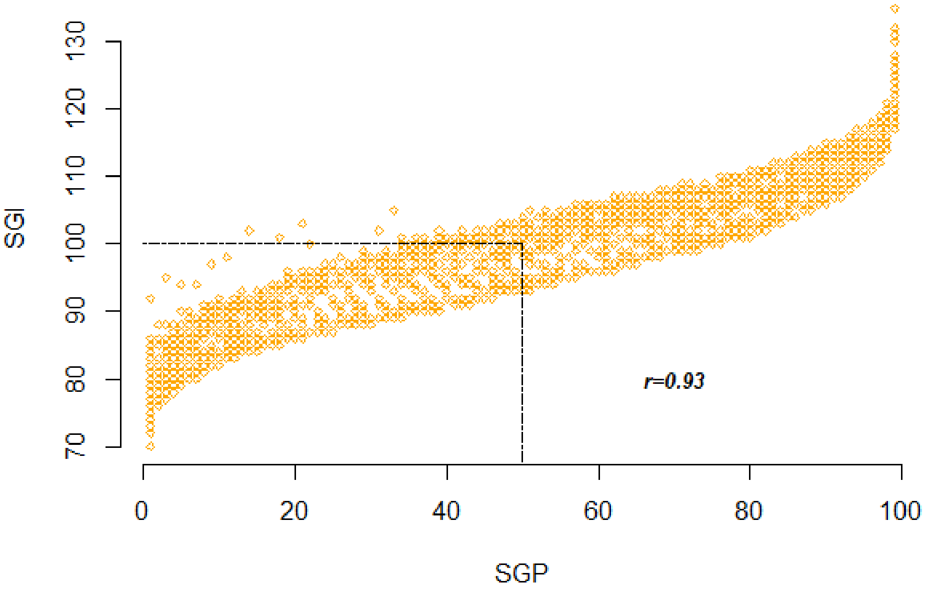


FIGURE 8.6: Relationship between Percent Meeting Grade 4 Math SPI Target and SGP Value

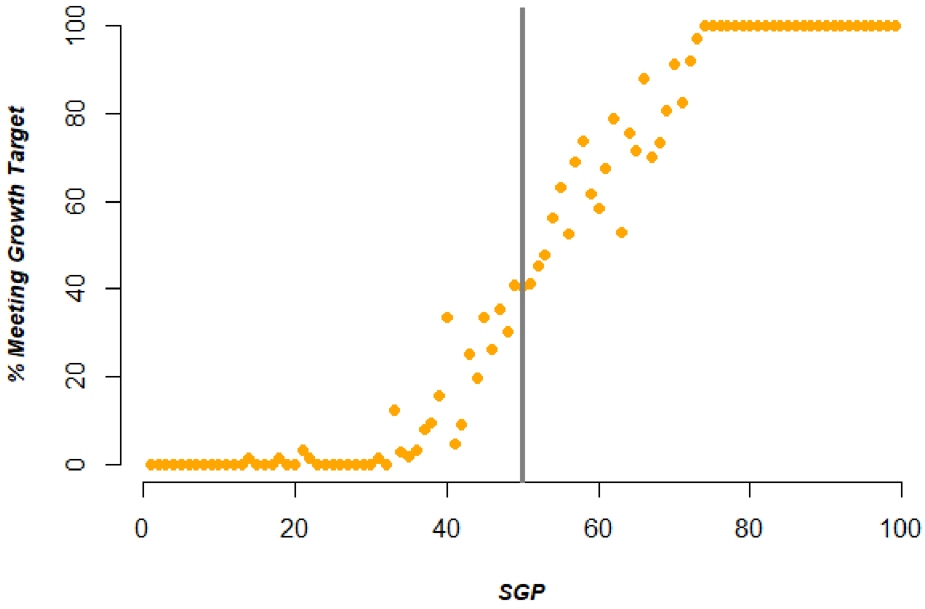
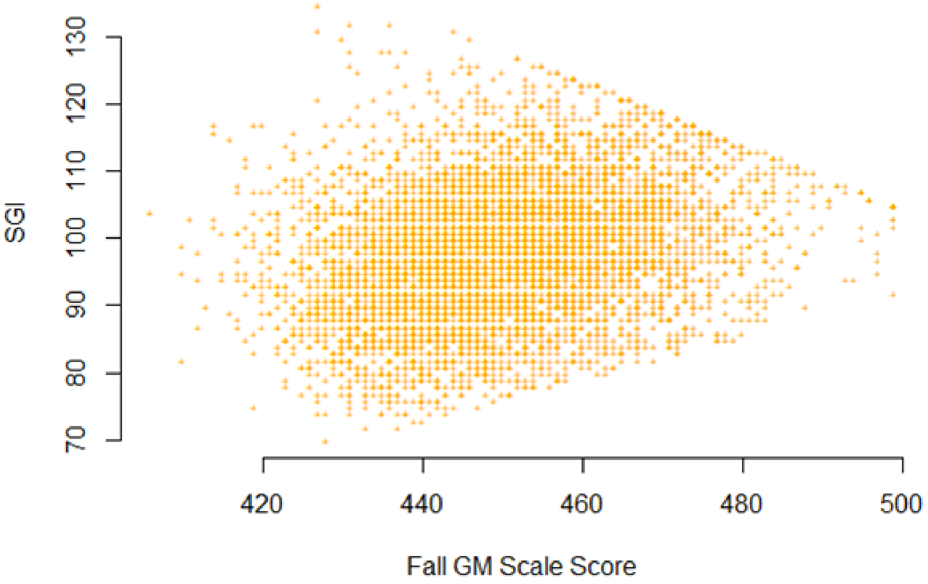


FIGURE 8.7: Relationship between Grade 4 Math SPI and Initial Test Score



- measure absolute growth as opposed to relative growth, and therefore are independent of how other students perform
- support reporting growth of a wider range than the SGP scale can support
- are more sensitive to growth of students at the low– and high–achieving ends

8.4. Summary

The GM assessments are standards–based interim assessments. Their purpose is to provide ongoing information to help the cause of students learning. The results of the GM are intended to guide decisions in the area of improving student achievement in reading and mathematics. Other information, along with the results of the assessments, can help teachers and school and district leaders make decisions that will have a positive impact on individual achievement and will accelerate growth. Validity and reliability evidence provided in this technical overview lends support to the use of the GM for instructional improvement. For example, each of the special validity studies supports use of the GM for decisions related to instruction: (1) The special study on the SPI shows that GM has the ability to appropriately reflect growth. (2) The special study on the CSEM shows that the GM is able to produce precise estimates of student ability at low, middle, and high ends of the scale. (3) The special study on performance levels shows that substantive performance levels can be described by the GM. (4) The information provided on growth and placement metrics in the special studies section facilitates the validity of use of the GM by supporting appropriate interpretations of GM metrics.

Validity is not an all–or–nothing property of a test; rather, validity evidence must be documented for a specific purpose and in the context of how the test scores will be interpreted and used. Much of the information contained in this technical overview is, in and of itself, documentation of the validity of the GM tests for their stated purpose. This chapter provides a summary of the evidence presented elsewhere in the technical overview and provides some additional types of validity evidence relevant to content and the interpretation of scores.

9. REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/https://doi.org/10.1007/BF02310555

Divgi, D. (1980). *Dimensionality of binary items: Use of a mixed model.*

Feldt, L., & Brennan, R. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (p. 105). The American Council on Education; National Council on Measurement in Education.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347–360.

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128–138.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 16–22. https://doi.org/https://doi.org/10.1111/j.1745-3992.2002.tb00081.x

Haertel, E. H. (2008). Standard setting. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (Vol. 7, pp. 139–154). Routledge.

Haertel, E. H., Beimers, J. N., & Miles, J. A. (2012). The briefing book method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (Second Edition, Vol. 14, pp. 283–300). Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *A look at psychometrics in the netherlands.*

Kettler, R. J. (2012). Testing accommodations: Theory and research to inform practice. *International Journal of Disability, Development and Education*, *59*(1), 53–66.

Kim, D.-H., Schneider, C., & Siskind, T. (2009). Examining the underlying factor structure of a statewide science test under oral and standard administrations. *Journal of Psychoeducational Assessment*, *27*(4), 323–333.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160. https://doi.org/http://dx.doi.org/10.1007/BF02288391

Linacre, J. M., & Wright, B. D. (2000). *Winsteps*. http://www.winsteps.com/index.htm

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher*, *31*(6), 3–16.

Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series*, *1956*(1), i–22. https://doi.org/https://doi.org/10.1002/j.2333-8504.1956.tb00058.x

Lord, F. M. (1962). Test reliability—a correction. *Educational and Psychological Measurement*, *22*(3), 511–512. https://doi.org/10.1177/001316446202200308

Miles, J. A., Beimers, J. N., & Way, W. D. (2010). *The modified briefing book standard setting method: Using validity data as a basis for setting cut scores.* American educational research association (AERA) and the national council on measurement in education (NCME). http://images.pearsonassessments.com/images/tmrs/tmrs_rg/Miles_themodifiedbriefingbookstandardsettingmethod_AERA_NCME2010.pdf

O'Malley, K., Miles, J. A., & Keng, L. (2012). From Z to A: Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (Second Edition, Vol. 14, pp. 283–300). Routledge.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*

Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test analysis modules*. https://CRAN.R-project.org/package=TAM

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments. Synthesis report.*

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.