TECHNICAL REPORT

for the
2010 Modified Pennsylvania System of School Assessment

Provided by
Data Recognition Corporation

Glossary of Common Terms

The following table contains some terms used in this technical report and their meanings. Some of these terms are used universally in the assessment community, and some of these terms are used commonly by psychometric professionals. A glossary of accommodation terms as applied to the PSSA is provided in Chapter Ten.

Table G–1. Glossary of Terms

Term	Common Definition
Ability	In the context of scaling, a latent-trait characteristic indicating the level of an individual on a particular construct or competence in a particular area. Following Rasch literature, ability is used as a generic term for the construct that is being measured by test. Competence, achievement, learning and status are alternative terms that are sometimes used, but all are subject to some degree of misinterpretation.
Adjacent Agreement	A score/rating difference of one (1) point in value usually assigned by two different raters under the same conditions (e.g., two independent raters give the same paper scores that differ by one point).
Alternate Forms	Two or more versions of a test that are considered exchangeable, i.e., they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. More specific terminology applies depending on the degree of statistical similarity between the test forms (e.g., parallel forms, equivalent forms, and comparable forms) where parallel forms refers to the situation in which the test forms have the highest degree of similarity to each other.
Average	A measure of central tendency in a score distribution that usually refers to the arithmetic mean of a set of scores. In this case, it is determined by adding all the scores in a distribution and then dividing the obtained value by the total number of scores. Sometimes people use the word average to refer to other measures of central tendency such as the median (the score in the middle of a distribution) or mode (the score value with the greatest frequency).
Bias	In a statistical context, bias refers to any source of systematic error in the measurement of a test score. In discussing test fairness, bias may refer to construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers (e.g., gender, ethnicity). Attempts are made to reduce bias by conducting item fairness reviews and various differential item functioning (DIF) analyses, detecting potential areas of concern, and either removing or revising the flagged test items prior to the development of the final operational form of the test. Also see Differential Item Functioning.
Constructed-Response Item	See open-ended item.
Content Validity Evidence	Evidence regarding the extent to which a test provides an appropriate sampling of a content domain of interest (e.g., assessable portions of a state's Grade 6 mathematics curriculum in terms of the knowledge, skills, objectives, and processes sampled).

Term	Common Definition
Core-Linking Item	Items that are utilized during the linking process (see Linking). They are a subset of the PSSA-M operational items and so they: 1) are the same on all test forms for any grade/subject area test, and 2) contribute to student total raw scores and scaled scores.
Criterion-Referenced Interpretation	When a score is interpreted as a measure of a student's performance as with respect to an expected level of mastery, educational objective, or standard. The types of resulting score interpretations provide information about what a student knows or can do with respect to a given content area.
Cut Score	A specified point on a score scale such that scores at or above that point are interpreted or acted upon differently from scores below that point. For example, a score designated as the minimum level of performance needed to pass a competency test. One or more cut scores can be set for a test that results in dividing the score range into various proficiency level ranges. Methods for establishing cut scores vary. See Performance Level Setting.
Decision Consistency	The extent to which classifications based on test scores would match the decisions based on scores from a second, parallel form, of the same test. It is often expressed as the proportion of examinees that are classified the same way from the two test administrations.
Differential Item Functioning (DIF)	A statistical property of a test item in which different groups of test takers (who have the same total test score) have different average item scores or, in some cases, different rates of choosing various item options. Also see Bias.
Distractor	An incorrect option in a multiple-choice item (also called a foil).
Equating	The strongest of several linking methods used to establish comparability between scores from multiple tests. Equated test scores should be considered exchangeable. Consequently, the criteria needed to refer to a linkage as equating are strong and somewhat complex (equal construct and precision, equity, and invariance). In practical terms, it is often stated that it should be a matter of indifference to a student if he/she takes any of the equated tests. See also Linking.
Equating Block (EB) Items	The PSSA-M uses multiple test forms for each grade/subject area test. Each form is composed of operational (OP) items, equating block (EB) items, and field test (FT) items. EB items are utilized during the linking process (see Linking). Each test form includes a set of EB items. EB items are not part of any student scores.
Error of Measurement	The amount by which the score actually received (an observed score) differs from a hypothetical true score. Also see Standard Error of Measurement.
Exact Agreement	When identical scores/ratings are assigned by two different raters under the same conditions (e.g., two independent raters give a paper the same score).

Term	Common Definition
Field Test (FT) Items	The PSSA-M uses multiple test forms for each grade/subject area test. Each form is composed of operational (OP) items, equating block (EB) items, and field test (FT) items. An FT item is a newly-developed item that is ready to be tried out to determine its statistical properties (e.g., see *P*-value and Point-Biserial Correlation). Each test form includes a set of FT items. FT items are not part of any student scores.
Frequency	The number of times that a certain value or range of values (score interval) occurs in a distribution of scores.
Frequency Distribution	A tabulation of scores from low to high or high to low showing the number and/or percent of individuals who obtain each score or who fall within each score interval or category.
Infit/Outfit	Statistical indicators of the agreement of the data and the measurement model. See also Outfit/Infit.
Item Difficulty	For the Rasch model, the dichotomous item difficulty represents the point along the latent trait continuum where an examinee has a 0.50 probability of making a correct response. For a polytomous item, the difficulty is the average of the item's step difficulties (see Step Difficulty).
Key	The correct response option for a multiple-choice item.
Linking	A generic term referring to one of a number of processes by which scores from one or more tests are made comparable to some degree. Linking includes several classes of transformations (equating, scale alignment, prediction, etc.). Equating is associated with the strongest degree of comparability (exchangeable scores). Other linkages may be very strong, but fail to meet one or more of the strict criteria required of equating. Also see Equating.
Logit	The fundamental unit of measurement in the Rasch model used to express both item difficulties and person locations. When expressing person locations, logits are invariably transformed into Scale Scores through a simple linear transformation before reporting (also see Scaled Score). When expressing item difficulties, logits are transformed *p*-value (also see *P*-value). The logit difficulty scale is inversely related to *p*-values. A higher logit value would represent a relatively harder item, while a lower logit value would represent a relatively easier item.
Mean	Also referred to as the arithmetic mean of a set of scores, mean is found by adding all the score values in a distribution and dividing by the total number of scores. For example, the mean of the set {66, 76, 85, and 97} is 81. The value of a mean can be influenced by extreme values in a score distribution.
Measure	A Rasch estimate (or calibration) for a parameter, i.e., a person ability-parameter estimate, or an item difficulty-parameter estimate.

Term	Common Definition
Median	The middle point or score in a set of rank-ordered observations that divides the distribution into two equal parts such that each part contains 50 percent of the total data set. More simply put, half of the scores are below the median value and half of the scores are above the median value. As an example, the median for the following ranked set of scores {2, 3, 6, 8, 9} is 6.
Multiple-Choice Item (MC)	A type of item format that requires the test taker to select a response from a group of possible choices, one of which is the correct answer (or key) to the question posed. Also see Open-Ended Item.
N-count	Sometimes designated as N or n, it is the number observations (usually individuals or students) in a particular group. Some examples include: the number of students tested, the number of students tested from a specific subpopulation (e.g., females), the number of students who attained a specific score, etc. In the following set {23, 32, 56, 65, 78, 87}, n = 6.
Open-ended Item (OE)	An open-ended (OE) item—referred to by some as a constructed-response (CR) item—is an item format that requires examinees to create their own responses, which can be expressed in various forms, (e.g., written paragraph, created table/graph, formulated calculation). Such items are frequently scored using more than two score categories, that is, polytomously (e.g., 0, 1, 2, and 3). This format is in contrast to when students make a choice from a supplied set of answers options (e.g., multiple-choice items (MC) which are typically dichotomously scored as right = 1 or wrong = 0). When interpreting item difficulty and discrimination indices it is important to consider whether an item is polytomously or dichotomously scored.
Operational Item	The PSSA-M uses multiple test forms for each grade/subject area test. Each form is composed of operational (OP) items, equating block (EB) items, and field test (FT) items. OP items are the same on all forms for any grade/subject area test. Student total raw scores and scaled scores are based exclusively on the OP items.
Outfit/Infit	Statistical indicators of the agreement of the data and the measurement model. Infit and Outfit are highly correlated, and both are highly correlated with the point-biserial correlation. Underfit can be caused when low-ability students correctly answer difficult items (perhaps by guessing or atypical experience) or high-ability students incorrectly answer easy items (perhaps because of carelessness or gaps in instruction). Any model expects some level of variability, so overfit can occur when nearly all low-ability students miss an item while nearly all high-ability students get the item correct.
Percent Correct	When referring to an individual item, the percent correct is the item's *p*-value expressed as a percent (instead of a proportion). When referring to a total test score, it is the percentage of the total number of points that a student received. The percent correct score is obtained by dividing the student's raw score by the total number of points possible and multiplying the result by 100. Percent Correct scores are often used in criterion-referenced interpretations and are generally more helpful if the overall difficulty of a test is known. Sometimes Percent Correct scores are incorrectly interpreted as Percentile Ranks.

Term	Common Definition
Percentile	The score or point in a score distribution at or below which a given percentage of scores fall. It should be emphasized that it is a value on the score scale, not the associated percentage (although sometimes in casual usage this misinterpretation is made). For example, if 72 percent of the students score at or below a Scaled Score of 1500 on a given test, then the Scaled Score of 1500 would be considered the 72nd percentile. As another example, the median is the 50th percentile.
Percentile Rank	The percentage of scores in a specified distribution falling at/below a certain point on a score distribution. Percentile Ranks range in value from 1 to 99, and indicate the status or relative standing of an individual within a specified group, by indicating the percent of individuals in that group who obtained equal or lower scores. An individual's percentile rank can vary depending on which group is used to determine the ranking. As suggested above, Percentile and Percentile Rank are sometimes used interchangeably; however strictly speaking, a percentile is a value on the score scale.
Performance Level Descriptors	Descriptions of an individual's competency in a particular content area, usually defined as ordered categories on a continuum, often labeled from Below Basic-M to Advanced-M, that constitute broad ranges for classifying performance. The exact labeling of these categories, and narrative descriptions, may vary from one assessment or testing program to another.
Performance Level Setting	Also referred to as standard setting, a procedure used in the determination of the cut scores for a given assessment that is used to measure students' progress towards certain performance standards. Standard setting methods vary (e.g., modified Angoff, Bookmark Method, etc.), but most use a panel of educators and expert judgments to operationalize the level of achievement students must demonstrate in order to be categorized within each performance level.
Point-Biserial Correlation	In classical test theory this is an item discrimination index. It is the correlation between a dichotomously scored item and a continuous criterion, usually represented by the total test score (or the corrected total test score with the reference item removed). It reflects the extent to which an item differentiates between high-scoring and low-scoring examinees. This discrimination index ranges from –1.00 to +1.00. The higher the discrimination index (the closer to +1.00), the better the item is considered to be performing. For multiple-choice items scored as 0 or 1, it is rare for the value of this index to exceed 0.5
P-value	An index indicating an item's difficulty for some specified group (perhaps grade). It is calculated as the proportion (sometimes percent) of students in the group who answer an item correctly. *P*-values range from 0.0 to 1.0 on the proportion scale. Lower values correspond to more difficult items and higher values correspond to easier items. *P*-values are usually provided for multiple-choice items or other items worth one point. For open-ended items or items worth more than one point, difficulty on a *p*-value-like scale can be estimated by dividing the item mean score by the maximum number of points possible for the item. Also see Logit.

Term	Common Definition
Raw Score	Sometimes abbreviated as RS—it is an unadjusted score usually determined by tallying the number of questions answered correctly, or by the sum of item scores (i.e., points). (Some rarer situations might include formula-scoring, the amount of time required to perform a task, the number of errors, application of basal/ceiling rules, etc.). Raw scores typically have little or no meaning by themselves and require additional information—like the number of items on the test, the difficulty of the test items, norm-referenced information, or criterion-referenced information.
Reliability	The expected degree to which test scores for a group of examinees are consistent over exchangeable replications of an assessment procedure, and therefore, considered dependable and repeatable for an individual examinee. A test that produces highly consistent, stable results (i.e., relatively free from random error) is said to be highly reliable. The reliability of a test is typically expressed as a reliability coefficient or by the standard error of measurement derived by that coefficient.
Reliability Coefficient	A statistical index that reflects the degree to which scores are free from random measurement error. Theoretically, it expresses the consistency of test scores as the ratio of true score variance to total score variance (true score variance plus error variance). This statistic is often expressed as correlation coefficient (e.g., correlation between two forms of a test) or with an index that resembles a correlation coefficient (e.g., calculation of a test's internal consistency using Coefficient Alpha). Expressed this way, the reliability coefficient is a unitless index. The higher the value of the index (closer to 1.0), the greater the reliability of the test. Also see Standard Error of Measurement.
Scaled Score	A mathematical transformation of a raw score developed through a process called scaling. Scaled scores are most useful when comparing test results over time. Several different methods of scaling exist, but each is intended to provide a continuous and meaningful score scale across different forms of a test.
Selected-Response Item	See multiple-choice item.
Spiraling	A packaging process used when multiple forms of a test exist and it is desired that each form be tested in all classrooms or other grouping unit (e.g., schools) participating in the testing process. This process allows for the random distribution of test booklets to students. For example, if a package has four test forms labeled A, B, C, & D, the order of the test booklets in the package would be: A, B, C, D, A, B, C, D, A, B, C, D, etc.

Term	Common Definition
Standard Deviation (SD)	A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statistic is always greater than or equal to zero. If all of the scores in a distribution are identical, the standard deviation is equal to zero. The further the scores are away from each other in value, the greater the standard deviation. This statistic is calculated using the information about the deviations (distances) between each score and the distribution's mean. It is equivalent to the square root of the variance statistic. The standard deviation is a commonly used method of examining a distribution's variability since the standard deviation is expressed in the same units as the data.
Standard Error of Measurement (SEM)	Abbreviated SEM, it is the amount an observed score is expected to fluctuate around the true score. As an example, across replications of a measurement procedure, the true score will not differ by more than plus or minus one standard error from the observed score about 68 percent of the time (assuming normally distributed errors). The SEM is frequently used to obtain an idea of the consistency of a person's score in actual score units, or to set a confidence band around a score in terms of the error of measurement. Often a single SEM value is calculated for all test scores. On other occasions, however, the value of the SEM can vary along a score scale. Conditional standard errors of measurement (CSEMs) provide an SEM for each possible scaled score.
Step Difficulty	Step difficulty is a parameter estimate in Master's partial credit model (PCM) that represents the relative difficulty of each score step (e.g., going from a score of 1 to a score of 2). The higher the value of a particular step difficulty, the more difficult a particular step is relative to other score steps (e.g., is it harder to go from a 1 to a 2, or to go from a 2 to a 3).
Strand	On score reports, a strand often refers to a set of items on a test measuring the same contextual area (e.g., Number Sense in Mathematics). Items developed to measure the same reporting category would be used to determine the strand score (sometimes called subscale score).
Technical Advisory Committee (TAC)	A group of individuals, most often professionals in the field of testing, that are either appointed or selected to make recommendations for and to guide the technical development of a given testing program.
Validity	The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by the purposed uses of a test. There are various ways of gathering validity evidence.

Preface: An Overview of Modified Assessments from 2008 to the Present

The Pennsylvania System of School Assessment with Modified Academic Achievement Standards (PSSA-M) is a statewide system designed to meet the *No Child Left Behind Act of 2001* (NCLB) requirement that all students be included in state assessment and accountability systems. The target population consists of those students who function above the one percent of students with the most severe cognitive impairments who are eligible to take the Pennsylvania Alternate System of Assessment (PASA), but whose disabilities inhibit their ability to respond to the standard PSSA, even with accommodations. The Pennsylvania Academic Assessment Anchor Content Standards, further delineated by the Eligible Content for Mathematics, Reading and Science, is the basis for test development. To facilitate students' ability to demonstrate their grade-level content knowledge and skills, revisions were made to assessment tasks, (e.g., items, passages, graphics/stimuli, scenarios) with the goal of minimizing or removing processing effects (e.g., cognitive, linguistic) or physical challenges related to students' disabilities without significant alteration of the assessed construct.

The introduction of an operational mathematics modified assessment in 2010 moved closer to reality with a major standalone field test at Grades 4–8, and 11 in May of 2009. Operational modified assessments for reading and science, scheduled for implementation in spring 2011, underwent item development in 2009 and field testing in 2010.

To assist the reader in navigating through the year-to-year developmental activity of the PSSA-M, tables are presented along with explanatory text. Provided is an overview of the subject areas assessed, time of year the testing activity took place, and the type of testing that occurred (e.g., operational, field testing, Grade 12 retest).

ASSESSMENT ACTIVITIES OCCURRING IN THE 2008–09 SCHOOL YEAR

Table P–1 shows the plan for the field testing of modified assessments for mathematics during the 2008–09 school year. Following the spring operational assessment of the PSSA, a separate, standalone field test of items developed for Pennsylvania Assessment Anchors and Eligible Content in mathematics was conducted at Grades 4 through 8, and Grade 11. Item development for these new assessments took place during 2008.

Table P–1. Field Testing of Modified Assessments During the 2008–09 School Year

Grade	Assessment Activity	Date
4	Standalone field test in mathematics modified	May 2009
5	Standalone field test in mathematics modified	May 2009
6	Standalone field test in mathematics modified	May 2009
7	Standalone field test in mathematics modified	May 2009
8	Standalone field test in mathematics modified	May 2009
11	Standalone field test in mathematics modified	May 2009

ASSESSMENT ACTIVITIES OCCURRING IN THE 2009–10 SCHOOL YEAR

Table P–2 shows the assessment plan for modified assessments during the 2009–10 school year. The mathematics modified assessments became operational for Grades 4 through 8 and Grade 11 and were incorporated in the administration of the PSSA as a test version for eligible students with disabilities. There was an April test window with a make-up period extending through the first week of May for all assessments. Field testing for mathematics was embedded as part of the operational assessments at each grade level. As for the regular PSSA, a fall retest opportunity at Grade 12 will be implemented for students taking the mathematics modified assessment.

Standalone field tests in reading and science modified were conducted subsequent to administration of the spring PSSA. Item development for these new assessments took place during 2009.

Table P–2. Operational Assessment and Field Testing During the 2009–10 School Year

Grade	Assessment Activity	Date
4	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
5	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
6	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
7	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
8	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
	Standalone field test in science modified	May 2010
11	Operational mathematics modified with embedded field test	April/May 2010
	Standalone field test in reading modified	May 2010
	Standalone field test in science modified	May 2010

ASSESSMENT ACTIVITIES PLANNED FOR THE 2010–11 SCHOOL YEAR

Table P–3 shows the assessment plan for modified assessments during the 2010–11 school year. This will be the second year for which the mathematics modified assessment is operational and the first year of implementation for the reading and science modified. There will be no embedded field testing as part of the operational modified assessments.

The modified assessments will become operational for reading at Grades 4 through 8 and Grade 11, and for science at Grades 8 and 11. As for the regular PSSA, a fall retest opportunity at Grade 12 will be implemented for students taking the mathematics modified assessment. A retest opportunity will become available in the fall of 2011 for students failing to reach the proficient level on the reading and/or science modified assessments.

Table P–3. Operational Assessment and Field Testing During the 2010–11 School Year (Planned)

Grade	Assessment Activity	Date
4	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
5	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
6	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
7	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
8	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
	Operational science modified	May 2011
11	Operational mathematics modified	March 2011
	Operational reading modified	March 2011
	Operational science modified	May 2011
12	Retest opportunity for students who as Grade 11 students in the spring of 2010 failed to reach at least the Proficient level in mathematics modified.	October/ November 2010

Chapter One: Background of the Modified Pennsylvania System of School Assessment (PSSA-M)

This brief overview of a decade of change in Pennsylvania's assessment program summarizes the state and federal regulations that have continued to shape the design and development of the program. Among the changes are those involving content structure for reading, mathematics, and writing, the addition of science to the subject areas assessed, the expansion of grade levels assessed for reading and mathematics, the implementation of an alternate assessment for students with very severe disabilities, and the implementation of a modified assessment for a group of IEP students whose disabilities inhibit their ability to respond to a regular assessment.

STATE AND FEDERAL REGULATIONS AFFECTING THE PSSA

The Pennsylvania System of School Assessment (PSSA) program underwent major structural changes in test content with the State Board of Education's adoption of the Pennsylvania Academic Standards for Reading, Writing, Speaking and Listening, and Mathematics in January 1999 (Pennsylvania State Board of Education, 1999). The Academic Standards, which are part of *Chapter 4 Regulations on Academic Standards and Assessment*, detailed what students should know (knowledge) and be able to do (skills) at various grade levels. Subsequently, the State Board approved a set of criteria defining Advanced, Proficient, Basic, and Below Basic levels of performance. Reading and mathematics performance level results were reported at both the student and school levels for the 2000 PSSA. At that point, the PSSA became a standards-based, criterion-referenced assessment measuring student attainment of the Academic Standards at Grades 5, 8, and 11. In 2003, a reading and mathematics assessment at Grade 3 was added. Act 16 of Pennsylvania Senate Bill 652 in 2000, redefined the PSSA to include science and State Board adoption of *Science and Technology Standards* on July 12, 2001, and the *Environment and Ecology Standards* on January 5, 2002, thereby laying the groundwork for a future science assessment. At the federal level, PL 107–110, the *No Child Left Behind Act of 2001* (NCLB) stipulated that states must develop reading and mathematics assessments in Grades 3–8 and at least once betweens Grades 9 and 12.

PURPOSES OF THE PSSA

Chapter 4 regulations stipulated that the purposes of the PSSA are to:

- Provide students, parents, educators, and citizens with an understanding of student and school performance.
- Determine the degree to which programs enable students to attain proficiency of academic standards.
- Provide results to school districts, including charter schools, and Career and Technical Centers (CTCs) for consideration in the development of strategic plans.
- Provide information to state policymakers, including the General Assembly, and the State Board, on how effective schools are in promoting and demonstrating student proficiency of the Academic Standards.

- Provide information to the general public on school performance.
- Provide results to school districts, including charter schools, and CTCs based on the aggregate performance of all students and for relevant subgroups, such as students with an IEP and for those without an IEP.

CHANGES IN 2005 AND BEYOND

Assessment in 2005 was marked by implementation of *Assessment Anchor Content Standards*, developed for reading and mathematics during the previous school year to clarify content structure, improve articulation between assessment and instruction, and improve test design and reporting. To meet the conditions of NCLB, assessment of reading and mathematics at Grades 4, 6, and 7 became operational in 2006, enabling Pennsylvania to more completely determine adequate yearly progress (AYP) at the state, district, and school level.

Although NCLB does not require states to conduct a writing assessment, Chapter 4 does include one, aligned to the Academic Standards and reported in terms of performance levels, for all students at three grade levels. The 2006 PSSA operational writing assessment involved a shift from Grades 6, 9, and 11 to Grades 5, 8, and 11 to provide better alignment to the end of elementary school and middle school. Also incorporated were mode-specific scoring guides for essay responses and stimulus-based revising/editing multiple-choice items.

In accordance with the NCLB requirement to implement an operational science assessment in 2008, a major test development effort took place during 2006, followed by a large-scale, standalone field test in April/May of 2007. Full implementation of an operational science assessment at Grades 4, 8, and 11 first occurred in April–May 2008, aligned to the *Pennsylvania Science Assessment Anchor Content Standards* and Eligible Content.

More information regarding the 2010 PSSA may be found in the *2010 PSSA Technical Report for Reading, Mathematics, Science, and Writing*. This report can be accessed by going to www.education.state.pa.us. On the left, click on "Programs," then "Programs O–R," then "Pennsylvania System of School Assessment (PSSA)" and then "Resource Materials."

STUDENTS WITH COMPLEX SUPPORT NEEDS: ALTERNATE ASSESSMENT

Although NCLB recommended that the same achievement standards be applied to all students, the U.S. Department of Education acknowledged that the same assessments are not universally appropriate. To better accommodate students with significant cognitive disabilities, intended for the lowest functioning 1% of the student population, the Department issued regulations permitting states to develop alternate achievement standards along with aligned assessments. In 2004 the *Pennsylvania Alternate System of Assessment* (PASA) was implemented to address the needs of these students. To be eligible for participation in the PASA a student must meet each of the following criteria for reading, mathematics, and science, and a school administered alternate assessment for writing: 1) enrolled in the assessed grade level for the subject area, 2) had a very severe cognitive disability, 3) required very intensive instruction, 4) required very extensive adaptation and support to perform or participate meaningfully, 5) required very substantial modification of the general education curriculum, and 6) participation in the general education curriculum differed very substantially in form and substance from that of other students (see *The 2009–2010 PSSA Handbook for Assessment Coordinators: Reading and Mathematics, Writing, and Science*, PDE, 2010, p.9), which may be accessed by going to www.education.state.pa.us. On the left side of the navigation bar, click on "Programs," then "Programs O–R," then

"Pennsylvania System of School Assessment (PSSA)" and then "Testing Accommodations & Security."

STUDENTS WITH DISABILITIES NEEDING A MODIFIED APPROACH: MODIFIED ASSESSMENT

Following the issuance of regulations permitting states to develop alternate assessments for the students with the most severe cognitive disabilities, further research along with the experience of state assessment programs identified a need to address the difficulties encountered by a small group of IEP students in responding optimally to the regular assessment instruments. The U.S. Department of Education responded to this recognition by issuing additional regulations in April 2007 permitting states to develop assessments for the approximately 2% of students with disabilities based on modified achievement standards. Students targeted are those whose disabilities are not severe enough to warrant taking an alternate assessment and yet interfere significantly with their ability to respond optimally on the regular state assessment. This modified assessment must be aligned to a set of modified achievement standards designed to measure the same grade-level content as the state's general assessment. To be eligible to take a modified assessment a student must meet a rigorous set of criteria such as the IEP addressing educational goals reflecting grade-level content standards along with provisions for monitoring student progress.

To address the unique needs of these students, and be in closer compliance with the NCLB intent that all students be included in state assessment and accountability systems, the *Pennsylvania System of School Assessment Modified* (PSSA-M) became operational in 2010 with a mathematics modified assessment at Grades 4–8, 11. It will be joined by operational modified assessments in reading at Grades 4–8, 11 and science at Grades 8 and 11 in the spring of 2011.

More information regarding the development and composition of the 2010 PSSA-M Mathematics test may be found in Chapter Two of this report. Information may also be found in the Pennsylvania Department of Education publication, *2009–2010 PSSA Assessment Handbook,* (see *Part Six: PSSA—Modified*). This handbook can be accessed by going to www.education.state.pa.us. On the left, click on "Programs," then "Programs O–R," then "Pennsylvania System of School Assessment (PSSA)" and then "Resource Materials."

Eligibility for the PSSA-M requires that a student 1) is not eligible for the PASA, 2) has a grade-level standards aligned IEP that clearly documents that the student requires significant instructional accommodations to successfully access grade level content, 3) demonstrates persistent academic difficulties with 4) lacks academic progress. More detailed information on the PSSA-M eligibility criteria may be accessed by going to www.education.state.pa.us. On the left side of the navigation bar, click on "Programs," then "Programs S–Z," then "Special Education." From the "Special Education" page click on "Assessment" to access the relevant documents.

Chapter Two: Test Development Overview of the Modified PSSA

OVERVIEW OF THE DEVELOPMENT PROCESS

The modified mathematics assessments were developed under the direction of the Pennsylvania Department of Education (PDE). The PSSA-M assessments were developed using the same rigorous and technically sound development steps that are used to develop the general education assessment, Pennsylvania System of School Assessment (PSSA). These technically sound development steps involve Pennsylvania educators in all stages of the process. The Pennsylvania educators from school districts throughout the Commonwealth of Pennsylvania selected to participate in the development process were those with both content-area teaching expertise (e.g., mathematics, reading, and science) as well as those with expertise in teaching students with disabilities. The key development steps PDE followed when developing the PSSA-M assessments included the following:

- Developing guidelines for revising and/or enhancing assessment questions
- Interviewing students and surveying teachers
- Revising and/or enhancing items to be more accessible to the given population of students
- Reviewing items by committees of Pennsylvania educators, including reviewing items for content alignment; rigor alignment; adherence to the principles of universal design; bias, fairness, and sensitivity; and adherence to technical quality or the standards for high-quality items
- Developing field test forms
- Field testing the items to determine whether or not the items did, in fact, lend themselves to being more accessible to the given population
- Scoring the open-ended or constructed-response items
- Reviewing the items to determine which items should be placed in the pool of items to be considered acceptable for operational testing
- Reviewing the final operational forms prior to being administered to students
- Defining the expectation of mastery on the PSSA-M assessments or what it means for a student to be proficient as determined by the standard-setting process
- Developing Modified Achievement Standards

ACADEMIC STANDARDS, ASSESSMENT ANCHOR CONTENT STANDARDS, AND ELIGIBLE CONTENT

PSSA-M Mathematics

The PSSA-M assessment follows the guidelines of the PSSA Assessment Anchor Content Standards and Eligible Content, which are based on the Pennsylvania Academic Standards. Although the Academic Standards indicate what students should know and be able to do, educator concerns regarding the number and breadth of Academic Standards led to an initiative by the Pennsylvania Department of Education (PDE) to develop Assessment Anchor Content Standards (Assessment Anchors) to indicate which parts of the Academic Standards (Instructional Standards) would be assessed on the PSSA and PSSA-M. Based on recommendations from Pennsylvania educators, the Assessment Anchors were designed as a tool to improve the articulation of curricular, instructional, and assessment practices. The Assessment Anchors clarify what is expected across each grade span and focus the content of the standards into what is assessable on a large-scale test. The Assessment Anchor documents also serve to communicate eligible content, also called assessment limits, or the range of knowledge and skills from which the PSSA and PSSA-M would be designed.

The Assessment Anchor's code is in an outline format. The code includes the content, grade level, Reporting Category, Assessment Anchor, descriptor (Sub-Assessment Anchor), and Eligible Content. Thus, M.4.A.1.1.1 would be: Math, Grade 4, Reporting Category A, Assessment Anchor 1, descriptor (Sub-Assessment Anchor) 1, and Eligible Content 1.

Each of the Assessment Anchors has one or more descriptors (Sub-Assessment Anchors) and Eligible Content varying to reflect grade-level appropriateness. The Assessment Anchors form the basis of the test design for the grades undergoing new test development. In turn, this hierarchy is the basis for organizing the total content scores (based on the core [common] sections).

A draft version of the assessment anchors and eligible content for mathematics and reading was submitted to Achieve, Inc., Washington, D.C., to conduct a special analysis to evaluate the degree of alignment with the Academic Standards. Preliminary feedback enabled PDE to make adjustments to improve the alignment as the Assessment Anchors took final form. These adjustments were reflected operationally starting with the 2007 PSSA.

The Assessment Anchor Content Standards as defined by the Eligible Content are the same for the PSSA-M as they are for the general PSSA. However, in the PSSA-M, items measuring the Assessment Anchors as defined by the Eligible Content have been modified (revised and/or enhanced), when appropriate. Modifications, such as reduced text, easier vocabulary, simplified tasks, and the addition of hint boxes, allow for items to be more accessible to the given population of students but still in line with measuring the Assessment Anchors as defined by the Eligible Content. In so doing, the PSSA-M reflects the same emphasis and patterns as the general PSSA while utilizing a similar style and format. However, the PSSA-M does contain fewer items. These modifications, including fewer items and revisions and enhancements to items, are designed to allow students with disabilities a more suitable assessment opportunity in which to demonstrate proficiency.

The complete set of Assessment Anchors and Eligible Content can be referenced at PDE's website: www.education.state.pa.us. From the menu in the left-hand column, select "Programs," "Programs O–R," "Pennsylvania System of School Assessment (PSSA)," and then "Assessment Anchors." In addition, see Appendix A for more information about how the Academic Standards are linked to the Reporting Categories, Assessment Anchors, and Eligible Content.

Mathematics Assessment Measures

In keeping with the alignment of the PSSA, the PSSA-M mathematics assessments at Grades 4, 5, 6, 7, 8, and 11 have five major reporting categories: Numbers and Operations, Algebraic Concepts, Geometry, Measurement, and Data Analysis and Probability. By organizing the Assessment Anchors into a five-category reporting structure, there is a similarity to the categories used by the National Council of Teachers of Mathematics (NCTM) and the National Assessment of Educational Progress (NAEP). See Appendix A for more information about how the Academic Standards are linked to the Reporting Categories, Assessment Anchors, and Eligible Content.

In keeping with the PSSA, the PSSA-M mathematics assessment also employs two types of test items: multiple-choice and open-ended. These item types assess different levels of knowledge and provide different kinds of information about mathematics achievement. Psychometrically, multiple-choice items are very useful and efficient tools for collecting information about a student's academic achievement. Open-ended performance tasks are less efficient in the sense that they generally generate fewer scoreable points in the same amount of testing time. They do, however, provide tasks that are more realistic and better sample higher-level thinking skills. The design of the PSSA-M attempts to achieve a reasonable balance between the two item types. Furthermore, well-constructed scoring guides have made it possible to include open-ended tasks in large-scale assessments such as the PSSA-M. Trained scorers can apply the scoring guides to efficiently score large numbers of student papers in a highly reliable way.

MATHEMATICS MULTIPLE-CHOICE ITEMS

The majority of the mathematics items included on the PSSA-M, much like the PSSA, are multiple-choice (selected-response) items. This item type is especially efficient for measuring a broad range of content. In the PSSA and PSSA-M mathematics assessments, each multiple-choice item has four response options, only one of which is correct. The student is awarded one point for choosing the correct response. Distractors typically represent incorrect concepts, incorrect logic, incorrect application of an algorithm, or computation errors. It is important to note that for the PSSA-M eliminating an answer option is not an allowable modification.

Multiple-choice items are used to assess a variety of skill levels, from short-term recall of facts to problem solving. PSSA and PSSA-M items involving application emphasize the requirement to carry out some mathematical process to find an answer, rather than simply recalling information from memory.

OPEN-ENDED TASKS FOR MATHEMATICS

For both the PSSA and the PSSA-M, open-ended, or constructed-response tasks, require students to read a problem description and develop an appropriate solution. The PSSA-M open-ended items are designed to be scaffolded, which means that they have several components to the overall task that may enable students to enter or begin the problem at different places. In some items, each successive component is designed to assess progressively more difficult skills or higher knowledge levels. Certain components ask students to explain their reasoning for engaging in particular mathematical operations or for arriving at certain conclusions. The types of tasks utilized do not necessarily require computations. Students may also be asked to perform such tasks as constructing a graph, shading some portion of a figure, or listing object combinations that meet specified criteria.

Open-ended tasks are especially useful for measuring students' problem-solving skills in mathematics. They offer the opportunity to present real-life situations that require students to solve problems using mathematics abilities learned in the classroom. Students must read the task carefully, identify the necessary information, devise a method of solution, perform the calculations, enter the solution directly in the answer document, and when required, offer an explanation. This provides insight into the students' mathematical knowledge, abilities, and reasoning processes.

For both the PSSA and the PSSA-M, open-ended mathematics items are scored on a 0–4 point scale with an item-specific scoring guideline. The item-specific scoring guideline outlines the requirements at each score point. Item-specific scoring guidelines are based on the General Description of Mathematics Scoring Guidelines for Open-Ended Items. The general guidelines describe a hierarchy of responses, which represent the five score levels. See Appendix B or the *PSSA-M Mathematics Item and Scoring Samplers* available on the PDE website.

The tables on the following page provide a high-level overview of the operational mathematics PSSA-M test plan as compared to the general education mathematics PSSA. In addition, a comparison of the reporting categories for the mathematics PSSA-M and the general education mathematics PSSA is also provided. The PSSA-M content test blueprints show the same emphasis and patterns as the PSSA. The test content blueprints also show the extent to which the same or consistent categories of content appear in the PSSA-M and the PSSA. The PSSA-M, however, as noted in Table 2–1, is a shorter test.

Table 2–1. Mathematics Operational Test Plan Summary: PSSA and PSSA-M

	Program	Grades	Number of MC Items per PSSA	Number of 4-point OE Items per PSSA	Total Number of Points (MC + OE) per PSSA
Mathematics	PSSA	4, 5, 6, 7, 8, and 11	60	3	72
	PSSA-M	4, 5, 6, 7, 8, and 11	30	2	38

Table 2–2. Mathematics Blueprint (percentage of total test points): PSSA and PSSA-M

Reporting Category	Program	Grade					
		4	5	6	7	8	11
Numbers and Operations	PSSA	43%–47%	41%–45%	28%–32%	20%–24%	18%–22%	12%–15%
	PSSA-M	43%–47%	41%–45%	28%–32%	20%–24%	18%–22%	12%–15%
Measurement	PSSA	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%
	PSSA-M	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%
Geometry	PSSA	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
	PSSA-M	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
Algebraic Concepts	PSSA	12%–15%	13%–17%	15%–20%	20%–27%	25%–30%	38%–42%
	PSSA-M	12%–15%	13%–17%	15%–20%	20%–27%	25%–30%	38%–42%
Data Analysis & Probability	PSSA	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
	PSSA-M	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%

Chapter Three: Item Development Process

The core portion of the 2010 PSSA-M operational administration is made up of items that were field tested in the 2009 PSSA-M standalone field test. Therefore the activities that led to the 2010 PSSA-M operational administration begin with the development of the draft test items that appeared in the 2009 PSSA-M standalone field test.

STEPS IN THE DEVELOPMENT PROCESS

A series of major activities took place in the development of the PSSA-M. These key activities included the initial development of the guidelines for item revision and/or enhancement; cognitive interviews; item revision and/or enhancement of items; content review; bias, fairness, and sensitivity review; field testing of items in spring 2009; item review with data; and final selection of items to compose the 2010 PSSA-M mathematics assessment for Grades 4, 5, 6, 7, 8, and 11. These activities are summarized in Table 3–1 below, and they are further described in the paragraphs that follow.

Table 3–1. PSSA-M Mathematics Development Timeline

Time Frame	Assessment	Activity
September 2008–January 2009	’09 FT for ’10 OP	Item modifications implemented in preparation for 2009 standalone field test
January 2009	’09 FT for ’10 OP	Item review and bias, fairness, and sensitivity review for candidate items for the 2009 standalone field test
February–March 2009	’09 FT for ’10 OP	Forms construction for the 2009 standalone field test
May 2009	Cognitive Interviews	Cognitive Interviews conducted in Pennsylvania schools
May 2009	’09 FT for ’10 OP	PSSA-M Mathematics Standalone Field Test
June–July 2009	’10 FT for ’11 OP	Item modifications (revisions and/or enhancements) implemented in preparation for 2010 embedded field test
July–August 2009	’10 FT for ’11 OP	Item review and bias, fairness, and sensitivity review for candidate items of the 2010 embedded field test
August 2009	’09 FT for ’10 OP	Statistical review of the 2009 field tested items
September 2009–January 2010	’10 OP & ’10 FT for ’11 OP	Forms construction for the 2010 operational assessment with embedded field test
April 2010	’10 OP & ’10 FT for ’11 OP	2010 operational assessment

Item Development Planning Meeting

Prior to the start of any item development work, DRC's test development staff meets with PDE's assessment office to discuss the test development plans for the next PSSA administration, including the test blueprint, the field test plan (including development counts), procedures, timelines, etc. With a complete development cycle lasting several years (from item authoring through field test, data review, and operational usage), the initial planning begins well in advance of the anticipated administration. For the 2010 PSSA-M operational administration, the initial planning meetings for the item modifying process for the 2009 field test occurred throughout 2008. Item modifying began in fall 2008, with the item review meetings occurring in early 2009. (See Table 3–1 for additional details.)

Review of the Items

In September 2008, a pool of mathematics items from Grades 4, 5, 6, 7, 8, and 11 was reviewed. The review of the items focused upon whether each item might lend itself well to revision and/or enhancement for possible field testing of PSSA-M items in spring 2009. The pool of candidate items was comprised of PSSA mathematics items field tested in spring 2008.

Training

To begin the process, WestEd and DRC selected and trained mathematics staff to review PSSA items for possible revising and/or enhancement. Qualified mathematics content experts were college graduates with teaching experience and a demonstrated base of knowledge in the content area. Many of these writers were content assessment specialists and curriculum specialists. The writers were trained individually and had previous experience in writing and modifying multiple-choice and open-ended response items. Prior to modifying items for the PSSA-M, the cadre of item writers was trained with regard to the following:

- Pennsylvania Academic Standards, Assessment Anchors, and Eligible Content
- Webb's Four Levels of Cognitive Complexity: Recall, Basic Application of Skill/Concept, Strategic Thinking, and Extended Thinking
- General scoring guidelines for each content area
- Specific and General Guidelines for Item Writing
- Bias, Fairness, and Sensitivity
- Principles of Universal Design
- Item Quality Technical Style Guidelines
- Reference Information
- Sample Items

In addition, staff with a background in special education (e.g., those certified in special education and/or those with teaching experience in working with students with disabilities) and/or those with a background in developing assessments for the given population were also members of the team.

Training of mathematics staff at WestEd and DRC began with the study and discussion of the information presented in the *Pennsylvania System of School Assessment-Modified (PSSA-M) Alternate Assessment Based on Modified Achievement Standards Item Revision and/or Enhancement Guidelines for the Spring, 2009 Field Test November 13, 2008*. These guidelines were developed by WestEd with support from DRC. They were reviewed and approved by PDE prior to item revision and/or item enhancement. The guidelines served as the basis for all item revision and/or enhancement. A summary of the guidelines is given in the next section. The full guidelines are found in Appendix D of this document. It is important to note that these guidelines do adhere to the Principles of Universal Design (Center for Universal Design, 1997). NCEO has produced seven elements of Universal Design as they apply to assessments (Johnstone, Altman, & Thurlow, 2006).

These elements served to guide PSSA-M item revision and/or enhancement and are clearly noted in the Guidelines for Item Revision and Enhancement, found in Appendix D. Further discussion related to universal design considerations can be found in Chapter Four.

Table 3–2 shows the actual number of Multiple-Choice (MC) and Open-Ended (OE) items revised and/or enhanced for field testing in spring 2009. In some cases, during the review of the items, the reviewers determined that an existing item did not lend itself to revision and/or enhancement. In Table 3–2, these are noted "As Is" with no modifications made to the item.

Table 3–2. Number of Mathematics Items (MC and OE) Revised and/or Enhanced

Grade	MC Modified	MC As Is	Total MC	OE Modified	Total Items per Grade
4	78	3	81	11	92
5	66	9	75	10	85
6	52	8	60	7	67
7	52	8	60	8	68
8	54	8	62	7	69
11	52	10	62	7	69
Total	**357**	**46**	**400**	**50**	**450**

SUMMARY OF GENERAL REVISION AND/OR ENHANCEMENT GUIDELINES

Under the direction of the Pennsylvania Department of Education (PDE), the revisions and/or enhancements to PSSA items for mathematics, reading, and science were purposefully and necessarily made in order to address the eligible students' need for accessibility when taking the PSSA-M. The initial phases of PSSA-M item revisions and/or enhancements relied on expert judgment (e.g., PDE content-area experts and special educators; Pennsylvania educators, including both content-area educators and those special education educators with expertise in teaching the target population of students with disabilities). In addition, all revised and/or enhanced items were field tested in spring 2009 for mathematics and in spring 2010 for reading and science. The additional data collected on item performance of each field test item further served to validate the design and the revisions and/or enhancements of the PSSA-M items. The

data also offered PDE guidance in the selection of revised and/or enhanced items for the PSSA-M operational assessments. The types of revisions and/or enhancements to items are provided below.

Revisions

Students who will be eligible for the PSSA-M generally have difficulty processing information. As a result, revisions to items included the following:

- Simplifying the language in order to reduce the cognitive load or the amount of complex information, without changing the construct or what the item was intended to measure.
- Simplifying the language in order to remove any words that might be irrelevant.

Enhancements: Providing Supports

Enhancements to items involved embedding a type of support (e.g., adding graphics or artwork, providing definitions or context clues, providing scaffolds, and/or other permissible ways students might need to access and demonstrate understanding of the assessed content). Enhancement supports to items included the following:

- Providing helpful hints designed to support students' processing of information.
- Providing additional graphics and/or artwork to support understanding.
- Segmenting passages/prompts/scenarios, when appropriate. When passages are segmented, items follow an order that parallels how information generally appears in the passage, prompt, and/or scenario. (For example, for the reading PSSA-M, when appropriate, students will be provided the same passage/prompt/scenario as the general education PSSA at a given grade level, but the passage will be segmented or divided into meaningful parts. Those items that apply directly to each segment will appear directly after or adjacent to the referenced section of the text).
- Providing scaffolds such as adding hints or thought boxes (visual cues) to provide a further definition of a word or words and terminology and/or to support the text or emphasize main ideas.
- Providing supports for a number of steps and/or operations. For example, in a multi-step mathematics item, as appropriate, sub-questions or steps to break up or help students think through multi-step problems/item are provided.
- Adding additional directions to explain a process or activity.
- Adding pre-reading information to clarify the purpose of a passage, prompt, or scenario, such as the topic of a science scenario.
- Embedding a formula (as appropriate for intention of the assessed standard).

Enhancements: Visual Display

Enhancements to items also involve the degree to which the item format can be altered (e.g., introducing bolding, underlining, and other text changes, as well as changes in font size) and still provide a reliable measure of the student's knowledge/skill. Enhancements involving item format included the following:

- Adding more space between letters and words if item validity was not affected.
- Having fewer items per page, when appropriate.
- Increasing the width of an item or line length (from two columns to one, single-column layout so that the text of the item spans the entire width of the page), when appropriate.
- Restructuring the stem of an item into a stacked format. (Facts or details related to the item were indented and placed into a stacked format as well.)
- Inserting bullets to organize complex information or inserting bullets to break complex text within an item stem into smaller parts.

ITEM AUTHORING AND TRACKING

Initially, items are generated with software-prepared PSSA-M item cards and used for preliminary sorting and reviewing. Although very similar, the PSSA-M item card for multiple-choice items differs from the PSSA-M item card for open-ended items in that the former has a location at the bottom of the card for comments regarding the distractors. Blank examples of these two cards are shown in Appendix E. In both instances a column against the right margin provides for codes to identify the subject area, grade, content categories, passage information (in the case of reading), item type, depth of knowledge (i.e., cognitive complexity), estimated difficulty, answer key (for MC items), and calculator use (for mathematics).

All items undergoing field testing in 2009 were entered into the DRC Item Development and Educational Assessment System (IDEAS), which is a comprehensive, secure, online item banking system. It accommodates item writing, item viewing and reviewing, and item tracking and versioning. IDEAS manages the transition of an item from its developmental stage to its approval for use within a test form. The system supports an extensive item history that includes item usage within a form, item-level notes, content categories and subcategories, item statistics from both classical and Rasch item analyses, and classifications derived from analyses of differential item functioning (DIF). A sample IDEAS Item Card is presented in Appendix E.

INTERNAL REVIEWS AND PDE REVIEWS

To ensure that the items revised and/or enhanced were sufficient in number and adequately distributed across subcategories and levels of difficulty, content specialists, editors, and special education experts were informed of the required quantities of items needed for the external review by committees of Pennsylvania educators. Based upon the training received, content experts and special education experts began the process of revising and/or enhancing items. As items were revised and/or enhanced, they were entered into the item banking system along with important information (e.g., grade level, assessment anchor, eligible content, depth of knowledge). Subsequently, as an integral part of the internal item revision and/or enhancement process, each item was reviewed by a team of content specialists, editors, and special education experts both at WestEd and DRC. Content specialists, editors, and special education experts

evaluated each item to make sure that the construct had not changed and that it still measured the intended Eligible Content and/or Assessment Anchor Content Standard. They also assessed each item to make certain that the item revisions and/or enhancements were appropriate to the intended grade and that they provided and cued only one correct answer. In addition, the difficulty level, depth of knowledge, graphics, language demand, and distractors were also evaluated. Other elements considered in this process included, but were not limited to Universal Design considerations, adherence to the PDE-approved item revision, and enhancement guidelines, bias, source of challenge, grammar/punctuation, and PSSA-M style.

Following this internal process, revised and/or enhanced items were submitted to mathematics content specialists at the Pennsylvania Department of Education for review. PDE staff then consulted with WestEd and DRC about any general issues (style, format, interpretation of assessment anchors and eligible content) and about the revisions and/or enhancements to specific items. Following PDE's review, the revised and/or enhanced items were prepared for the content review meetings and the bias, fairness, and sensitivity meetings conducted with Pennsylvania educators. Information concerning these external reviews by Pennsylvania educators is provided below.

Review by Committees of Pennsylvania Educators

Before the PSSA-M items were field tested, two committees at two separate meetings reviewed them. The first meeting was the Bias, Fairness, and Sensitivity Meeting, and the second was the Item Content Meeting. The Bias, Fairness, and Sensitivity Meeting was held in Harrisburg, PA, on January 12 and 13 of 2009, and the Item Content Meeting also held in Harrisburg, PA, took place January 14 through January 16 of 2009. Summaries, guidelines, and procedures for each meeting are presented below.

BIAS, FAIRNESS, AND SENSITIVITY REVIEW

Prior to 2009 field testing, all revised and/or enhanced PSSA-M items were submitted to a Bias, Fairness, and Sensitivity Committee for review. As stated above, this took place on January 12 and 13 of 2009. The committee members consisted of a cross-representation of ethnic groups. (See Table 3–3.) Members of the committee also had expertise with special needs students and English Language Learners. All members had served on previous Pennsylvania Bias, Fairness, and Sensitivity Committees. The committee's primary responsibility was to evaluate items as to acceptability with regard to bias, fairness, and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the potential for issues of bias, fairness, and/or sensitivity.

The expert, multi-ethnic committee composed of men and women was trained by DRC and WestEd staff to review items for bias, fairness, and sensitivity issues. Training materials included a PDE-approved manual developed by DRC (DRC, 2003–2009). The focus of the training was on security and confidentiality; fairness in testing ensuring balanced treatment; definition of bias; and types of bias including stereotyping, gender, regional or geographical, ethnic or cultural, socioeconomic or class, religious, ageism, persons with disabilities, experiential, and sensitivity.

PDE staff members also attended the review and served as reviewers of the process. All mathematics PSSA-M items were read by a cross-section of committee members. Each member noted bias, fairness, and/or sensitivity comments on tracking sheets and on the item, if needed, for clarification. Committee members individually categorized any concerns as related to ageism, disability, ethnicity/culture, gender, regional, religious, socioeconomic, or stereotyping. These categories then formed the framework through which recommendations for modification or rejection of items occurred during the subsequent committee consensus process. The committee then discussed each of the issues as a group and came to consensus as to which issues should represent the view of the committee. All consensus comments were then compiled, and the suggested actions on these items were recorded and submitted to PDE. This review followed security procedures. Items in binders were distributed for committee review by number and signed for by each member on a daily basis. All attendees, with the exception of PDE staff, were required to sign a confidentiality agreement. All materials not in use at any time were stored in a locked room at the DRC offices in Harrisburg, PA. Secure materials that did not need to be retained after the meeting were deposited in secure barrels, the contents of which were shredded.

Table 3–3. Demographic Composition of the 2009 Bias, Fairness, and Sensitivity Committees

Member #	Gender	Race/Ethnicity	Background
1.	Male	Caucasian American	PDE Staff Member
2.	Male	Caucasian American	PDE Staff Member
3.	Male	Caucasian American	PDE Staff Member
4.	Female	Caucasian American	PATTAN Staff Member
5.	Female	Caucasian American	PDE Staff Member
6.	Female	Caucasian American	PDE Staff Member
7.	Male	Hispanic American	PATTAN Staff Member
8.	Female	Hispanic American	Local Community Leader
9.	Male	African American	Retired School Superintendent and Teacher
10.	Male	African American	PDE Staff Member
11.	Female	African American	Pennsylvania Teacher and Mathematics Coach/Specialist
12.	Female	African American	National Consultant for Special Education
13.	Female	Asian American	ESOL/Bilingual Education Specialist
Totals	7 Females 6 Males	4 African Americans 1 Asian Americans 6 Caucasian Americans 2 Hispanic Americans	

The results from the Bias, Fairness, and Sensitivity Committee review of PSSA-M mathematics are summarized in Table 3–4.

Table 3–4. Number of Items—2009
Bias, Fairness, and Sensitivity Committee Review for PSSA-M Mathematics

Grade	PSSA-M Mathematics			
	Items brought to Bias Review	Accepted As Is	Accepted With Revision	Rejected
4	91	85	6	0
5	81	80	1	0
6	70	69	1	0
7	70	66	4	0
8	70	69	1	0
11	70	67	3	0
Total	452	436	16	0

ITEM CONTENT REVIEW

Prior to the 2009 field testing, all revised and/or enhanced items were also submitted to content committees for review. This meeting took place in Harrisburg, PA, from January 14 through January 16, 2009. The content committees consisted of Pennsylvania educators from school districts throughout the Commonwealth of Pennsylvania. The committee members were selected to have both content expertise as well as expertise in teaching students with disabilities and/or those students who may be administered the PSSA-M assessment. The primary responsibility of the content committee was to evaluate the revised and/or enhanced items with regard to the quality of the revision and/or enhancement, the content classification or whether or not the construct had changed, including grade-level appropriateness of the revision and/or enhancement, estimated difficulty, depth of knowledge, and source of challenge. With source of challenge (Webb, 2002), items were identified where the cognitive demand was focused on an unintended content, concept, or skill. In addition, source of challenge was considered if the reason that an answer could be given resulted from a cultural bias, an inappropriate reading level, a flawed graphic in an item revision and/or enhancement, or if the item still required specialized, non-content related knowledge to answer. Source of challenge could result in the student answering—either correctly or incorrectly—without actually demonstrating the intended content or skill. Committee members were asked to note any items with a source of challenge and to suggest additional revisions and/or enhancements to remove the source of challenge. They also suggested additional and/or other revisions to items and/or other enhancements to the items. In some cases when an item was deleted, the committee members reviewed a suggested replacement item provided by the facilitators. The committee also reviewed the items for adherence to the guidelines for item revision and/or enhancement and the Principles of Universal Design, including language demand and issues of bias, fairness, and sensitivity.

Committee members were approved by PDE, and PDE-approved invitations were sent to them by DRC. PDE also selected internal PDE staff members for attendance. The meeting commenced with a welcome by PDE and DRC. This was followed by a PowerPoint presentation by DRC and WestEd. The PowerPoint presentation introduced the goals of the meeting, security and confidentiality, overview of the PSSA-M, and PSSA-M strategies for revising and/or enhancing items, including what could not be considered. The life of a PSSA item, the life of a PSSA-M item, the item review process, content alignment, rigor-level alignment, technical design, universal design, roles and responsibilities, and questions were also included in the PowerPoint training. In addition, the training also included procedures and forms to be used for item content review. Unique to this item review training was the presentation of sample items which included presenting each parent item along with the modified child item. These parent items were shown so the committee could see how the item originated as a PSSA item.

After the training, committee members were divided into groups. WestEd content assessment specialists facilitated the reviews and were assisted by representatives from DRC and PDE. The members reviewed each item and then came to consensus and assigned a status to each item as a group: Approved, Accepted with Revision, Move to Another Assessment Anchor or Grade, or Rejected. All comments were recorded, and a master rating sheet was completed. Committee facilitators recorded the committee consensus on an Item Review Rating Sheet.

Security was addressed by adhering to a strict set of procedures. Items in binders were distributed for committee review by number and signed for by each member on a daily basis. All attendees, with the exception of PDE staff, were required to sign a confidentiality agreement. All materials not in use at any time were stored in a locked room. Secure materials that did not need to be retained were deposited in secure barrels, the contents of which were shredded.

As the committee members reviewed the items and completed the Item Rating Sheets, they used the *PSSA-M Item Review Criteria Guidelines* produced by DRC and approved by PDE. These guidelines are found in Appendix F of this report. All committees had between 13 and 15 participants. Committees included a mixture of veteran item reviewers, new reviewers, and special education teachers. In general, all participants had been exposed to special needs students and they paid close attention to what the special education teacher had to say about the items. There were good discussions among the members of the committees, and overall, they liked the modifications that were made to the items.

All committee-recommended edits were reviewed by PDE. Approved edits were provided to DRC. All PDE approved edits were made. The revised and/or enhanced items were then made available for the Cognitive Interviews.

COGNITIVE INTERVIEWS

As a part of the development process for the PSSA-M, Cognitive Interviews were also conducted. In order for the results of the Cognitive Interviews to help inform the item revision and/or enhancement process for the PSSA-M mathematics assessment, the interviews were conducted prior to field testing of the items. In addition to mathematics items, the Cognitive Interviews involved reading and science items. The following information summarizes the process used for the Cognitive Interviews. The introduction, study overview, and rationale for the PSSA-M cognitive interviews is based upon the *Cognitive Interviews in Pennsylvania: Report on Data Collection for the Pennsylvania System of School Assessment Alternate*

Assessment with Modified Achievement Standards (PSSA-M) Study (PDE, 2009).[1] Additional details found in this report include the method used to conduct the interviews; target sample size; characteristics of the districts selected to participate; the process of school and student recruitment; informed consent; interview process; item booklets; teacher survey; findings; frequency of responses; findings by cluster, linguistic enhancements, test design enhancements, typographic feature enhancements; challenges with terminology and vocabulary, and findings by item enhancement type.

In order to provide help in identifying the need for an additional alternate assessment, PDE requested additional information from the Cognitive Interviews including information concerning accommodations used during instruction, effective tasks/task types that might help students with disabilities demonstrate their knowledge and ability, corroboration of enhancement strategies employed on PSSA-M, preparing students with disabilities for the PSSA-M, and application of PSSA-M results.

1. Introduction

Data Recognition Corporation (DRC), in collaboration with WestEd, proposed to the Commonwealth of Pennsylvania a study intended to provide PDE with information the Department might want to consider when making decisions concerning the development of the PSSA-M. More specifically, DRC's subcontractor, WestEd, designed and conducted Cognitive Interviews with general education students and students with disabilities to examine the degree to which revision and/or enhancement strategies applied to PSSA-M items facilitated student access (their ability to understand and demonstrate their grade-level content understanding) to tested content. The Cognitive Interviews were conducted in Pennsylvania schools between May 11 and May 29, 2009. The sections below present an overview of the study, the Cognitive Interview methodology, and findings. Implications for future development of the PSSA-M also are presented.

2. Study Overview

The study systematically evaluated the strategies used to develop items to be field-tested and the degree to which these strategies facilitated students' abilities to demonstrate what they knew and could do. More specifically, this Cognitive Interview study intended to address the following question: What are the cognitive processes by which test items (or item types) are understood by students?

Data were collected from 252 students in Grades 4, 5, 6, 7, 8, and 11 enrolled in Pennsylvania public schools in five districts across the Commonwealth, and from teachers in those schools who work primarily with PSSA-M-eligible students. This process is further described below.

3. Rationale for Cognitive Interviews

In the study, Cognitive Interviews were conducted to examine the effectiveness of the item enhancement strategies currently used during development of the PSSA-M mathematics field test items. Reading and science items were also included in the study. The results provided information concerning the degree to which current enhancement strategies—which consist primarily of changes to item structure or

[1] This report is available upon request from PDE at 1-717-705-2343.

format—increase access to test items for students with disabilities (SWDs) and general education students.

Cognitive interviewing strategies were drawn from the family of process-tracing or verbal protocol models that can be used to confirm or verify hypotheses about access to tested content. They provided a forum for the researchers to test assumptions about the intent of an item or question. By microanalyzing the items (Solano-Flores & Trumbull, 2003), the researchers could simultaneously gather information about students' understanding of task expectations; their levels of mastery of the content; and the reasoning processes, problem-solving strategies, and adaptive skills students use when answering test questions (Ericsson & Simon, 1980, 1993; Paulsen & Levine, 1999).

During each Cognitive Interview, researchers observe students individually as they respond to test questions. As students attempt to answer each item or solve each problem, they are encouraged to articulate, or say out loud their interpretation of the task required and the steps or processes needed to complete the task (concurrent data collection). Student comments, observations, insights, and responses about directions, item stem, response choices, and graphics or stimuli help the researchers check assumptions about whether a test item is functioning as intended; that is, that the assessment task actually taps the cognitive processes that are intended to be assessed (National Research Council, 2001).

The Cognitive Interview process used in Pennsylvania was conducted in three steps (adapted from Sato, Rabinowitz, Gallagher & Huang, in press). In the first step, the student was introduced to the interview process and allowed to practice thinking aloud. In the second step, data was collected concurrently as the student spoke out loud as he/she attempted to answer each test question. Via prompts, the researcher interacted with the student to elicit verbal responses that described his/her understanding of the test question and strategies for answering it. In the third step, the retrospective stage of data collection, students were asked specific questions about the test item (probes) immediately after answering it. At this point, most students could look back, recall, and discuss what they did to answer the question or solve the problem; in this way, they could verify or clarify their earlier comments. Once the student responded to all test items, the researcher asked each student a set of follow-up questions to clarify or verify comments collected earlier and/or to probe deeper into the student's thinking processes about that item.

This multi-step process helped reveal the types of prior/background knowledge and/or requisite skills that may have supported students' abilities to respond to the item and to assess the consequences of their decisions (Kopriva, 2001). Data collected through the Cognitive Interview contributed to information that helped to validate the interpretations of test performance outcomes by indicating the degree to which students' demonstrated understanding concurred with the construct intended to be measured by the item. From these interviews, specific, richly descriptive data were collected. This data was then used to help inform decision-making about the strategies currently used to revise and/or enhance items for the PSSA-M so that these enhancements would appropriately facilitate student access to the assessed content.

Summary of Cognitive Interviews

As stated above, the purpose of the Cognitive Interview study was to systematically evaluate the strategies used to develop (revise and/or enhance) items for the PSSA-M and the degree to which these strategies facilitated students' ability to demonstrate what they know and can do. The study addressed the following question: What are the cognitive processes by which test items (or item types) are understood by students?

Test items used in this study reflected a range of revision and enhancement strategies intended to facilitate the access to assessed content of students eligible for the PSSA-M. Results of the study suggested that a number of the revision and enhancement strategies, such as those related to linguistic enhancements or test design enhancements, helped students with their performance on the items included in this study.

TEST CONTENT BLUEPRINT FOR 2010 PSSA-M MATHEMATICS ASSESSMENT

The PSSA-M, like the PSSA, is based on the Pennsylvania Academic Standards. The 2010 PSSA and PSSA-M reflect the new Assessment Anchors (PDE 2004), which were designed as a means of improving the articulation of curricular, instructional, and assessment practices. The Assessment Anchors serve to clarify the Academic Standards assessed on the PSSA and to communicate assessment limits, or the range of knowledge and skills from which the PSSA would be designed. Relevant to item development are the refinement and clarification embodied in the Assessment Anchors.

Table 3–5. Mathematics Blueprint (percentage of total test points): PSSA and PSSA-M

Reporting Category	Program	Grade					
		4	5	6	7	8	11
Numbers and Operations	PSSA	43%–47%	41%–45%	28%–32%	20%–24%	18%–22%	12%–15%
	PSSA-M	43%–47%	41%–45%	28%–32%	20%–24%	18%–22%	12%–15%
Measurement	PSSA	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%
	PSSA-M	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%	12%–15%
Geometry	PSSA	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
	PSSA-M	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
Algebraic Concepts	PSSA	12%–15%	13%–17%	15%–20%	20%–27%	25%–30%	38%–42%
	PSSA-M	12%–15%	13%–17%	15%–20%	20%–27%	25%–30%	38%–42%
Data Analysis & Probability	PSSA	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%
	PSSA-M	12%–15%	12%–15%	15%–20%	15%–20%	15%–20%	12%–18%

Operational Layout for 2010 PSSA-M Mathematics

The PSSA-M mathematics assessments for Grades 4 through 8 and Grade 11 are combined into one integrated test/answer booklet for each grade. The modified booklets contain scannable pages for multiple-choice (MC) responses, open-ended (OE) items with response spaces, and demographic data collection areas. All MC items are worth 1 point. OE items receive a maximum of 4 points (scale of 0–4).

For 2010, each test form contained common items (identical on all three forms) along with embedded field test items. The common items consisted of a set of core items taken by all students. The embedded field test items were unique to a form.

The 2010 PSSA-M was comprised of 3 forms per grade. All of the forms contained the common items identical for all students and sets of generally unique items that fulfill the purpose of field testing new items (FT items). Tables 3–6 and 3–7 display information about the test form layout.

Table 3–6. 2010 PSSA-M Mathematics Operational Test Plan Summary

Content Area	Year	Number of Common (Core) MC* Items per Form	Number of Common (Core) OE** Items per Form	Number of Field Test MC Items per Form	Number of Field Test OE Items per Form	Number of Forms per Grade
Mathematics	2010	30	2	8	1	3

*MC = Multiple-choice
**OE = Open-ended

Table 3–7. 2010 PSSA-M Math Operational Test Layout

Item Stage	Section 1	Section 2
Core	15 MC	15 MC
Core	1 OE	1 OE
Field test	1 OE	8 MC

Since an individual student's score is based solely on the common (or core) items, the total number of operational points is 38. The total score is obtained by combining the points from the core MC and OE portions of the test as follows:

Table 3–8. 2010 PSSA-M Mathematics Core Points

Student's Score	MC Items	Grades	OE Items	Total Score
Mathematics	30	4, 5, 6, 7, 8, & 11	2 items x 4 points=8 points	38

For more information concerning the process used to convert the operational layout into forms (form construction), see Chapter Six.

Linking for 2010 and 2011 PSSA-M Mathematics Assessments

Linking provides a statistical bridge between assessment administrations. The 2011 administration will be linked back to the 2010 administration through the use of linking items in the core (core-to-core link). In the PSSA-M, only multiple-choice items will be used for linking purposes. Open-ended items will not be repeated as linking items across cores.

MULTIPLE-CHOICE

For Grades 4, 5, 6, 7, 8, and 11, multiple-choice items will be repeated on the subsequent form for the purpose of linking.

The matter of linking will be treated more fully in Chapter Fifteen.

Test Sessions and Timing for 2010 PSSA-M Mathematics Assessment

The test window for the 2010 operational assessment, including make-ups, extended from April 7 through May 7, 2010. The mathematics assessments consisted of two sections. Test administration recommendations called for each section to be scheduled as one assessment session, and schools were not permitted to combine both sections into a single session. Administration guidelines stipulated that the sections be administered in the sequence in which they are printed in the test booklets. The following tables outline the assessment schedule and estimated times for each section. The estimated Student Testing times do not include time for administrative tasks that occur during the pre- and post-administration activities. These times are estimated separately. Times are approximate and are supplied to test administrators for scheduling purposes only.

Table 3–9. PSSA-M Mathematics—2010 Administration and Testing Times

Test Section	Suggested Times (In Minutes)			Grade Level Number of Items and Item Type					
	Administration (Total)	Administrative (Pre & Post)	Estimated Student Testing	4	5	6	7	8	11
1	75 to 90	15 to 20	60 to 70	15 MC 2 OE	15 MC 2 OE	15 MC 2 OE	15 MC 2 OE	15 MC 2 OE	15 MC 2 OE
2	90 to 105	15 to 20	75 to 85	23 MC 1 OE	23 MC 1 OE	23 MC 1 OE	23 MC 1 OE	23 MC 1 OE	23 MC 1 OE

During the assessment, students may request an extended assessment period if they indicate that they have not completed the task. Such requests are granted if the assessment administrator finds the request to be educationally valid. See Chapter Seven for more information about testing sessions.

Reporting Categories and Points Distributions for 2010 PSSA and PSSA-M Mathematics Assessments

The mathematics assessment results will be reported in five categories that approximately correspond to those advocated by the National Council of Teachers of Mathematics (NCTM). The code letters for these Assessment Anchor reporting categories are A–E and correspond to the following:

A. Numbers and Operations

B. Measurement

C. Geometry

D. Algebraic Concepts

E. Data Analysis and Probability

The distribution of test points into these five categories and their percentages of the total number of test points are shown in the following table.

Table 3–10. Mathematics Reporting Categories and Point Distributions

Grade	Reporting Categories					Total Points
	A: Numbers and Operations	B: Measurement	C: Geometry	D: Algebraic Concepts	E: Data Analysis & Probability	
4	43%–47% 16–18 points	12%–15% 5–6 points	12%–15% 5–6 points	12%–15% 5–6 points	12%–15% 5–6 points	38
5	41%–45% 16–17 points	12%–15% 5–6 points	12%–15% 5–6 points	13%–17% 5–6 points	12%–15% 5–6 points	38
6	28%–32% 11–12 points	12%–15% 5–6 points	15%–20% 6–8 points	15%–20% 6–8 points	15%–20% 6–8 points	38
7	20%–24% 8–9 points	12%–15% 5–6 points	15%–20% 6–8 points	20%–27% 8–10 points	15%–20% 6–8 points	38
8	18%–22% 7–8 points	12%–15% 5–6 points	15%–20% 6–8 points	25%–30% 10–11 points	15%–20% 6–8 points	38
11	12%–15% 5–6 points	12%–15% 5–6 points	12%–18% 5–7 points	38%–42% 14–16 points	12%–18% 5–7 points	38

The mathematics reporting categories are further subdivided for specificity and Eligible Content or limits. Each subdivision is coded by adding an additional numeral, such as A.1. These subdivisions are called Assessment Anchors and Eligible Content.

Assessment Anchor Content Standards Subsumed within Reporting Categories for 2010 Modified Mathematics Assessment

For mathematics, there are 16 Assessment Anchor Content Standards (Assessment Anchors) that occur at all grade levels (Grades 4 through 8 and 11), although they are not all assessed at each grade level. More specifically, the number targeted for assessment by grade level are 12 at Grade 4; 13 at Grade 5; 12 at Grade 6; 14 at Grade 7; 13 at Grade 8; and 13 at Grade 11.

Mathematics scores are based on the core (common) sections. Also reported are the student's mathematics performance levels. See Appendix C for a summary by grade.

TEST DEVELOPMENT CONSIDERATIONS FOR THE PSSA-M

Alignment to the PSSA Assessment Anchors and Eligible Content (or, in the case of writing, strong alignment with the PSSA Academic Standards), grade-level appropriateness (reading/interest level, etc.), depth of knowledge, cognitive level, item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology were major considerations in the item development process. The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) and the *Principles of Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the development process. In addition, DRC's *Bias, Fairness, and Sensitivity Guidelines* were used for developing items. All items were reviewed for fairness by bias and sensitivity committees and for content by Pennsylvania educators and field-specialists. Items were also reviewed for adherence to the Principles of Universal Design by representatives from the National Center for Educational Outcomes (NCEO) as well as adherence to the guidelines outlined in the Pennsylvania publication *Principles, Guidelines and Procedures for Developing Fair Assessment Systems: Pennsylvania Assessment Through Themes* (PATT).

Bias, Fairness, and Sensitivity

At every stage of the item and test development process, DRC employs procedures that are designed to ensure that items and tests meet Standard 7.4 of the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999).

> *Standard 7.4: Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.*

To meet Standard 7.4, DRC employs a series of internal quality steps. DRC provides specific training for test developers, item writers, and reviewers on how to write, review, revise, and edit items for issues of bias, fairness, and sensitivity (as well as for technical quality). Training also includes an awareness of and sensitivity to issues of cultural diversity. In addition to providing internal training in reviewing items in order to eliminate potential bias, DRC also provides external training to the review panels of minority experts, teachers, and other stakeholders.

DRC's guidelines for bias, fairness, and sensitivity include instruction concerning how to eliminate language, symbols, words, phrases, and content that might be considered offensive by members of racial, ethnic, gender, or other groups. Areas of bias that are specifically targeted include, but are not limited to: stereotyping, gender, regional/geographic, ethnic/cultural, socioeconomic/class, religious, and experiential, as well as biases against a particular age group

(ageism) and against persons with disabilities. DRC catalogues topics that should be avoided, and maintains balance in gender and ethnic emphasis within the pool of available items.

Universal Design

As stated above, the Principles of Universal Design were incorporated throughout the item development process to allow participation of the widest possible range of students in the PSSA. The following checklist was used as a guideline:

- Items measure what they are intended to measure.
- Items respect the diversity of the assessment population.
- Items have a clear format for text.
- Stimuli and items have clear pictures and graphics.
- Items have concise and readable text.
- Items allow changes to other formats, such as Braille, without changing meaning or difficulty.
- The arrangement of the items on the test has an overall appearance that is clean and well organized.

A more extensive description of the application of Principles of Universal Design is described in Chapter Four.

Depth of Knowledge

An important element in statewide assessment is the alignment between the overall assessment system and the state's standards. A methodology developed by Norman Webb (1999) offers a comprehensive model that can be applied to a wide variety of contexts. With regard to the alignment between standards statements and the assessment instruments, Webb's criteria include five categories, one of which deals with content. Within the content category is a useful set of levels for evaluating depth of knowledge (DOK). According to Webb (1999, p.7–8) "depth-of-knowledge consistency between standards and assessments indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards." The four levels of cognitive complexity (depth of knowledge) are as follows:

- Level 1: Recall
- Level 2: Skill/Concept
- Level 3: Strategic Thinking
- Level 4: Extended Thinking

Depth-of-knowledge levels were incorporated in the item writing and review process, and items were coded with respect to the level they represented. Generally, multiple-choice items are written to DOK levels 1 and 2, and open-ended items are written to DOK level 3.

Test Item Readability

Careful attention was given to the readability of the items to make certain that the assessment focus of the item did not shift based on the difficulty of reading the item. The issue of readability was addressed for all items during the final editing of items and at the Item Content Review. Vocabulary was also addressed at the Bias, Fairness, and Sensitivity Review, although the focus was on how certain words or phrases may represent a possible source of bias or issues of fairness or sensitivity.

TEST DEVELOPMENT PROCESS

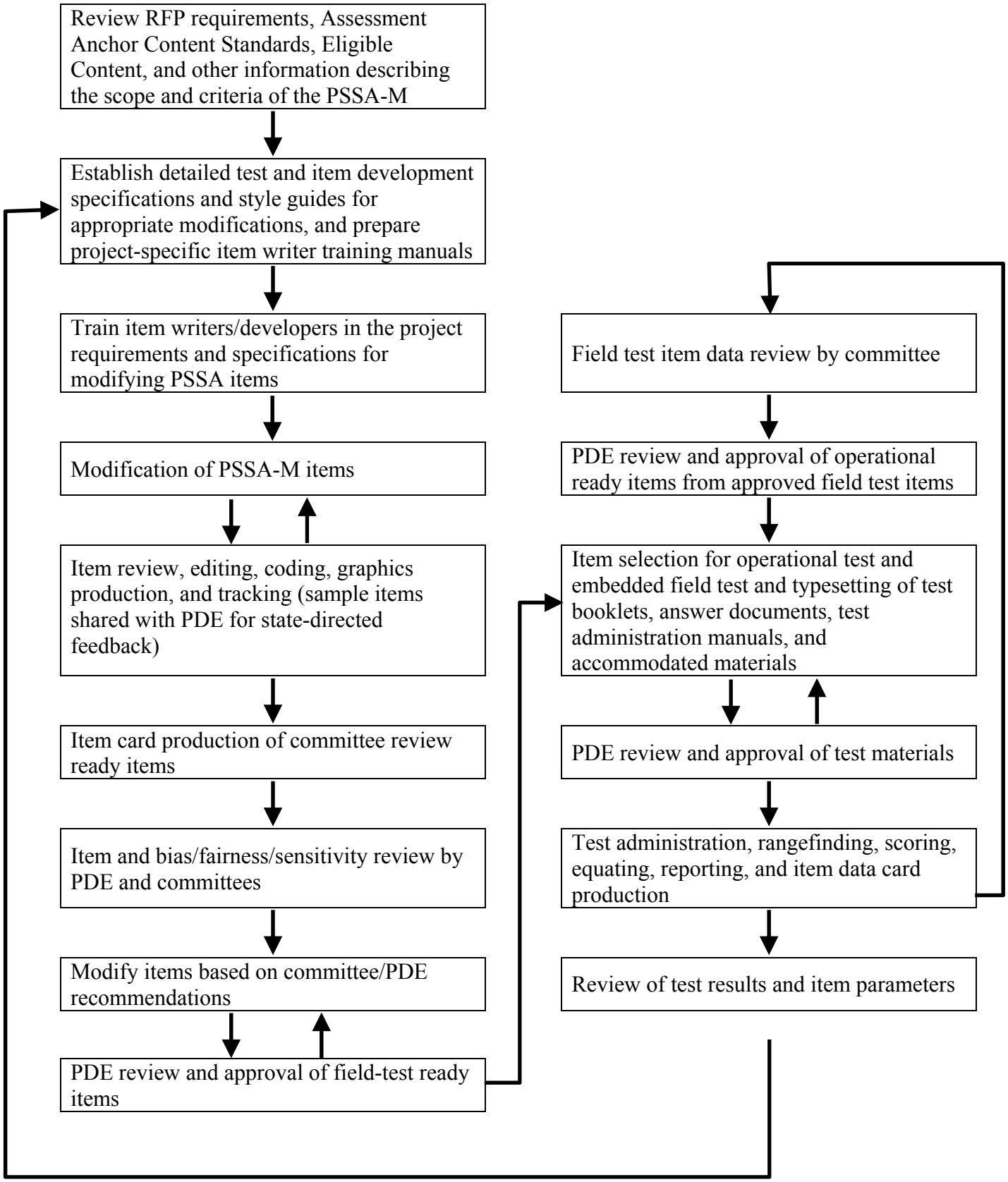
The item development process follows a logical timeline, which is outlined below in Figure 3–1. On the front-end of the schedule, tasks are generally completed with the goal of presenting field test candidate items to committees of Pennsylvania educators. On the back-end of the schedule, all tasks lead to the field test data review.

Figure 3–1. Item and Test Development Cycle and Timeline (2009–2010 only)

Steps in Development Cycle	Timeline to/from New Item Review		
Development planning	Fall	⇩	-12 to -4 months
Initial item modifying	Fall	⇩	-3 to -2 months
Internal reviews and PDE reviews	Fall/Winter	⇳	-3 to -1 months
Bias, Fairness, and Sensitivity Review	Winter	⇩	+/- 0 months
Newly Modified Item Content Review	**Winter**	⇨	**+/- 0 months**
Post-review resolution and clean-up	Winter	⇩	+/- 0 months
Build test forms	Spring	⇩	+0 to +1 months
Internal form reviews and PDE reviews	Spring	⇳	+1 to +2 months
Form printing, packaging, and shipping	Spring	⇩	+2 to +3 months
Test administration	Spring	⇩	+4 months
Material/data processing, rangefinding, and scoring	Spring/Summer	⇩	+4 to +7 months
Field Test Item Data Review	**Summer**	⇨	**+7 months**
Select operational items	Summer/Fall	⇩	+8 to +10 months

A process flowchart that illustrates the interrelationship among the steps in the process is shown in Figure 3–2. In addition, a detailed process table describing the item and test development processes also appears in Appendix D.

Figure 3–2. DRC Item and Test Development Process for PSSA-M



Chapter Four: Universal Design Procedures Applied in the Modified PSSA Test Development Process

Universally designed assessments allow participation of the widest possible range of students and contribute to valid inferences about participating students. Principles of Universal Design are based on the premise that each child in school is a part of the population to be tested and that testing results should not be affected by disability, gender, race, or English language ability (Thompson, Johnstone & Thurlow, 2002). At every stage of the item and test development process, including the 2009 field test, procedures were employed to ensure that items and subsequent tests were designed and developed using the elements of universally designed assessments developed by the National Center for Educational Outcomes (NCEO).

Federal legislation addresses the need for universally designed assessments. The *No Child Left Behind Act* (Elementary and Secondary Education Act) requires that each state must "provide for the participation in [statewide] assessments of all students" [Section 1111(b)(3)(C)(ix)(l)]. Both Title 1 and IDEA regulations call for universally designed assessments that are accessible and valid for all students, including students with disabilities and English Language Learners. The benefits of universally designed assessments not only apply to these groups of students, but to all individuals with wide-ranging characteristics.

DRC's test development team was trained in the elements of Universal Design as it relates to developing large-scale statewide assessments. Team leaders were trained directly by NCEO, and other team members were subsequently trained by team leaders. Committees involved in content review included some members who were familiar with the unique needs of students with disabilities and English Language Learners. Likewise some members of the Bias, Fairness, and Sensitivity Committee were conversant with these issues. What follows are the Universal Design guidelines followed during all stages of the item development process for the PSSA-M.

ELEMENTS OF UNIVERSALLY DESIGNED ASSESSMENTS

After a review of research relevant to the assessment development process and the principles of Universal Design (Center for Universal Design, 1997), NCEO has produced seven elements of Universal Design as they apply to assessments (Thompson, Johnstone & Thurlow, 2002). These elements served to guide PSSA-M item development.

- **Inclusive Assessment Population**

  The PSSA-M is intended for students with disabilities functioning above the lowest 1% of the population, but not at a level that allows them to access the general Pennsylvania System of School Assessment (PSSA). The PSSA-M utilizes modified items designed to allow students with disabilities to demonstrate proficiency on the assessment.

- **Precisely Defined Constructs**

  An important function of well-designed assessments is that they actually measure what they are intended to measure. The Pennsylvania Assessment Anchor Content Standards (Assessment Anchors) provided clear descriptions of the constructs to be measured by the PSSA-M at the assessed grade levels. Universally designed assessments must remove all non-construct-oriented cognitive, sensory, emotional, and physical barriers.

- **Accessible, Non-biased Items**

  DRC conducted both internal and external reviews of items and test specifications to ensure that they did not create barriers because of lack of sensitivity to disability, culture, or other subgroups. Items and test specifications were developed by a team of individuals who understand the varied characteristics of items that might create difficulties for any group of students. Accessibility is incorporated as a primary dimension of test specifications, so that accessibility was woven into the fabric of the test rather than being added after the fact.

- **Amenable to Accommodations**

  Even though items on universally designed assessments are accessible for most students, there are some students who continue to need accommodations. This essential element of universally designed assessment requires that the test is compatible with accommodations and a variety of widely-used adaptive equipment and assistive technology. (See the section on Assessment Accommodations on page 36 in this chapter.)

- **Simple, Clear, and Intuitive Instructions and Procedures**

  Assessment instructions should be easy to understand, regardless of a student's experience, knowledge, language skills, or current concentration level. Knowledge questions that are posed within complex language can invalidate the test if students cannot understand how they are expected to respond to a question. To meet this guideline, directions and questions were prepared in simple, clear, and understandable language that underwent multiple reviews.

- **Maximum Readability and Comprehensibility**

  A variety of guidelines exist to ensure that text is maximally readable and comprehensible. These features go beyond what is measured by readability formulas. Readability and comprehensibility are affected by many characteristics, including student background, sentence difficulty, text organization, and others. All of these features were considered as item text was developed.

  Plain language is a concept now being highlighted in research on assessments. Plain language has been defined as language that is straightforward and concise. The following strategies for editing text to produce plain language were used during the editing process of the newly modified PSSA-M items:

  - Reduction of excessive length
  - Use of common words
  - Avoidance of ambiguous words
  - Avoidance of irregularly spelled words
  - Avoidance of proper names
  - Avoidance of inconsistent naming and graphic conventions
  - Avoidance of unclear signals about how to direct attention

- **Maximum Legibility**

  Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. Bias results when tests contain physical features that interfere with a student's focus on or understanding of the constructs that test items are intended to assess. A style guide which included dimensions of style consistent with Universal Design was developed and updated annually (DRC 2004–2009) and was utilized with PDE approval.

GUIDELINES FOR UNIVERSALLY DESIGNED ITEMS

All modified and reviewed test items adhered closely to the following guidelines for Universal Design. Item writers and reviewers used a checklist during the item development process to ensure that each aspect was attended to. For information on the checklist, see the Universal Design section in Chapter Three of this report.

1. **Items measure what they are intended to measure.** Item writing training included assuring that writers and reviewers had a clear understanding of Pennsylvania's Academic Standards and the Assessment Anchors. During all phases of test development, items were presented with content-standard information to ensure that each item reflected the intended Assessment Anchor. Careful consideration of the content standards was important in determining which skills involved in responding to an item were extraneous and which were relevant to what was being tested. In certain types of items an additional skill is necessary, such as the mathematics test, which requires the student to read.
2. **Items respect the diversity of the assessment population.** To develop items that avoid content that might unfairly advantage or disadvantage any student subgroup, item writers, test developers, and reviewers were trained to write and review items for issues of bias, fairness, and sensitivity. Training also included an awareness of, and sensitivity to, issues of cultural and regional diversity.
3. **Items have a clear format for text.** Decisions about how items are presented to students must allow for maximum readability for all students. Appropriate fonts and point sizes were employed with minimal use of italics, which is far less legible and is read considerably more slowly than standard typeface. Captions, footnotes, keys, and legends were at least a 13-point size. Legibility was enhanced by sufficient spacing between letters, words, and lines. Blank space around paragraphs and between columns and staggered right margins were used.
4. **Stimuli and items have clear pictures and graphics.** When pictures and graphics were used, they were designed to provide essential information in a clear and uncluttered manner. Illustrations were placed directly next to the information to which they referred, and labels were used where possible. Sufficient contrast between background and text, with minimal use of shading, increased readability for students with visual difficulties. Color was not used to convey important information.

5. **Items have concise and readable text.** Linguistic demands of stimuli and items can interfere with a student's ability to demonstrate knowledge of the construct being assessed. During item writing and review, the following guidelines were used.
   - Simple, clear, commonly-used words were used whenever possible.
   - Extraneous text was omitted.
   - Vocabulary and sentence complexity were appropriate for the grade level assessed.
   - Technical terms and abbreviations were used only if related to the content being measured.
   - Definitions and examples were clear and understandable.
   - Idioms were avoided unless idiomatic speech was being assessed.
   - The questions to be answered were clearly identifiable.
6. **Items allow changes to format without changing meaning or difficulty.** A Braille version of the PSSA-M was available at each assessed grade. Attention was given to using items that allow for Braille. Specific accommodations were permitted such as signing to a student, the use of oral presentation under specified conditions, and the use of various assistive technologies. A Spanish version for the PSSA-M mathematics was available for use by English Language Learners who would benefit from this accommodation.
7. **The test has an overall appearance that is clean and organized.** Images, pictures, and text that may not be necessary (e.g., sidebars, overlays, callout boxes, visual crowding, shading) and that could be potentially distracting to students were avoided. Also avoided were purely decorative features that did not serve a purpose. Information was organized in a manner consistent with an academic English framework with a left-right, top-bottom flow.

ITEM DEVELOPMENT

DRC and WestEd worked closely with the Pennsylvania Department of Education to help ensure that PSSA-M tests comply with nationally recognized Principles of Universal Design. We supported the implementation of accommodations on large-scale statewide assessments for students with disabilities. In addition to the Principles of Universal Design as described in the Pennsylvania Technical Report, DRC and WestEd applied to each content area assessment the standards for test accessibility as described in *Tests Access: Making Tests Accessible for Students with Visual Impairments—A Guide for Test Publishers, and State Assessment Personnel* (Allman, 2004). To this end, we embraced the following precepts:

- Test directions were carefully worded to allow for alternate responses to open-ended questions.
- During item and bias reviews, test committee members were made aware of the Principles of Universal Design and of issues that may adversely affect students with disabilities with the goal of ensuring that PSSA-M tests are bias free for all students.

- With the goal of ensuring that the PSSA-M tests are accessible to the widest range of diverse student populations, PDE instructed DRC and WestEd to limit item types that are difficult to format in Braille, and that may become distorted when published in large print. DRC and WestEd were instructed to limit the following on the PSSA-M.
  - Mathematics: complicated tessellations, a chart or graph that extends beyond one page.
  - Reading: graphics and illustrations that are not germane to the content presented.
  - All content areas: unnecessary boxes and framing of text, unless enclosing the text provides necessary context for the student; use of italics (limited to only when it is absolutely necessary, such as with variables).

ITEM FORMATTING

For all content areas, DRC formatted PSSA-M tests to maximize accessibility for all students by using text that is in a 13-point size and font style that is easily readable. DRC limited shading, spacing, graphics, charts, and number of items per page so that there was sufficient white space on each page. Whenever possible, we ensured that graphics, pictures, diagrams, charts, and tables were positioned on the page with the associated test items. We use high contrast for text and background where possible to convey pertinent information. Tests were published on dull-finish paper to avoid the glare encountered on glossy paper. DRC paid close attention to the binding of the PSSA-M test booklets to ensure that they lie flat for two-page viewing and ease of reading and handling.

DRC ensured consistency across PSSA and PSSA-M assessments by following these Principles of Universal Design:

- High contrast and clarity is used to convey detailed information.
- Typically, shading is avoided; when necessary for content purposes, 10 percent screens are used as the standard.
- Overlaid print on diagrams, charts, and graphs is avoided.
- Charts, graphs, diagrams, and tables are clearly labeled with titles and with short descriptions where applicable.
- Only relevant information is included in diagrams, pictures, and graphics.
- Symbols used in keys and legends are meaningful and provide reasonable representations of the topic they depict.
- Pictures that require physical measurement are true to size.

ASSESSMENT ACCOMMODATIONS

While universally designed assessments provide for participation of the widest range of students, many students require accommodations in order to participate in the regular assessment. Clearly, the intent of providing accommodations for students is to ensure that students are not unfairly disadvantaged during testing and that the accommodations used during instruction, if appropriate, are made available as students take the test. The literature related to assessment accommodations is still evolving and often focuses on state policies regulating accommodations rather than on providing empirical data that supports the reliability and validity of the use of accommodations. On a yearly basis, the Pennsylvania Department of Education examines accommodations policies and current research to ensure that valid, acceptable accommodations are available for students. An accommodations manual for the PSSA and PSSA-M entitled *PSSA & PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans* (PDE, January 2010) was developed for use with the 2010 PSSA and PSSA-M.

The manual can be accessed by going to www.education.state.pa.us. On the left, click on "Programs," then "Programs O–R," then "Pennsylvania System of School Assessment (PSSA)" and then "Testing Accommodations & Security."

In addition, Spanish-language versions, translated from the original English versions, were made available for the PSSA-M mathematics assessments. The Spanish translation versions are discussed further in Chapter Six.

Chapter Five: Field Test Leading to the 2010 Core

EMBEDDED FIELD TEST ITEMS

All core items appearing on the 2010 assessment came from the Spring 2009 standalone field test. The purpose of administering field-test items is to obtain statistics for them so they can be reviewed before becoming operational. Based on the statistical review, many of the field test items tested in the 2009 PSSA-M standalone field test were selected for use as common core items in the 2010 PSSA-M.

Table 5–1. 2009 Spring PSSA-M Field Test

Grades	No. of FT MC per Form	No. of FT OE per Op. Form	Total No. of Forms	Total No. of FT MC	Total No. of FT OE	Total No. of Field Test Items
4	16	2	3	48	6	54
5	16	2	3	48	6	54
6	16	2	3	48	6	54
7	16	2	3	48	6	54
8	16	2	3	48	6	54
11	16	2	3	48	6	54

More information on the field test designs can be found in specific portions of Chapter Three.

STATISTICAL ANALYSIS OF ITEM DATA

All field-tested items were analyzed statistically following conventional item analysis methods. For MC items, traditional or classical item statistics included the corrected point-biserial correlation (Pt.Bis.) for the correct and incorrect responses (distractors), percent correct (*p*-value), and the percent responding to incorrect responses. For OE items the statistical indices included the item-test correlation, the point-biserial correlation for each score level, percent in each score category or level, and the percent of non-scoreable responses.

In general, more capable students are expected to respond correctly to easy items and less capable students are expected to respond incorrectly to difficult items. If either of these situations does not occur, the item will be reviewed by DRC test development staff and committees of Pennsylvania educators to determine the nature of the problem and the characteristics of the students affected. The primary way of detecting such conditions is through the point-biserial correlation coefficient for dichotomous (MC) items and the item-total correlation for polytomous (OE) items. In each case the statistic will be positive if the total test mean score is higher for the students who respond correctly to MC items (or attain a higher OE item score) and negative when the reverse is true.

Item statistics are used as a means of detecting items that deserve closer scrutiny, rather than being a mechanism for automatic retention or rejection. Toward this end, a set of criteria was used as a screening tool to identify items that needed a closer review by committees of Pennsylvania educators. For an MC item to be flagged, the criteria included any of the following:

- Point-biserial correlation for the correct response of less than 0.25
- Point-biserial correlation for any incorrect response greater than 0.0
- Percent correct less than 0.3 or greater than 0.9
- Percent responding to any incorrect responses greater than the percent correct
- Gender DIF code of either C- or C+
- Any ethnic DIF code of C-

For an OE item to be flagged, the criteria included any of the following:

- Gender DIF code of B-, B+, C- or C+
- Any ethnic DIF code of B- or C-

Item analysis results for MC and OE field-test items are presented in Appendix I.

REVIEW OF ITEMS WITH DATA

In the preceding section on Statistical Analysis of Item Data, it was stated that test development content-area specialists used certain statistics from item and DIF analyses of the 2009 field test to identify items for further review. Specific flagging criteria for this purpose were specified in the previous section. Due to the PSSA-M program being in its initial stages, however, it was determined that all of the PSSA-M mathematics items, both multiple-choice and open-ended, would be brought to the data review for approval.

The review of the items with data was conducted by over 70 Pennsylvania educators (teachers and PDE staff) broken out into grade-level committees. The review took place on August 5, 2009. In this session, committee members were first trained by a representative from DRC's psychometrics staff with regard to the statistical indices used in item evaluation. This was followed by a discussion with examples concerning reasons that an item might be retained regardless of the statistics. The committee review process involved a brief exploration of possible reasons for the statistical profile of an item (e.g., possible bias, grade appropriateness, and instructional issues) and a decision regarding acceptance. DRC and WestEd content-area test development specialists facilitated the review of the items. Each committee reviewed the pool of field test items and made recommendations on each item. Further discussion on how this information was used is covered in Chapter Six.

Table 5–2. 2009 Data Review Committee Results

Assessment	Grade	No. of Items in 2009 Field Test	Field Test Items Examined at 2009 Data Review Committee			Field Test Items Rejected by 2009 Data Review Committee*			Items Classified as Rejected from 2009 Field Test (all sources)**		
			No. of		% of FT	No. of		% of FT	No. of		% of FT
			MC	OE		MC	OE		MC	OE	
	4	54	48	6	100%	0	0	0%	0	0	0%
	5	54	48	6	100%	1	0	1.9%	1	0	1.9%
	6	54	48	6	100%	3	0	5.6%	3	0	5.6%
	7	54	48	6	100%	0	0	0%	0	0	0%
	8	54	48	6	100%	0	0	0%	0	0	0%
	11	54	48	6	100%	5	1	11.1%	5	1	11.1%
Total		**324**	**288**	**36**	**100%**	**9**	**1**	**3%**	**9**	**1**	**3%**

*Rejected as result of statistics

**Data Review Committee, PDE, and DRC

Chapter Six: Operational Forms Construction for 2010

FINAL SELECTION OF ITEMS AND 2010 PSSA-M FORMS CONSTRUCTION

When the final selection of items for the operational 2010 test was ready to begin, the candidate items that emerged from the spring 2009 field test had undergone multiple reviews, including:

- Reviews by DRC and WestEd content-area test development specialists and curriculum specialists
- Formal bias, fairness, and sensitivity review by the Bias, Fairness, and Sensitivity Committee consisting of an expert, multi-ethnic group of men and women with members also having expertise with special needs students and English Language Learners
- Formal review by the content committees consisting of Pennsylvania educators, including teachers as well as district personnel
- PDE review
- Item data review by members of the PDE subject-area teacher committees

The end product of the above process was an item status designation for each field tested item. All items having an item status code of Acceptable/Active were candidates to be selected for the 2010 PSSA-M. To have an item status code of Acceptable/Active meant that the item met the following criteria:

- Appropriately aligned with its designated Assessment Anchor Content Standard (Assessment Anchor) and sub-classifications
- Acceptable in terms of bias/fairness/sensitivity issues, including differential item functioning (for gender and race)
- Free of psychometric flaws, including a special review of flagged items

Next, all relevant information regarding the acceptable items, including associated graphics, was entered into the item banking system known as IDEAS (Item Development and Education Assessment System). From IDEAS and other database sources, Excel files were created for each content area at each grade. These files contained all relevant content codes and statistical characteristics. IDEAS also created a card displaying each acceptable item, any associated graphic, and all relevant content codes and item statistics for use by the content-area test development specialists and psychometric services staff.

DRC test development specialists reviewed the test design blueprint, including the number of items per strand for each content-area test.

Psychometricians provided content-area test development specialists with an overview of the psychometric guidelines for forms construction, including guidelines for selecting linking items to link to previous test forms.

Senior DRC content-area test development specialists reviewed all items in the operational pool to make an initial selection for common (core) positions according to test blueprint requirements and psychometric guidelines. Changes to items were not encouraged since alterations could affect how an item performs on subsequent testing.

For the common items, this meant that the combination of MC and OE items would yield the appropriate range of points while tapping an appropriate variety of the Assessment Anchors and related Eligible Content within each Reporting Category. Items selected in the first round were examined with regard to how well they went together as a set. Of particular concern were the following:

- One item providing cues as to the correct answer to another item
- Context redundancy (e.g., mathematics items with a sports context)
- Presence of clang (distractors not unique from one another)
- Diversity of names and artwork for gender and ethnicity

The first round of items was then evaluated for statistical features such as an acceptable point-biserial correlation and whether correct answers were distributed equally—that is, whether approximately 25 percent of correct answers appeared in each of the four possible positions (A, B, C, or D). Selected items that were deemed psychometrically less advantageous in contrast to the overall psychometric characteristics of the core resulted in a search by the senior reviewer for suitable replacements. At this point, the second round of items was analyzed. If necessary, this iterative process between content-based selections and statistical properties continued in an effort to reach the best possible balance.

Once the recommendations were finalized for the common/core items, they were submitted to PDE for review. Department staff provided feedback, which could be in the form of approval or recommendations for replacing certain items. Any item replacement was accomplished by the collective effort of the test development specialists, psychometricians, and PDE staff until final PDE approval.

LINKING THE 2010 OPERATIONAL TO THE 2011 OPERATIONAL

The 2010 Operational PSSA-M Mathematics test will be linked with the 2011 Operational PSSA-M Mathematics test using core-to-core linking items (items that are repeated from one operational form to another).

In the selection of the core-to-core linking items (part of the overall core pull), content considerations will remain relevant, together with statistical features, such as an acceptable point-biserial correlation and whether the items, as a collection, had an average logit value and a test characteristic curve approximating that of the previous administration.

SPECIAL FORMS USED IN THE 2010 PSSA-M

Braille and Large Print

Students with visual impairments were able to respond to test materials that were available in either Braille or large print. At each grade level assessed, one form was selected for the creation of a Braille and a large print edition. School district personnel ordered Braille or large print assessment materials directly from DRC. They could also contact the Pennsylvania Training and Technical Assistance Network (PaTTAN) for technical assistance regarding students with visual impairments.

School personnel were directed to transcribe all student answers (for MC and OE items) into scannable answer documents exactly as the student responded. No alterations or corrections of student work were permitted, and the answer document had to have the identical numerical form designation.

Spanish Translation of the Mathematics Assessments

School personnel had the option of having Spanish-speaking students who had been enrolled in schools in the United States for less than three years respond to a Spanish version of the PSSA for mathematics only. The original translation of the items and the *Directions for Administration Manual* was initiated by Second Language Testing Incorporated and completed by DRC. After discussions with PDE and Second Language Testing Incorporated, the mathematics booklets for Grades 4, 5, 6, 7, 8, and 11 were designed with a modified over/under format, with the Spanish presented directly above or to the left of the English. To assist the presentation of the two languages on the same page, the English portion was presented in italics and in a smaller font. Those students using this accommodated version of the mathematics assessment could write their answers in English, Spanish, or a combination of both Spanish and English, with the highest possible score from those combinations recorded for the student.

Spanish translated versions of the PSSA-M mathematics assessment were used by a total of 17 students at Grades 4, 5, 6, 7, 8, and 11 in 2010.

Instructions for the appropriate use of these special forms are detailed in the *PSSA & PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans* (PDE, January 2010).

This document can be accessed by going to www.education.state.pa.us. On the left, click on "Programs," then "Programs O–R," then "Pennsylvania System of School Assessment (PSSA)" and then "Testing Accommodations & Security."

Chapter Seven: Test Administration Procedures

Test Sessions, Test Sections, Test Timing, and Test Layout

The PSSA-M Mathematics test utilizes a single consumable booklet. When a single scannable answer booklet is used, the contents of the answer booklet and the test booklet are combined into one integrated booklet. This organization allows the students who are taking the modified tests to maintain the flow and directions of the test without having to manage two separate booklets.

The 2010 PSSA-M tests consisted of two untimed sections. Testing-time recommendations were given, but the estimated times were meant to provide a general guideline for timing rather than absolute testing times.

Table 7–1. PSSA-M Mathematics Test Section Information

Grade	No. of Sections per Test	No. of MC items Section 1	No. of OE items Section 1	No. of MC items Section 2	No. of OE items Section 2	Primary Testing Window	Make-up Testing Window
4	2	15	2	23	1	April 7–16	April 19–May 7
5	2	15	2	23	1	April 7–16	April 19–May 7
6	2	15	2	23	1	April 7–16	April 19–May 7
7	2	15	2	23	1	April 7–16	April 19–May 7
8	2	15	2	23	1	April 7–16	April 19–May 7
11	2	15	2	23	1	April 7–16	April 19–May 7

Table 7–2. PSSA-M Mathematics Duration and Testing Load by Grade

Assessment	Grade	Total No. of MC Items per Form per Admin	Total No. of OE Items per Form per Admin	Total Estimated Administration Time per Form (in Minutes)
Mathematics	**4**	38	3	165 to 190
	5	38	3	165 to 190
	6	38	3	165 to 190
	7	38	3	165 to 190
	8	38	3	165 to 190
	11	38	3	165 to 190

Test administrators were instructed that each section in a form should be scheduled as one assessment session. In addition, they were also told to not combine multiple sections into a single session. Test administrators were also instructed to administer the sections in the sequence in which they were printed in the booklets. In all cases, individual assessment sections had to be completed within one school day.

Test administrators were advised to use a testing location that was separate from the administration of the general PSSA Reading and Math assessment. For 2010, students who participated in the PSSA-M assessment had to also participate in the general PSSA Reading assessment. Students were allowed to complete both sections of the PSSA-M math before completing the three general PSSA reading sections. Alternating the PSSA-M math sections with the PSSA reading sections was also an option for the test administrators, as long as the subject sections were administered in the sequence of the booklet.

Since not all students would finish the assessment sections at the same time, test administrators were advised to use the flexibility of the time limits to the students' advantage. For example, test administrators managed the testing time so that students did not feel rushed while they were taking any assessment section, and no student was penalized because he or she worked slowly. It was equally stressed to test administrators that a student should not be given an opportunity to waste time. Students were told to close their booklets when they finished the section of the assessment in which they had been working. Students who finished early were allowed to sit quietly or read for pleasure until all students had finished. Students with special requirements and/or abilities (i.e., physical, visual, auditory, or learning disabilities as defined by their IEP or service contracts) and students who just worked slowly may have required extended time. Special assessment situations were arranged for these students. When all students in a testing session indicated that they had finished an assessment section, test administrators ended the section and began the next section or allowed the students to return to regular activities.

Scheduled extended time was provided by a test administrator as needed, and students could request extended time if they indicated that they had not completed the task. Such requests were granted if the test administrator found the request to be educationally valid. Test administrators were advised that not permitting ample time for students to complete the assessment would possibly impact the students' and schools' performance.

As a general guideline, however, when all students indicated that they had finished a section, that section was closed. Students requiring time beyond the majority of the student population were allowed to continue immediately following the regularly scheduled session in another setting. When such accommodations were made, school personnel ensured that students were monitored at all times to prevent sharing of information. Students were not permitted to continue a section of the assessment after a significant lapse of time from the original session.

For PSSA-M Mathematics at Grades 7, 8, and 11, test administrators were asked to print out and distribute a copy of the individual grade's formula sheets. The formula sheets were posted at www.education.state.pa.us. [Test administrators were given the following instructions: First click on "Programs" in the left navigation bar, select "Programs O–R," select "Pennsylvania System of School Assessment (PSSA)," and finally click on "Resource Materials." The formula sheets are listed under "Mathematics Resources: 2009–2010."]

Additional information concerning testing time and test layouts can be found in Chapter Three.

TESTING WINDOW

The testing windows for the 2010 PSSA-M operational assessments were as follows:

- Primary testing window – April 7 through April 16, 2010
- Make-up testing window – April 19 through May 7, 2010

Additional information concerning testing time and test layouts can be found in Chapter Three.

SHIPPING, PACKAGING, AND DELIVERY OF MATERIALS

There were two shipments sent out by DRC for the 2010 PSSA-M operational assessment:

- Shipment one contained the *Handbook for Assessment Coordinators* and the *Directions for Administration Manual*s, for each grade tested at a school participating in the mathematics assessment. Shipment one was delivered by March 10, 2010.
- Shipment two contained the administrative materials (e.g., Return Shipping labels, District/School labels, Do Not Score labels, and Student Precode labels) and secure materials (e.g., consumable test/answer booklets) for each grade tested at a school participating in the mathematics assessment. Shipment two was delivered by March 24, 2010.

DRC ensured that all assessment materials were assembled correctly prior to shipping. DRC operations staff used the automated Operations Materials Management System (Ops MMS) to assign secure materials to a school at the time of ship out. This system used barcode technology to provide an automated quality check between items requested for a site and items shipped to a site. A shipment box manifest was produced for and placed in each box shipped. DRC operations staff double-checked all box contents with the box manifest prior to the box being sealed for shipment to ensure accurate delivery of materials. DRC operations staff performed lot acceptance sampling on both shipments. Districts and schools were selected at random and examined for correct and complete packaging and labeling. This sampling represented a minimum of 10 percent of all shipping sites.

DRC's materials management system, along with the systems of shippers, allowed DRC to track materials from DRC's warehouse facility to receipt at the district, school, or testing site. All DRC shipping facilities, materials processing facilities, and storage facilities are secure. Access is restricted by security code. Non-DRC personnel are escorted by a DRC employee at all times. Only DRC inventory control personnel have access to stored secure materials. DRC employees are trained in and made aware of the high level of security that is required.

DRC packed 55,973 modified assessment booklets and 20,055 modified mathematics *Directions for Administration Manual*s for 3382 testing sites. DRC used United Parcel Service (UPS) and Advanced Shipping Technologies to deliver the secure materials to the testing sites.

MATERIALS RETURNED

DRC used UPS for all returns. The materials return windows for the PSSA-M were as follows:

- Primary return window – April 14 through April 20, 2010
- Make-up return window – April 19 through May 7, 2010

TEST SECURITY MEASURES

Test security is essential to obtaining reliable and valid scores for accountability purposes. A test security affidavit was sent to all sites that received PSSA testing material. Every principal or director is to sign and return the test security affidavit with the return of the testing material. The purpose of the affidavit was to serve as a tool to document that the individuals responsible for administering the assessments both understood and acknowledged the importance of test security and accountability. The test security affidavit attested that all security measures were followed concerning the handling of secure materials.

SAMPLE MANUALS

Copies of the *Handbook for Assessment Coordinators* and the *Directions for Administration Manual*s can be found on the PDE website at www.education.state.pa.us.

TESTING WINDOW ASSESSMENT ACCOMMODATIONS

Three accommodations manuals: *PSSA & PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans, Accommodations for English Language Learners, and Accommodations Guidelines for All Students* were developed for use with the 2010 PSSA-M. Additional information regarding assessment accommodations can be found in Chapter Four of this report.

Chapter Eight: Processing and Scoring

RECEIPT OF MATERIALS

Receipt of PSSA-M test materials began on April 14, 2010 and concluded on May 14, 2010, with all make-up tests. DRC's Operations Materials Management System (Ops MMS) was utilized to receive assessment materials securely, accurately, and efficiently. This system features innovative automation and advanced barcode scanners. Captured data were organized into reports, which provided timely information with respect to suspected missing material.

The first step in the Ops MMS was the Box Receipt System. When a shipment arrived at DRC, the boxes were removed from the carrier's truck and passed under a barcode reader, which read the barcode printed on the return label and identified the district and school. If the label could not be read automatically, a floor operator entered the information into the system manually. The data collected in this process were stored in the Ops MMS database. After the barcode data were captured, the boxes were placed on a pallet and assigned a corresponding pallet number.

Once the box receipt process was completed, the materials separation phase began. Warehouse personnel opened the boxes and materials were sorted by grade and status (used or unused answer booklet) into new boxes. Once filled, a sorted box's documents were loaded into an automated counter, which recorded a booklet count for each box. An on-demand DRC box label was produced that contained a description of each box's contents and quantity in both barcode and human-readable format. This count remained correlated to the box as an essential quality control step throughout secure booklet processing and provided a target number for all steps of the check-in process.

Once labeled, the sorted and counted boxes proceeded to booklet check-in. This system used streamfeeder automation to carry documents past oscillating scanners that captured data from up to two representative barcodes and stored it in the Ops MMS database.

The secure booklet check-in operator used a hand scanner to scan the counted box label. This procedure identified the material type and quantity parameters for what the Ops MMS should expect within a box. The box's contents were then loaded into the streamfeeder.

The documents were fed past oscillating scanners that captured both the security code and precode from the booklets. A human operator monitored an Ops MMS screen, which displayed scan errors, an ordered accounting of what was successfully scanned, and the document count for each box.

When all materials were scanned and the correct document count was reached, the box was sealed and placed on a pallet. If the correct document count was not reached, or if the operator encountered difficulties with material scanning, the box and its contents were delivered to an exception handling station for resolution.

This check-in process occurred immediately upon receipt of materials; therefore, DRC provided feedback to districts and schools regarding any missing materials based on actual receipts versus expected receipts. Sites that had 100 percent of their materials missing after the date they were due to DRC were contacted and any issues were resolved.

Throughout the process of secure booklet check-in, DRC project management ran a daily missing materials report. Every site that was missing any number of booklets was contacted by DRC. Results of these correspondences were recorded for inclusion in a final Missing Materials Report if the missing booklets were not recorded by the testing site. DRC produced the Missing

Materials Report for PDE upon completion of secure booklet check-in. The report listed all schools in each participating district along with security barcodes for any booklets not returned to DRC.

After scannable materials (used booklets) were processed through booklet check-in, the materials became available to the DRC Document Processing Center Log-in staff for document log-in. The booklets were logged-in using the following process:

- A DRC scannable barcode batch header was scanned, and a batch number was assigned to each box of booklets.
- The DRC box label barcode was scanned into the system to link the box and booklets to the newly created batch and to create a Batch Control Sheet.
- The DRC box label barcode number, along with the number of booklets in the box, was printed on the Batch Control Sheet for document tracking purposes. All documents that were linked to the box barcode were assigned to the batch number and tracked through all processing steps. As documents were processed, DRC staff dated and initialed the Batch Control Sheet to indicate that proper processing and controls were observed.

Before the booklets were scanned, all batches went through a quality inspection to ensure batch integrity and correct document placement.

After a quality check in the DRC Document Processing Center log-in area, the spines were cut off the scannable documents, and the pages were sent to DRC's Imaging and Scoring System.

SCANNING OF MATERIALS

Customized scanning programs for all scannable documents were prepared to read the booklets and to format the scanned information electronically. Before materials arrived, all image scanning programs went through a quality review process that included scanning of mock data from production booklets to ensure proper data collection.

DRC's image scanners were calibrated using a standard deck of scannable pages with 16 known levels of gray. On a predefined page location, the average pixel darkness was compared to the standard calibration to determine the level of gray. Marks with an average darkness level of 4 or above on a scale of 16 (0 through F) were determined to be valid responses, per industry standards. If multiple marks were read for a single item and the difference of the grayscale reads was greater than four levels, the lighter mark was discarded. If the multiple marks had fewer than four levels of grayscale difference, the response was flagged systematically and forwarded to an editor for resolution.

DRC's image scanners read selected-response, demographic, and identification information. The image scanners also used barcode readers to read pre-printed barcodes from a label on the booklet.

The scannable documents were automatically fed into the image scanners where pre-defined processing criteria determined which fields were to be captured electronically. Constructed response images were separated out for image-based scoring.

During scanning, a unique serial number was printed on each sheet of paper. This serial number was used for document integrity and to maintain sequencing within a batch of answer documents.

A monitor randomly displayed images, and the human operator adjusted or cleaned the scanner when the scanned image did not meet DRC's strict quality standards for image clarity.

All images passed through a process and a software clean-up program that despeckled, deskewed, and desmeared the images. A random sample of images was reviewed for image quality approval. If any document failed to meet image quality standards, the document was returned for rescanning.

Page scan verification was performed to ensure that all pre-defined portions of the booklets were represented in their entirety in the image files. If a page was missing, the entire booklet was flagged for resolution.

After each batch was scanned, booklets were processed through a computer-based editing program to detect potential errors as a result of smudges, multiple marks, and omits in predetermined fields. Marks that did not meet the pre-defined editing standards were routed to editors for resolution.

Experienced DRC Document Processing Center editing staff reviewed all potential errors detected during scanning and made necessary corrections to the data file. The imaging system displayed each suspected error. The editing staff then inspected the image and made any needed corrections using the unique serial number printed on the document during scanning.

Upon completion of editing, quality control reports were run to ensure that all detected potential errors were reviewed again and a final disposition was determined.

Before batches of booklets were extracted for scoring, a final edit was performed to ensure that all requirements for final processing were met. If a batch contained errors, it was flagged for further review before being extracted for scoring and reporting.

During this processing step, the actual number of documents scanned was compared to the number of booklets assigned to the box during book receipt. Count discrepancies between book receipt and booklets scanned were resolved at this time.

Once all requirements for final processing were met, the batch was released for scoring and student level processing.

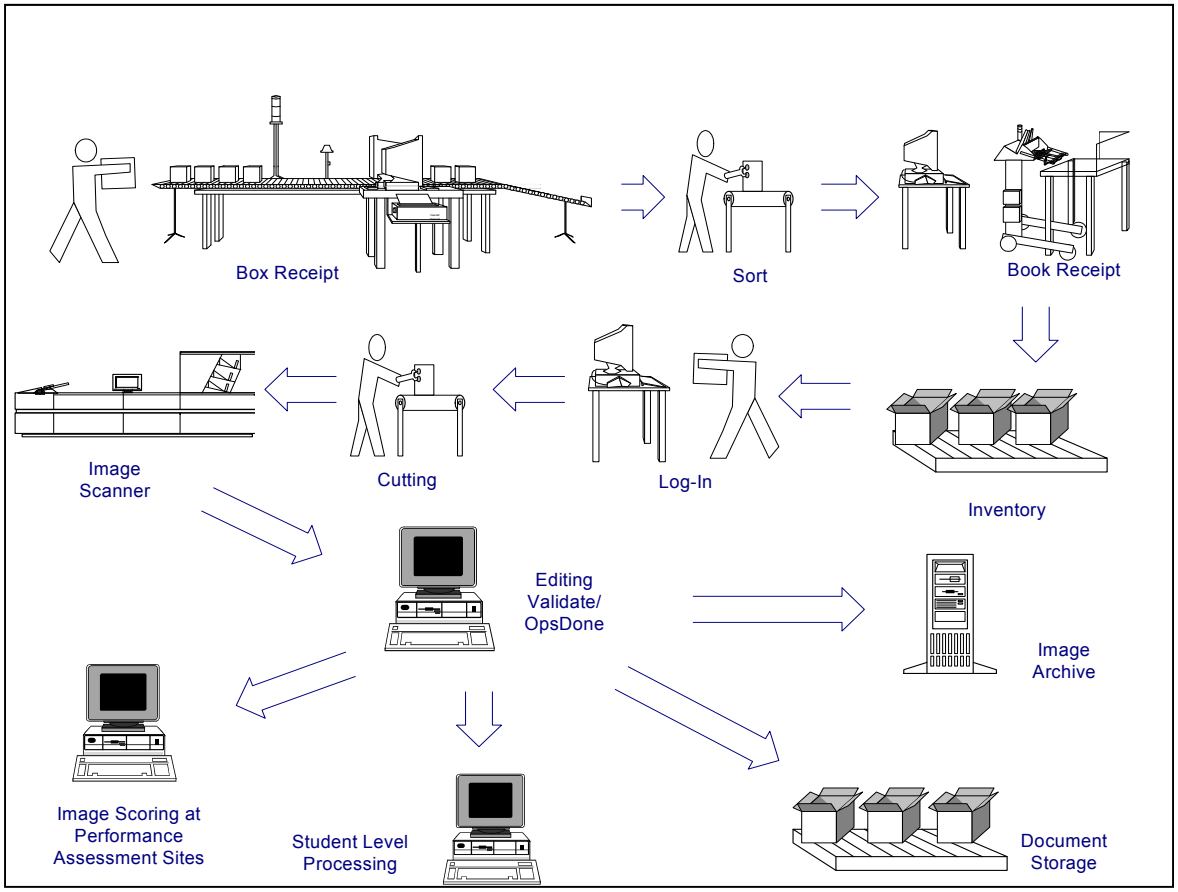
Table 8–1 shows the number of modified booklets received through booklet check-in and the number of modified booklets that contained student responses that were scanned and scored.

Table 8–1. Counts of 2010 PSSA-M Materials Received – Grades 4, 5, 6, 7, 8, and 11

	Booklets Received	Used Booklets Scanned	Total Booklets Shipped
Grade 4 Modified Math	9308	2206	9324
Grade 5 Modified Math	9896	2593	9917
Grade 6 Modified Math	9080	2764	9101
Grade 7 Modified Math	8822	2868	8843
Grade 8 Modified Math	9225	3095	9248
Grade 11 Modified Math	9642	3634	9664

Figure 8–1 illustrates the production workflow for DRC's Ops MMS and Image Scanning and Scoring System from receipt of materials through all processing of materials and the presentation of scanned images for scoring.

Figure 8–1. Workflow System



MATERIALS STORAGE

Upon completion of processing, student response documents were boxed for security purposes and final storage:

- Project-specific box labels were created containing unique customer and project information, material type, batch number, pallet/box number, and the number of boxes for a given batch.
- Boxes were stacked on pallets that were labeled with the project information and a list of the pallet's contents before delivery to the Materials Distribution Center for final secure storage.
- Materials will be destroyed one year after contract year ends with PDE written approval.

SCORING MULTIPLE-CHOICE ITEMS

The scoring process included the scoring of multiple-choice items against the answer key and the aggregation of raw scores from the constructed responses. A student's raw score is the actual number of points achieved by the student for tested elements of an assessment. From the raw scores, the scale scores were calculated.

The student file was scored against the finalized and approved multiple-choice answer key. Items were scored as right, wrong, omitted, or double-gridded (more than one answer was bubbled for an item). Sections of the test were evaluated as a whole and an attempt status was determined for each student for each subject. The score program defined all data elements at the student level for reporting.

RANGEFINDING

After student answer documents were received and processed, DRC's Performance Assessment Services (PAS) staff assembled groups of responses that exemplified the different score points represented in the 0–4 item-specific scoring guidelines for modified mathematics.

Examples of student essays were identified from the operational writing assessment and student responses were also pulled for the new 2010 field test items. Once examples for all score points were identified, sets were put together for each item, and copies were made for each rangefinding participant. Rangefinding committees consisted of Pennsylvania educators, PDE staff members, DRC Test Development staff, and DRC Performance Assessment Services staff. The Modified Math Rangefinding Meeting was held on July 7, 2010 at the Hilton, Harrisburg.

Rangefinding meetings began in a joint session with a review of the history of the 2010 assessment and then broke into grade-level groups. Copies of student responses were presented to the committees, one item at a time. The committees initially reviewed and scored the student samples together to ensure that everyone was interpreting the scoring guidelines consistently. Committee members then went on to score responses independently, and those scores were discussed until a consensus was reached. Only responses for which a high agreement rate among committee members was attained were chosen as training materials for DRC readers. Discussions of responses included the mandatory use of scoring guideline language, assuring PDE and all involved that the score point examples clearly illustrated the specific requirements of each score level. DRC PAS staff made notes of how and why the committees arrived at score

point decisions, and this information was used by the individual scoring directors in reader training.

DRC and PDE discussed scoring guideline edits that the committees suggested. Changes approved by PDE were then made by DRC Test Development staff, and the scoring guidelines were used by PAS staff in the preparation of materials and training of readers.

READER RECRUITMENT/QUALIFICATIONS

DRC retains a number of readers from year to year. This pool of experienced readers was used to staff the approximate 2,800 readers who were needed for scoring the 2010 PSSA, including the modified mathematics. To complete the reader staffing for this project, DRC placed advertisements in local papers and also utilized a variety of websites. Open houses were held and applications for reader positions were screened by the DRC recruiting staff. Candidates were personally interviewed and a mandatory, on-demand writing sample and a mathematics sample were collected, along with references and proof of a four-year college degree. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing expertise in mathematics. Since readers had to have a strong content-specific background, the reader pool consisted of educators and other professionals who were valued for their experience, but who were also required to set aside their own biases about student performance and accept the scoring standards.

LEADERSHIP RECRUITMENT/QUALIFICATIONS

Scoring directors and team leaders were chosen by the content specialists from a pool consisting of experienced individuals who were successful readers and leaders on previous DRC contracts and had strong backgrounds in scoring mathematics. Those selected demonstrated organization, leadership, and management skills. A majority of the scoring directors and team leaders had at least five years of leadership experience on large-scale assessments, including the PSSA. All scoring directors, team leaders, and readers were required to sign confidentiality agreements before any training or handling of secure materials began.

Each room of readers was assigned a scoring director. This individual was monitored by the project director and project content specialist and led the handscoring for the duration of the project. The scoring director assisted in rangefinding, worked with supervisors to create training materials, conducted the team leader training, and was responsible for training the readers. The scoring director also made sure that reports were available and interpreted reports for the readers. The scoring director supervised the team leaders.

Team leaders assisted the scoring director with reader training and monitoring by working with their teams in small group discussions and answering individual questions that readers may not have felt comfortable asking in a large group. Once readers had qualified, the team leaders were responsible for maintaining the accuracy and workload of team members. The ongoing monitoring identified those readers who were having difficulty scoring accurately and resulted in the reader receiving one-on-one retraining or being paired with a stronger reader. This process corrected any inaccuracies in scoring or, if not, that reader was released from the project and any responses they scored were rescored by other readers.

TRAINING

As part of the training for the modified mathematics common items for 2010, DRC's PAS staff assembled the approved scoring guidelines and the scored student responses from training materials that were identified by rangefinding committees, into sets used for training the readers. The same process was used to assemble field test training materials upon completion of the Field Test Rangefinding. Responses that were relevant in terms of the scoring concepts they illustrated were annotated for use in an anchor set. The item-specific scoring guidelines for modified mathematics served as the readers' constant reference.

Training and qualifying sets consisted of examples of student responses reviewed by the rangefinding committees. Responses were selected to show the readers the range of each score point, such as, high, mid, and low 4, 3, 2, 1. Examples of 0s were included as well. Readers were instructed on how to apply the guidelines and were required to demonstrate a clear comprehension of each anchor set by performing well on the training materials that were presented for each grade and item. This helped to train the readers to recognize the various ways that a student could respond to earn the score point that was outlined and defined in the item specific scoring guidelines. All ranges of score points were represented and clearly annotated in the Anchor Set, which was used for reference by the readers throughout the scoring of the project.

The scoring director conducted the team leader training before the reader training. This training followed the same procedures as the reader training, but qualifying standards were more stringent because of the responsibilities required of the team leaders. During team leader training, all materials were reviewed and discussed, and anticipated reader questions and concerns were addressed. Team leaders were required to annotate all of their training responses with the official annotations received from the content committee members at the rangefinding meetings. To facilitate scoring consistency, it was imperative that each team leader imparted the same rationale for each response that other team leaders used. Once the team leaders qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader for each 8–10 readers ensured adequate monitoring of the readers.

The 2010 assessment included the opportunity for students to respond in Spanish to the Grades 4–8 and 11 modified mathematics items. The scoring director responsible for this was a bilingual Hispanic with a strong mathematics background who had also worked with the PSSA for over ten years. Everyone who read these responses was bilingual and hired specifically to score the Spanish portion of the assessment. They were required to meet the same training and scoring standards that were set for the readers of the English version of the assessment.

Reader training began with the scoring director providing an intensive review of the scoring guidelines and anchor papers to all readers. Next, the readers practiced by independently scoring the responses in the training sets. Afterwards, the scoring director or team leaders led a thorough discussion of each set in either a room-wide or small-group setting.

Once the scoring guidelines, anchor sets, and all the training sets were discussed, readers were required to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. The true scores were those assigned by the rangefinding committees and agreed upon by the content staff of PDE and DRC. If readers could not accept or recognize the rationale and purpose for assigning the correct score and did not perform accordingly on the training materials, it was determined that they did not understand the parameters of scoring correctly. Readers who failed to achieve the 70 percent level of exact

agreement were given additional training to acquire the highest degree of accuracy possible. Readers who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score any student responses. These readers were removed from the pool of potential scorers in DRC's imaging system and released from the project.

HANDSCORING PROCESS

Student responses were scored independently and by multiple readers. All responses were read once with 10 percent double read, and additional read behinds above the 10 percent double reads were done to ensure reliability. The data collected from this 10 percent double read was used to calculate the exact and adjacent agreement rates in the Scoring Summary Reports. The responses that were used for the 10 percent read behind were randomly chosen by the imaging system at the item level.

Readers scored the imaged student responses on PC monitors at the DRC Scoring Centers in Columbus, Ohio and Plymouth and Woodbury, Minnesota. Readers were seated at tables with two imaging stations at each table. Image distribution was controlled, thus ensuring that student images were sent to designated groups of readers qualified to score those items. Imaged student responses were electronically separated for routing to individual readers by item, and readers were only provided with student responses that they were qualified to score. Readers read each response and keyed in the score.

To handle possible alerts (i.e., student responses indicating potential issues related to the student's safety and well-being that may require attention at the state or local level), the imaging system allows readers to forward responses needing attention to the scoring director. These alerts are reviewed by the project director, who then notifies the students' schools and PDE of the occurrences. However, PDE does not receive the students' responses or any other identifying information on the students. Also, at no time does the reader know anything about the students' personal identities. There were no alerted student responses during the scoring of the 2010 modified mathematics.

Once handscoring was completed, PAS compiled anecdotal reviews of the field test items for all grade levels. This information was presented to DRC's Test Development group.

HANDSCORING VALIDITY PROCESS

One of the quality/training tools PAS utilized to ensure reader accuracy was the validity process. The goal of the validity process is to ensure that scoring standards are maintained. Specifically, the objective is to make sure that scorers rate student responses in a manner consistent with statewide standards both within a single administration of the PSSA modified mathematics and across consecutive administrations. This scoring consistency was maintained, in part, through the validity process.

The validity process began with the selection of scored responses from the initial field test. The content specialist for modified mathematics selected 40 validity papers for each core constructed response (CR) item. These 40 papers were drawn from a pool of exemplars (responses that are representative of a score point and have been verified by the scoring director and the content specialist). The scores on validity papers are considered criterion or true scores.

The validity papers were then implemented to test reader accuracy. The responses were scanned into the imaging system and dispersed intermittently to the readers. By the end of the project, readers had scored all 40 validity papers for any items they were qualified to score. Readers were unaware that they were being fed pre-scored responses and assumed that they were scoring live student responses. This helped bolster the internal validity of the process. It is important to note that all readers who received validity papers had already successfully completed the training/qualifying process.

Next, the scores that the readers assigned to the validity papers were compared to the true scores in order to determine the validity of the readers' scores. For each item, the percentage of exact agreement as well as the percentage of high and low scores was computed. This data was accessed through the Validity Item Detail Report. The same kind of data was also computed for each specific reader. This data was accessed through the Validity Reader Detail Report. Both of these can be run as a daily or cumulative report.

The Validity Reader Detail Report was used to identify particular readers for retraining. If a reader on a certain day generated a lower rate of agreement on a group of validity papers, it was immediately apparent in the Validity Reader Detail Report. A lower rate of agreement was defined as anything below 70 percent exact agreement with the true scores. Anytime a reader's validity agreement rate fell below 70 percent, this cued the scoring director to examine that reader's scoring. First, the scoring director tried to figure out what kind of validity papers the reader was scoring incorrectly. This was done to determine if there was any kind of a trend (e.g., tending to go low on the 1–2 line). Once the source of the low agreement was determined, the reader was retrained. If it was determined that the reader had been scoring live papers inaccurately, then his/her scores were purged for that day and the responses were re-circulated and scored by other readers.

The cumulative Validity Item Detail Report was utilized to identify potential room-wide trends in need of correction. For instance, if a particular validity response with a true score of three was given a score of two by a significant number of readers, that trend would be revealed in the Validity Item Detail Report. To correct a trend of this sort, the scoring director would look for student responses similar to the validity paper being scored incorrectly. Once located, these responses would be used in room-wide training, usually in the form of an annotated handout or a short set of papers without printed scores given to readers as a recalibration test.

Validity was employed on all core modified mathematics CR items. Each 40 paper validity set was formulated to mimic the score point distribution that the item generated during its previous administration. Each validity set included at least five examples of each score point. Examples of different types of responses were included to ensure that readers were tested on the full spectrum of response types.

The exact reader agreement rate generated during the validity process was often higher than the inter-rater agreement rate for the same item. The reason for this discrepancy has to do with how validity sets are formulated. The 40 validity papers for each item, chosen by the content specialist, are intended to cover the full breadth of each score point. For example, each validity set contains examples of high, low and middle twos. This scope ensures that the validity process is truly valid in terms of addressing the complete spectrum of response types. However, certain types of responses are generally not included in validity sets. These include line papers (i.e., examples of score points that are so close to the adjacent score point that readers are instructed to consult with a supervisor before assigning a score) and responses that, because of poor word

choice/writing, are difficult to understand. The reason for these exclusions is that confusing/line/illegible papers often do not impart a teachable lesson. Since these types of papers are usually unique, any potential lesson the paper might teach would apply only to that specific response. Conversely, the papers in validity sets are chosen because they represent common response-types and teach lessons that can be applied to other similar papers. Due to this distinction, validity sets generate a slightly higher agreement rate than is normally generated during operational scoring. PAS could change this practice in the future by including some truly line papers in validity sets. However, this change might also have the unintended effect of confusing the training process. Since line papers are already included in the training sets given to readers prior to qualifying, putting them into validity sets might complicate the process unnecessarily.

QUALITY CONTROL

Reader accuracy was monitored throughout the scoring session by producing both daily and on-demand reports, ensuring that an acceptable level of scoring accuracy was maintained. Inter-reader reliability was tracked and monitored with multiple quality control reports that were reviewed by quality assurance analysts. These reports and other quality control documents were generated at the handscoring centers and were reviewed by the scoring directors, team leaders, content specialists, and project directors. The following reports and documents were used during the scoring of the constructed responses:

The Scoring Summary Report (includes two related reports)

1. **The Reader Monitor Report** monitored how often readers were in exact agreement with one another and ensured that an acceptable agreement rate was maintained. This report provided daily and cumulative exact and adjacent inter-reader agreement on the 10 percent that was double read.
2. **The Score Point Distribution Report** monitored the percentage of responses given each of the score points. For example, the mathematics daily and cumulative reports showed how many 0s, 1s, 2s, 3s, and 4s a reader had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each reader so that production rates could be monitored.

The Item Status Report monitored the progress of handscoring. This report tracked each response and indicated the status (e.g., needs second read or complete). This report ensured that all responses were scored by the end of the project.

The Read-Behind Report identified all responses scored by an individual reader. This report was useful if any responses needed rescoring because of possible reader drift.

The Validity Reports tracked how the readers performed by comparing pre-determined scored responses to readers' scores for the same set of responses. If the readers' scoring fell below the 70 percent determined agreement rate, remediation occurred. Readers who did not retrain to the required level of agreement were released from the project.

The Read-Behind Log was used by the team leader/scoring director to monitor individual reader reliability. Student responses were randomly selected and team leaders read scored items from each team member. If the team leader disagreed with the reader's score, remediation occurred. This proved to be a very effective type of feedback because it was done with live items scored by a particular reader.

Recalibration Sets were used throughout the scoring sessions to ensure accuracy by comparing each reader's scores with the true scores on a preselected set of responses. Recalibration sets helped to refocus readers on Pennsylvania scoring standards. This check made sure there was no change in the scoring pattern as the project progressed. Readers failing to achieve 70 percent agreement with the recalibration true scores were given additional training to achieve the highest degree of accuracy possible. Readers who were unable to recalibrate were released from the project. The procedure for creating and administering recalibration sets was similar to the one used for training sets.

Table 8–2 shows exact and adjacent agreement rates of readers on the core constructed responses for the modified mathematics items in the 2010 PSSA. All student responses were read once with a 10 percent double read. The data collected from this 10 percent double read was used to calculate the exact and adjacent agreement rates.

Table 8–2. Inter-rater Agreement for 2010 PSSA Modified Mathematics Grades 4–8 & 11 Constructed Response Items and Validity

0–4 possible score points

Modified Mathematics	Common Item	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent Agreement	% Exact Validity Agreement
Grade 4	1	98	2	100	96
	2	100	0	100	92
Grade 5	1	96	4	100	89
	2	97	3	100	97
Grade 6	1	88	12	100	91
	2	99	1	100	91
Grade 7	1	96	4	100	96
	2	94	5	99	98
Grade 8	1	89	11	100	92
	2	87	13	100	98
Grade 11	1	93	7	100	94
	2	95	5	100	93

Table 8–3 shows the distribution of scores for the modified mathematics items. All modified mathematics items are scored with a 0–4 score point range. B=blank and NS=non-scoreable.

Table 8–3. Percentages Awarded for Each Possible Score Point
2010 PSSA Modified Mathematics Grades 4–8 and 11

Modified Mathematics	Common Item	%0	%1	%2	%3	%4	%B/NS
Grade 4	1	34	29	9	12	14	0
	2	31	32	21	9	6	0
Grade 5	1	61	21	8	7	2	1
	2	31	24	11	12	22	1
Grade 6	1	41	39	12	5	3	1
	2	16	54	21	7	1	1
Grade 7	1	26	30	16	17	11	1
	2	74	4	6	7	8	2
Grade 8	1	25	40	24	6	4	1
	2	33	37	16	10	3	2
Grade 11	1	49	35	7	2	3	4
	2	36	49	5	4	0	6

Chapter Nine: Description of Data Sources and Sampling Adequacy

This chapter describes the data sources (e.g., *n*-counts, characteristics of students) used for the various analysis procedures discussed in the remaining chapters of this technical report. Statistical analysis is conducted at several points for the PSSA-M: 1) an early analysis for quality control purposes; 2) analyses associated with the late-stage calibration, scaling, and linking process (e.g., impact results); 3) analyses used for item banking; and, 4) analyses for the technical report. Very detailed information regarding the attributes of students used for AYP reporting is provided in Chapter Ten[2].

PRIMARY STUDENT FILTERING CRITERIA

For many data files, the primary means of filtering students for inclusion/exclusion from any data analysis are based on the state reporting criteria which are outlined below. Within the state reporting rules are separate attempt criteria for individual subject areas. The attempt criteria are discussed more fully below.

State Reporting Criteria

The state reporting criteria are as follows:

- Student must be enrolled for the full academic year.
- Student must be attributed to a public district/school (state).
- Student must receive a score (i.e., met the subject attempt logic—see additional information below).
- Student is not a home school student.
- Student is not a foreign exchange student.
- Student is not a first year ELL student.

PSSA-M ATTEMPT CRITERIA

For all data sources, only students who meet the attempt criteria are included. The attempt criteria required a minimum of four items (multiple-choice or open-ended items) to be completed in each subject area section of the test booklets. Counts were based on operational items only.

[2] This data file was delivered to PDE on July 29, 2010.

KEY VALIDATION DATA

These data are only mentioned for the sake of completeness as no formal results from these data are provided in this technical document. An analysis on all MC items is conducted early in the scoring process to ensure that the items are performing as expected. This is an important quality check that is always done for the PSSA-M. This analysis is usually (but not always) done using all students from early-return schools. The sample does not need to be state representative for these quality checks. Available student data typically suffices as long as there is reasonable variability in the total test scores of students.

For 2010 this data included all public school students who: 1) had their MC items scanned and scored by April 30 and 2) met preliminary attempt criteria (i.e., attempt was determined based on MC items only). Note that the full state reporting criteria were not in effect for this file (only attribution to a public school based on tested site and preliminary attempt criteria were used to filter students).

CALIBRATION DATA

Calibration data included students who met the preliminary state reporting criteria (including attempt criteria) by May 23, 2010. The state reporting criteria were preliminary meaning that attributions and final PIMS information were not complete by this time. No sampling was undertaken in this data (i.e., it included all students who met the above criteria with operational test scores up to this point).

ITEM BANK DATA

The item-bank data included students who met the state reporting criteria, pre-AYP appeals (including attempt criteria) by July 15, 2010. No sampling was undertaken in this data (i.e., it included all students who met the above criteria with scored field test data up to that point). The data banked for field-test items were based on this data file.

FINAL DATA

This file included all students who met state reporting criteria, pre-AYP appeals (including attempt criteria) by July 15[3] for all subject areas. No sampling was undertaken in this data (i.e., it includes all students in the data file). The final data is pre-appeals data, meaning that schools had not yet had an opportunity to correct certain fields within the data during the AYP appeals process (e.g., student ethnicity). The majority of the results included in this technical report were derived using the final data file.

[3] The AYP reporting file was delivered to PDE on July 29, 2010. Most analyses in this report were conducted on stripped-down version of that data file (i.e., some data elements were removed to reduce file size). Hence, the two different file dates.

FINAL *N*-COUNTS FOR ALL DATA SOURCES

The *n*-counts for all data sources are provided in Table 9–1.

Table 9–1. Data Source *N*-Counts

	Key Validation	Calibration	Item Bank	Final
4	1587	2129	2169	2169
5	1919	2514	2551	2552
6	2060	2642	2698	2700
7	2094	2746	2814	2817
8	2240	2944	3012	3019
11	2158	3433	3532	3536

Chapter Ten: Summary Demographic, Program, and Accommodation Data for the 2010 PSSA Modified

ASSESSED STUDENTS

As stated in the Preface, the target population for the PSSA-M consists of those public school students with an IEP and history of low academic achievement whose disabilities inhibit their capacity to respond to the standard PSSA, even with accommodations, but function above the one percent of students with the most severe cognitive impairments who qualify for the Pennsylvania Alternate System of Assessment (PASA).

Eligibility for the PSSA-M requires that a student 1) is not eligible for the PASA, 2) must have a grade-level standards aligned IEP that clearly documents that the student requires significant instructional accommodations to successfully access grade level content, 3) demonstrates persistent academic difficulties with 4) a lack of academic progress. More detailed information on the PSSA-M eligibility criteria may be accessed by going to www.education.state.pa.us. On the left side of the navigation bar, click on "Programs," then "Programs S–Z," then "Special Education." From the "Special Education" page click on "Assessment" to access the relevant documents.

Results for Chapter Ten are presented in tables with a numbering system that includes a letter designating a subject area. For the 2010 PSSA-M Technical Report, each table number contains an "M" for Mathematics as this was the only subject area assessed.

Table 10–1M provides a summary of the assessed students for mathematics. Presented on the first line is the total number of non-blank answer documents processed by grade level for the 2010 PSSA-M. This number pertains to the total number of records on the student file and is typically less than the "Used Booklets Scanned" column shown in Table 8–1. The reason for the difference is that completely blank answer booklets (no student name and no items responded to) get removed from the initial batch of materials scanned. See Chapter Eight for more details on processing. The second line shows the number and percent of students with a PSSA-M score in mathematics, followed by the number and percent not receiving a score. The final line gives the number of students contributing to state summary statistics, which is especially relevant for all tables following 10–2M. (See the section of this chapter entitled "Composition of Sample Used in Subsequent Tables" for additional explanation.)

Table 10–1M. Students Assessed on the 2010 PSSA-M: Mathematics

	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Number of non-blank answer documents processed	2,206	2,590	2,758	2,862	3,092	3,629
Students with mathematics scores	2,203 99.9	2,580 99.6	2,741 99.4	2,853 99.7	3,063 99.1	3,572 98.4
Number processed but not assessed (without a total score)	3 0.1	10 0.4	17 0.6	9 0.3	29 0.9	57 1.6
Students with mathematics scores used in state summaries	2,169	2,552	2,700	2,817	3,019	3,536

As may be observed from Table 10–1M, not all students were assessed. Although there are a variety of reasons for this, the major ones pertained to:

- Extended absence from school that continued beyond the assessment window.
- Being absent without make-up for at least one section of the mathematics test.
- Failure of a student to meet the attempt criteria on one or more mathematics test sections and no exclusion code was marked by school personnel. For mathematics, the attempt criteria required a minimum of four items to be completed in each test section.
- Medical emergency.
- Other reasons (includes parental request due to religious reasons, students who are court-agency placed, students with multiple reasons coded, and the category of other).

The numbers of students without test scores for these reasons are presented in Table 10–2M.

Table 10–2M. Counts of Students without Scores on the 2010 PSSA-M: Mathematics

Reason for Non-Assessment: Mathematics	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Extended absence from school	1 33.3	1 10.0	1 5.9	0 0.0	4 13.8	11 19.3
Absent without make-up	0 0.0	0 0.0	1 5.9	2 22.2	2 6.9	3 5.3
Non-Attempt for mathematics	0 0.0	4 40.0	9 52.9	5 55.6	15 51.7	28 49.1
Medical emergency	1 33.3	4 40.0	4 23.5	2 22.2	5 17.2	5 8.8
Other reasons	1 33.3	1 10.0	2 11.8	0 0.0	3 10.3	10 17.5
Total not assessed for mathematics	3	10	17	9	29	57

COMPOSITION OF SAMPLE USED IN SUBSEQUENT TABLES

Students included in the following demographic analyses were those who contributed to state summary statistics, using the Pre-Appeals (AYP) individual student data file provided to the Pennsylvania Department of Education on July 29, 2010. Students not included in the present state summary data were those who were 1) enrolled in a Pennsylvania school after October 1, 2009, 2) coded as ELL and enrolled after March 27, 2009, 3) a foreign exchange student, 4) home schooled, 5) enrolled in a non-public school, or 6) do not have a mathematics test score.

Demographic data for students taking the PSSA-M for mathematics is presented in Table 10–3M. Results for accommodations received are presented in Tables 10–4M through 10–7M.

COLLECTION OF STUDENT DEMOGRAPHIC INFORMATION

Data for analyses involving demographic characteristics were obtained primarily from information supplied by school district personnel through the Pennsylvania Information Management System (PIMS) and subsequently transmitted to DRC. Updates of attribution data for AYP were carried out through the DRC Attribution System. Some data such as accommodation information is marked directly on the student answer document at the time the PSSA-M is administered.

DEMOGRAPHIC CHARACTERISTICS

Frequency data for each category is presented in Table 10–3M. Percentages are based on students with a score in mathematics, which are shown at the bottom of the table. Included are students receiving education in a non-traditional setting, such as court-agency placed.

Table 10–3M. Demographic Characteristics of Students Taking the 2010 PSSA-M: Mathematics

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Gender						
Female	912 42.0	1,069 41.9	1,076 39.9	1,156 41.0	1,196 39.6	1,332 37.7
Male	1,249 57.6	1,471 57.6	1,616 59.9	1,647 58.5	1,803 59.7	2,163 61.2
Race/Ethnicity						
American Indian or Alaskan Native	4 0.2	6 0.2	7 0.3	3 0.1	8 0.3	5 0.1
Asian or Pacific Islander	27 1.2	19 0.7	31 1.1	36 1.3	22 0.7	27 0.8
Black/African American non-Hispanic	415 19.1	465 18.2	511 18.9	511 18.1	554 18.4	700 19.8
Latino/Hispanic	193 8.9	223 8.7	233 8.6	236 8.4	251 8.3	232 6.6
White non-Hispanic	1,492 68.8	1,783 69.9	1,874 69.4	1,987 70.5	2,138 70.8	2,496 70.6
Multi-Racial/Ethnic	28 1.3	42 1.6	36 1.3	27 1.0	26 0.9	36 1.0

Demographic or Educational Characteristic	**Gr. 4**	**Gr. 5**	**Gr. 6**	**Gr. 7**	**Gr. 8**	**Gr. 11**
	N / Pct	**N / Pct**	**N / Pct**	**N / Pct**	**N / Pct**	**N / Pct**
Educational Category and Other Demographic Groups						
IEP (not gifted)	2,169 100.0	2,552 100.0	2,700 100.0	2,817 100.0	3,019 100.0	3,536 100.0
Student exited IEP in last 2 years	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Title I	524 24.2	613 24.0	548 20.3	425 15.1	451 14.9	358 10.1
Title III Served	60 2.8	65 2.5	76 2.8	48 1.7	57 1.9	12 0.3
Title III Not Served	30 1.4	30 1.2	23 0.9	26 0.9	19 0.6	13 0.4
Migrant Student	1 0.0	3 0.1	11 0.4	3 0.1	5 0.2	1 0.0
ELL (enrolled after 3-27-09)	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
ELL (enrolled before 3-27-09)	90 4.1	95 3.7	99 3.7	74 2.6	76 2.5	25 0.7
Exited ESL/bilingual program and in first year of monitoring	7 0.3	7 0.3	3 0.1	6 0.2	8 0.3	2 0.1
Exited ESL/bilingual program and in second year of monitoring	2 0.1	2 0.1	6 0.2	3 0.1	1 0.0	2 0.1
Former ELL no longer monitored	11 0.5	16 0.6	21 0.8	24 0.9	40 1.3	42 1.2
Economically Disadvantaged	1,291 59.5	1,526 59.8	1,499 55.5	1,584 56.2	1,621 53.7	1,687 47.7
Enrollment						
Current Enrollment in school of residence after 10-1-09	60 2.8	47 1.8	48 1.8	81 2.9	65 2.2	97 2.7
Current Enrollment in district of residence after 10-1-09	33 1.5	21 0.8	24 0.9	44 1.6	29 1.0	41 1.2
Current Enrollment as PA resident after 10-1-09	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Enrolled in school of residence after 10-1-08 but on/before 10-1-09	450 20.7	643 25.2	1,100 40.7	880 31.2	545 18.1	603 17.1
Enrolled in district of residence after 10-1-08 but on/before 10-1-09	239 11.0	278 10.9	292 10.8	306 10.9	335 11.1	389 11.0
Education in Non-Traditional Settings						
Court / agency placed	2 0.1	5 0.2	4 0.1	17 0.6	17 0.6	44 1.2
Students with mathematics scores used in state summaries	2,169	2,552	2,700	2,817	3,019	3,536

TEST ACCOMMODATIONS PROVIDED

School personnel supplied information regarding accommodations that a student may have received while taking the PSSA-M. Accommodations, classified in terms of presentation, response, setting, and timing, enable students to better manage disabilities that hinder their ability to learn and respond to assessments. An accommodations manual for the PSSA entitled *PSSA and PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans* (PDE, Revised 1/11/2010) was developed for use with the 2010 PSSA and PSSA-M. The manual can be accessed by going to www.education.state.pa.us. On the left side of the navigation bar, click on "Programs," then "Programs O–R," then "Pennsylvania System of School Assessment (PSSA)" and then "Testing Accommodations & Security."

The frequency with which these accommodations were utilized is summarized separately for each accommodation category in Tables 10–4M through 10–7M. Table values are based on all scored students who contributed to state summary statistics. Note that a glossary of accommodation terms as applied to the PSSA is provided in Table 10–10 at the end of this chapter.

PRESENTATION ACCOMMODATIONS RECEIVED

Presentation Accommodations are those that provide alternate ways for students to access and process printed instructional material and assessments. These include auditory, tactile, visual, and combined auditory/visual modes of presentation. The number of presentation accommodations provided for mathematics in the 2010 assessment was 12. As depicted in Table 10–4M, the actual frequencies are quite low, generally representing less than five-tenths of one percent of assessed students statewide. The most notable exceptions were test directions read aloud and, test items/questions read aloud.

RESPONSE ACCOMMODATIONS RECEIVED

Response Accommodations permit students to complete assignments, tests, and activities in different ways to solve or organize problems using some type of assistive device or organizer. The number of response accommodations provided for mathematics was 12. The frequency with which these accommodations were utilized is summarized in Table 10–5M. The actual frequencies are quite low, most representing less than one percent of assessed students statewide.

SETTING ACCOMMODATIONS RECEIVED

Setting Accommodations permit a change in location in which a student receives instruction or participates in an assessment. There were four categories of setting accommodations in 2010. As depicted in Table 10–6M, small group testing, and testing in a separate setting were the most commonly used accommodations.

TIMING ACCOMMODATIONS RECEIVED

Timing Accommodations involve a change in the allowable length of time to complete assignments or assessments, including the way in which time is organized. There were four categories of timing accommodations in 2010. As depicted in Table 10–7M, the most common accommodation was scheduled extended time.

Table 10–4M. Incidence of Presentation Accommodations Received on the 2010 PSSA-M: Mathematics

Type of Presentation Accommodation	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Braille Format	0 0.0	0 0.0	1 0.0	0 0.0	1 0.0	1 0.0
Large Print Format	8 0.4	7 0.3	6 0.2	6 0.2	9 0.3	2 0.1
Electronic Screen Reader	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Test directions read aloud (provided by live reader)	931 42.9	908 35.6	752 27.9	601 21.3	644 21.3	450 12.7
Test directions signed, interpreted for ELL student, or recorded	12 0.6	13 0.5	12 0.4	16 0.6	14 0.5	9 0.3
Test items/questions read aloud (provided by live reader) or signed	1,360 62.7	1,457 57.1	1,104 40.9	808 28.7	762 25.2	296 8.4
Test items / questions interpreted for ELL student	13 0.6	8 0.3	6 0.2	9 0.3	13 0.4	4 0.1
Amplification device	9 0.4	2 0.1	2 0.1	1 0.0	0 0.0	1 0.0
Magnification device	3 0.1	1 0.0	2 0.1	0 0.0	1 0.0	0 0.0
Reading windows, reading guides	22 1.0	22 0.9	1 0.0	0 0.0	2 0.1	2 0.1
Other (per *Accommodations Guidelines*)	29 1.3	44 1.7	32 1.2	30 1.1	38 1.3	25 0.7
Spanish version for mathematics	0 0.0	2 0.1	4 0.1	1 0.0	2 0.1	0 0.0

Table 10–5M. Incidence of Response Accommodations Received on the 2010 PSSA-M: Mathematics

Type of Response Accommodations	Gr.4	Gr.5	Gr.6	Gr.7	Gr.8	Gr.11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Test administrator marked multiple-choice responses	51 2.4	31 1.2	22 0.8	20 0.7	9 0.3	8 0.2
Test administrator scribed open-ended	123 5.7	85 3.3	44 1.6	25 0.9	16 0.5	12 0.3
Test administrator transcribed student responses	48 2.2	23 0.9	20 0.7	16 0.6	13 0.4	6 0.2
Qualified interpreter for ELL	0 0.0	0 0.0	0 0.0	0 0.0	1 0.0	1 0.0
Typewriter, word processor or computer	3 0.1	1 0.0	3 0.1	1 0.0	10 0.3	4 0.1
Brailler / Notetaker	0 0.0	0 0.0	0 0.0	0 0.0	1 0.0	0 0.0
Augmentative communication device	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Audio recording of student responses	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Electronic Screen Reader	0 0.0	0 0.0	0 0.0	1 0.0	0 0.0	0 0.0
Manipulative	38 1.8	15 0.6	12 0.4	3 0.1	1 0.0	5 0.1
Translation dictionary for ELL student	0 0.0	0 0.0	3 0.1	1 0.0	0 0.0	0 0.0
Other (approved by PDE)	15 0.7	27 1.1	7 0.3	25 0.9	9 0.3	7 0.2

Table 10–6M. Incidence of Setting Accommodations Received on the 2010 PSSA-M: Mathematics

Type of Timing Accommodation	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Hospital/Home Testing	4 0.2	6 0.2	2 0.1	4 0.1	6 0.2	11 0.3
Separate Setting	991 45.7	971 38.0	745 27.6	772 27.4	813 26.9	663 18.8
Small Group Testing	1,606 74.0	1,821 71.4	1,751 64.9	1,673 59.4	1,829 60.6	1,794 50.7
Other (PDE approved)	21 1.0	19 0.7	17 0.6	27 1.0	21 0.7	19 0.5

Table 10–7M. Incidence of Timing Accommodations Received on the 2010 PSSA-M: Mathematics

Type of Timing Accommodation	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Scheduled Extended Time	589 27.2	570 22.3	463 17.1	417 14.8	421 13.9	518 14.6
Requested Extended Time	46 2.1	56 2.2	68 2.5	89 3.2	113 3.7	187 5.3
Multiple Test Sessions	101 4.7	80 3.1	80 3.0	115 4.1	88 2.9	141 4.0
Changed Test Schedule	46 2.1	43 1.7	35 1.3	52 1.8	18 0.6	43 1.2

ACCOMMODATION RATE

The incidence of students receiving one or more of the 32 accommodations available for mathematics is provided in Table 10–8M. The category of Non-Accommodated indicates that a student did not receive any accommodation during the testing.

The general pattern of findings provided in Table 10–8M reveals a consistently high percent of students receiving an accommodation, which diminished across grade levels, with the exception of Grade 8.

Table 10–8M. Accommodation Rate on the 2010 PSSA-M: Mathematics

Student Subgroup	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct	N / Pct
Non-Accommodated	393 18.1	537 21.0	734 27.2	907 32.2	919 30.4	1,531 43.3
Accommodated	1,776 81.9	2,015 79.0	1,966 72.8	1,910 67.8	2,100 69.6	2,005 56.7
	2,169	2,552	2,700	2,817	3,019	3,536

The Incidence of Accommodations and ELL Status

By definition students qualifying to take the PSSA-M assessment have an IEP along with a history of very low achievement. These students often receive various accommodations to assist them in accessing and responding optimally in assessment situations. As observed in Tables 10–4M through 10–7M, the most frequently occurring accommodations for assessed students were:

- Test directions read aloud
- Test items/questions read aloud or signed (mathematics and science only)
- Tested in separate setting
- Small group testing
- Scheduled extended time

Because the accommodations with the largest frequencies can potentially supply the most stable data when broken out for subgroup analysis, these were selected for display in Table 10–9M. For purposes of this analysis, an English Language Learner (ELL) was a student classified as ELL and enrolled in a U.S. school on or before March 27, 2009. All other assessed students, including those who exited an ESL/bilingual program and in the first or second year of monitoring were regarded as Not ELL. Students coded as ELL and enrolled in a U.S. school after March 27, 2009 are excluded from state summary statistics as stated earlier in this chapter.

A cross-tabulation between each of the selected accommodations and ELL status revealed a nearly equal number of times that one group had a larger percentage receiving an accommodation than the other group. Table 10–9M displays the number and percent of Non-ELL and ELL students receiving five selected accommodations for each of the six grade levels. In comparing the two groups with respect to the percentage of students receiving an accommodation there was little consistency in terms of type of accommodation or grade level. Of the 30 possible comparisons Non-ELL students received the larger percentage of accommodations in 14 instances, ELL students in 12 instances, and in four remaining instances the difference was less than one percent.

Table 10–9M. Incidence of Non-ELL and ELL Students Receiving Selected Accommodations: Mathematics

Accommodation Received	Non-ELL Students		ELL Students	
Gr. 4	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	888	42.7	43	47.8
Mathematics test items/ questions read aloud or signed	1,298	62.4	62	68.9
Tested in separate setting	950	45.7	41	45.6
Small group testing	1,535	73.8	71	78.9
Scheduled extended time	555	26.7	34	37.8
Column N for Gr. 4	2,079		90	
Gr. 5	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	875	35.6	33	34.7
Mathematics test items/ questions read aloud or signed	1,412	57.5	45	47.4
Tested in separate setting	942	38.3	29	30.5
Small group testing	1,756	71.5	65	68.4
Scheduled extended time	551	22.4	19	20.0
Column N for Gr. 5	2,457		95	
Gr. 6	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	721	27.7	31	31.3
Mathematics test items/ questions read aloud or signed	1,060	40.8	44	44.4
Tested in separate setting	724	27.8	21	21.2
Small group testing	1,698	65.3	53	53.5
Scheduled extended time	443	17.0	20	20.2
Column N for Gr. 6	2,601		99	

Accommodation Received	**Non-ELL Students**		**ELL Students**	
Gr. 7	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	593	21.6	8	10.8
Mathematics test items/ questions read aloud or signed	793	28.9	15	20.3
Tested in separate setting	757	27.6	15	20.3
Small group testing	1,641	59.8	32	43.2
Scheduled extended time	406	14.8	11	14.9
Column N for Gr. 7	2,743		74	
Gr. 8	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	632	21.5	12	15.8
Mathematics test items/ questions read aloud or signed	740	25.1	22	28.9
Tested in separate setting	804	27.3	9	11.8
Small group testing	1,786	60.7	43	56.6
Scheduled extended time	413	14.0	8	10.5
Column N for Gr. 8	2,943		76	
Gr. 11	**N**	**Pct**	**N**	**Pct**
Test directions read aloud	444	12.6	6	24.0
Mathematics test items/ questions read aloud or signed	294	8.4	28	8.0
Tested in separate setting	655	18.7	8	32.0
Small group testing	1,778	50.6	16	64.0
Scheduled extended time	513	14.6	5	20.0
Column N for Gr. 11	3,511		25	

GLOSSARY OF ACCOMMODATIONS TERMS

Table 10–10 provides a brief description of accommodations terms as used in the PSSA and PSSA-M. School personnel identified the accommodations that a student received by marking the relevant bubble(s) in the student answer document as noted in the left column. The right column contains an explanation abstracted from the *PSSA and PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans* (PDE, Revised 1/11/2010, pages 24–46).

Table 10–10. Glossary of Accommodations Terms as Applied in the PSSA and PSSA-M

Type of Testing Accommodation	Explanation
Student used the following Presentation Accommodations	
Braille format	Students may use a Braille format of the test. Answers must then be transcribed into the answer booklet without alteration.
Large print format	Students with visual impairments may use a large print format. Answers must then be transcribed into the answer booklet without alteration.
Magnification device	Devices to magnify print may be used for students with visual impairments and/or print disabilities.
Reading windows, reading guides	Students with visual impairments may use reading windows and reading guides in all assessments.
Electronic screen reader (PDE approval required)	Students with a severe visual disability may use an electronic screen reader; however, PDE must approve the program and functions prior to the test window.
Sign language interpreter	Deaf/hearing impaired students may receive test directions from a qualified interpreter. Signing is also permitted for: essay prompts (writing), items/questions (mathematics, science only).
Qualified interpreter for ELL student	An interpreter may translate directions or clarify instructions for the assessments. They may translate, not define specific words or test questions on the mathematics and science tests. On the reading test interpreters may only translate directions and may not translate or define words in the passage or test questions.
Test directions read aloud, signed, or recorded	Directions for all PSSA tests may be read aloud, signed or presented by audio recording.
Test items/questions read aloud or signed	Students unable to decode text visually may have items / questions read aloud for mathematics and science only; however, words may not be defined.
Test prompts recorded	Writing essay prompts may be presented by audio recording.

Type of Testing Accommodation	Explanation
Amplification device	In addition to hearing aids, students may require an amplification device to enhance clarity.
Other (PDE approval required)	Other presentation accommodations indicated in the *Accommodation Guidelines* may be provided; however, PDE approval is required prior to the test window.
Spanish version for mathematics and science	Students whose first language is Spanish and who have been enrolled in U.S. schools for fewer than 3 years, may take this version.
Student used the following Response Accommodations	
Brailler / Note taker (per *Accommodations Guidelines*)	Students using this device as part of their regular instructional program may use it on the PSSA; however, without thesaurus, spell- or grammar checker, etc.
Test administrator scribed open-ended responses at student's direction	A test administrator may record word-for-word exactly what a student dictated directly into the PSSA test booklet. This includes MC and OE responses for reading, mathematics and science. For writing, this includes MC items only.
Test administrator marked multiple-choice responses at student's direction	A test administrator may mark an answer booklet at the direction of a student. (e.g., a student may point to a multiple-choice answer with the test administrator marking the response in the answer booklet).
Test administrator transcribed (copied) student responses. (per *Accommodations Guidelines*)	For writing prompts the test administrator may transcribe handwriting that is extremely difficult to read. On reading, mathematics, or science illegible handwriting may be transcribed for open-ended items only.
Qualified Interpreter for ELL student (translated, transcribed, and/or scribed student responses)	A qualified interpreter may interpret a student's non-English oral responses into written English for Mathematics and science assessments. Interpreters are not permitted to make corrections or change the meaning of the response.
Augmentative communication device	Students with severe communication difficulties may use a special device to convey responses, which must be transcribed into the test booklet by the test administrator.
Typewriter, word processor or computer (per *Accommodations Guidelines*)	An allowable accommodation as a typing function only for students with identified need. Supports such as dictionaries, thesauri, spell checkers and grammar checkers must be turned off. Answers must then be transcribed into the answer booklet without alteration.

Type of Testing Accommodation	Explanation
Audio recording of student responses (per *Accommodations Guidelines*)	An electronic recording device may be used to record responses, which must be transcribed into the test booklet by the test administrator. (Students who are unable to use a pencil or have illegible handwriting may answer reading, mathematics, and writing multiple-choice questions orally. Answers must be recorded in the answer booklet without alteration during the testing period.)
Manipulative (Cranmer Abacus, number line)	An adaptive calculator or a Cranmer Abacus may be used for the calculator portion of the test only. Eligible students are only those with blindness, low vision, or partial sight.
Translation dictionary for ELL student	A word-to-word dictionary that translates native language to English (or vice versa) without word definitions or pictures is allowed on any portion of the mathematics test and open-ended section of the reading test (but not for the reading passage or multiple-choice items). Cannot be used on any section of the writing test.
Electronic screen reader (PDE approval required)	Students with blindness or extremely low vision may use computer software that converts text to synthesized speech or Braille.
Other (per *Accommodations Guidelines* or PDE approval)	Other accommodations may be appropriate and available if they do not compromise the integrity of the assessment. Documentation must be provided to PDE.
Student used the following Setting Accommodations	
Hospital/home testing	A student who is confined to a hospital or to home during the testing window may be tested in that environment.
Tested in a separate setting	A separate room may be used to reduce distraction.
Small group testing	Some students may require a test setting with fewer students or a setting apart from all other students.
Other (per *Accommodations Guidelines* or PDE approval)	Other accommodations may be appropriate and available if they do not compromise the integrity of the assessment. Documentation must be provided to PDE.

Type of Testing Accommodation	Explanation
Student used the following Timing Accommodations	
Scheduled extended time	Extended time may be allotted for each section of the test as a planned accommodation to enable students to finish.
Student-requested extended time	A student may request extended time if working productively.
Multiple test sessions	Multiple test sessions (breaks within a test section) may be scheduled for the completion of each test section; however, a test section must be completed within one school day.
Changed test schedule	Students whose disabilities prevent them from following a regular, planned test schedule may follow an individual schedule, enabling test completion.

Chapter Eleven: Classical Item Statistics

This chapter provides an overview of the two most familiar item-level statistics obtained from any classical (traditional) item analysis: item difficulty and item discrimination. The following results pertain only to operational PSSA-M items (i.e., those that contributed to a student's total test score). Related information is discussed elsewhere in this document. Specifically, Rasch item statistics are discussed in Chapter Twelve and test-level statistics in Chapter Seventeen. An analysis of item omit rates is also provided.

ITEM-LEVEL STATISTICS

Appendix I provides classical item statistics for all PSSA-M items. Results are organized by subject and grade. These statistics represent the item characteristics most often used to determine if an item functioned properly and/or how a group of students performed on a particular item. The item statistics in the appendices include: *p*-values for MC items and item means for OE items (indicators of item difficulty); point-biserial correlations for MC items and item-test correlations for OE items (indicators of item discrimination); and the proportion selecting each MC item option or earning each OE item score point.

ITEM DIFFICULTY

Item difficulty is an important consideration for the PSSA-M tests because of the ranging achievement levels of students in Pennsylvania (Below Basic-M, Basic-M, Proficient-M, and Advanced-M). At the most general level, an item's difficulty is indicated by its mean score in some specified group (e.g., grade level).

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (n). For multiple-choice items, student scores are represented by 0's and 1's (0 = wrong, 1 = right). With 0–1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. So, this is also the proportion correct for the item, or as it is better known, the *p*-value. In theory, *p*-values can range from 0.0 to 1.0 on the proportion-correct scale. For example, if an item has a *p*-value of 0.89, it means 89 percent of the students answered the item correctly. Additionally, this value might also suggest that: 1) the item was relatively easy, and/or 2) the students who attempted the item were relatively high achievers. In other words, item difficulty and student ability are somewhat confounded.

For OE items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score (four points in the case of mathematics). Sometimes a pseudo *p*-value is provided for an OE item. This is done by dividing the mean item score by the maximum possible item score.

The minimum and maximum extremes of the difficulty scale are never seen in applied practice. However, understanding what those values are helps illustrate that relatively lower values correspond to more difficult items and that relatively higher values correspond to easier items. (Because of this, some assert that this index would be better referred to as the item's easiness.)

ITEM DISCRIMINATION

Discrimination is an important consideration for the PSSA-M because the use of more discriminating items on a test is associated with more reliable test scores. This means that score estimates will be more precise (i.e., there will be smaller confidence intervals around the scores) and that more accurate performance level placements will be made. The issues of reliability, confidence intervals, and performance level classifications are further discussed in Chapter Eighteen.

At the most general level, item discrimination indicates an item's ability to differentiate between high and low achievers. It is expected that students with high ability (i.e., those who perform well on the PSSA-M overall) would be more likely to answer any given PSSA-M item correctly, while students with low ability (i.e., those who perform poorly on the PSSA-M overall) would be more likely to answer the same item incorrectly. For the PSSA-M tests, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate discrimination. (As commonly practiced, DRC removes the item score from the total score so that the resulting correlations will not be spuriously high.) The correlation coefficient can range from -1.0 to +1.0. If the aforementioned expectation is met (i.e., high-scoring students tend to get the item right while low-scoring students do not) the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero) meaning the item is a good discriminator between high and low ability students. This should be the case for all PSSA-M operational test items.

In summary, the correlation will be positive in value when the mean test score of the students answering the item correctly is higher than the mean test score of the students answering the item incorrectly[4]. In other words, this indicates that students who did well on the total test tended to do well on the item as well. However, an interaction can exist between item discrimination and item difficulty. Items answered correctly (or incorrectly) by a large proportion of examinees (i.e., they have extreme *p*-values) can have reduced power to discriminate, and thus, can have lower correlations.

Finally, discrimination for dichotomous MC items is typically referred to as the point-biserial correlation coefficient. For OE items, the term item-test correlation is sometimes used.

DISCRIMINATION ON DIFFICULTY SCATTERPLOTS

Figure 11–1 contains a series of scatterplots showing item discrimination values (Y-axis) and item difficulty (X-axis) for each grade. Note that pseudo *p*-values (described above) are used to measure the difficulty for the OE items. These plots provide maximum information about item discrimination and difficulty in a single visual image for each PSSA-M test. This is because the X- and Y-axes visually represent many important univariate distributional indices:

- The minimum and maximum values are listed.
- Mean scores are indicated by the dot.
- P_{25}, P_{50}, and P_{75} are indicated by the raised/indented portions of the axes.
- Marginal "rugs" indicate the density of the individual data points.

[4] It is legitimate to view the point-biserial correlation as a standardized mean difference. A positive value indicates students who chose that response had a higher mean score than the average student; a negative value indicates students who chose that response had a lower than average mean score.

The bivariate relationship between item discrimination (item-test correlations) and difficulty (item mean scores) is reflected by the scatterplots in these figures. However, it is often the case that items with extreme difficulties can have lower discrimination values, as can be revealed in those scatterplots.

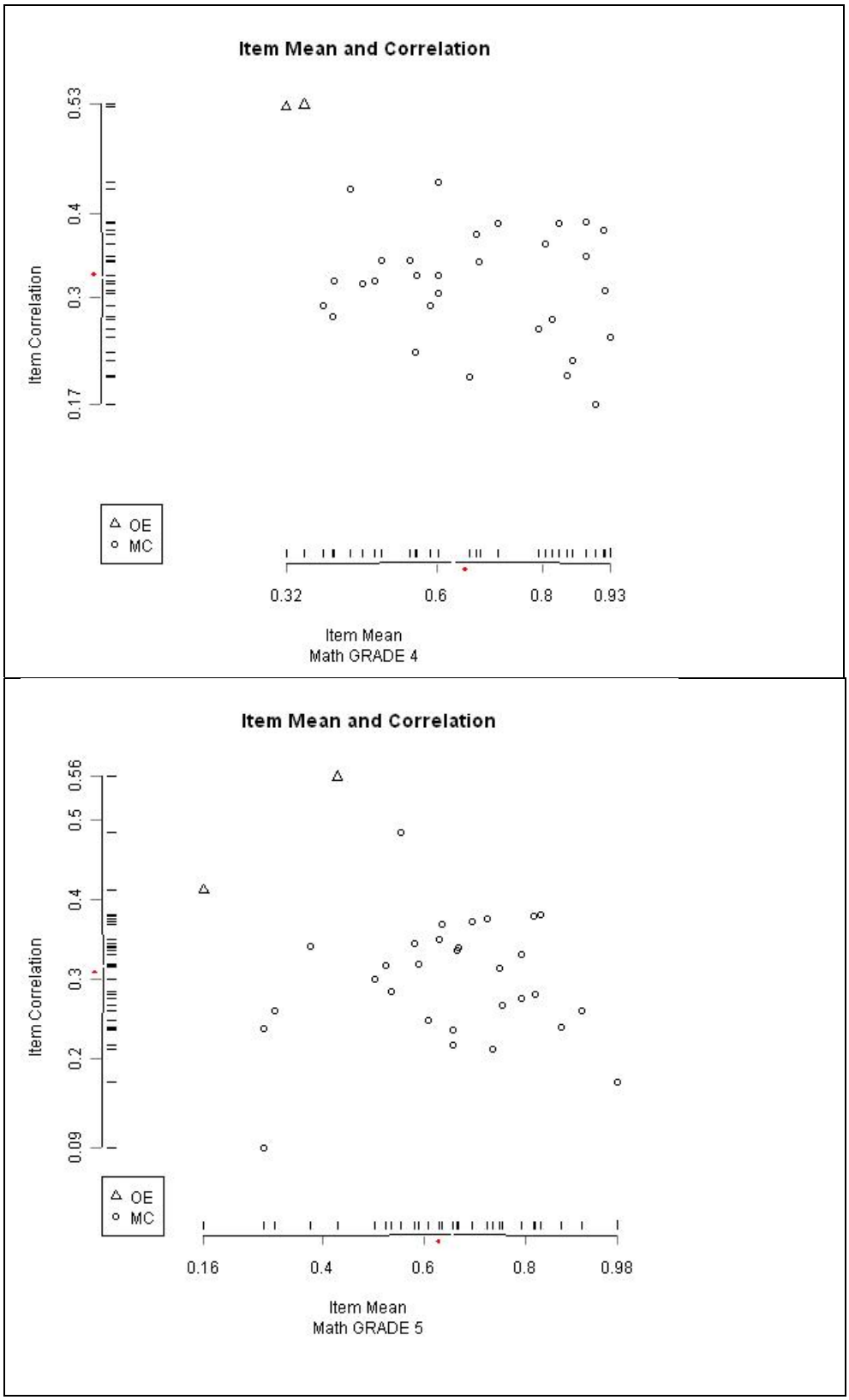
OBSERVATIONS AND INTERPRETATIONS

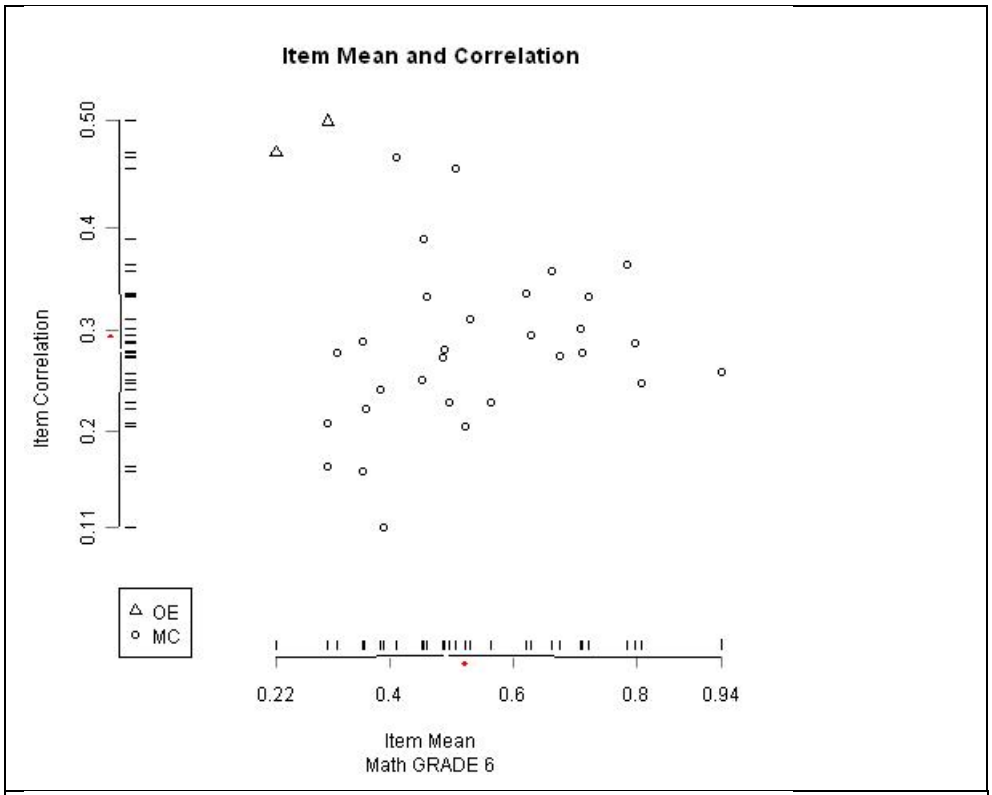
From the difficulty distributions illustrated in the scatterplots, a wide range of item difficulties appeared on each PSSA-M test, which was a desired goal. To support the visuals, Table 11–1 provides break-out results for the MC and OE items. Additional summary statistics (for the MC items only) are provided in Table 11–2. The mean *p*-values for the MC items ranged from about 0.52–0.68, while the mean proportion-correct values for the OE items ranged from about 0.18–0.33. These means were generally lower than 0.65 (a typical mean *p*-value on the general PSSA tests). Relatively speaking, this suggests that the PSSA-M items were somewhat challenging for most students taking the PSSA-M, particularly at the higher grade levels. As noted earlier, lower *p*-values can reflect that the items are more difficult or that the achievement level of the students is lower (or both).

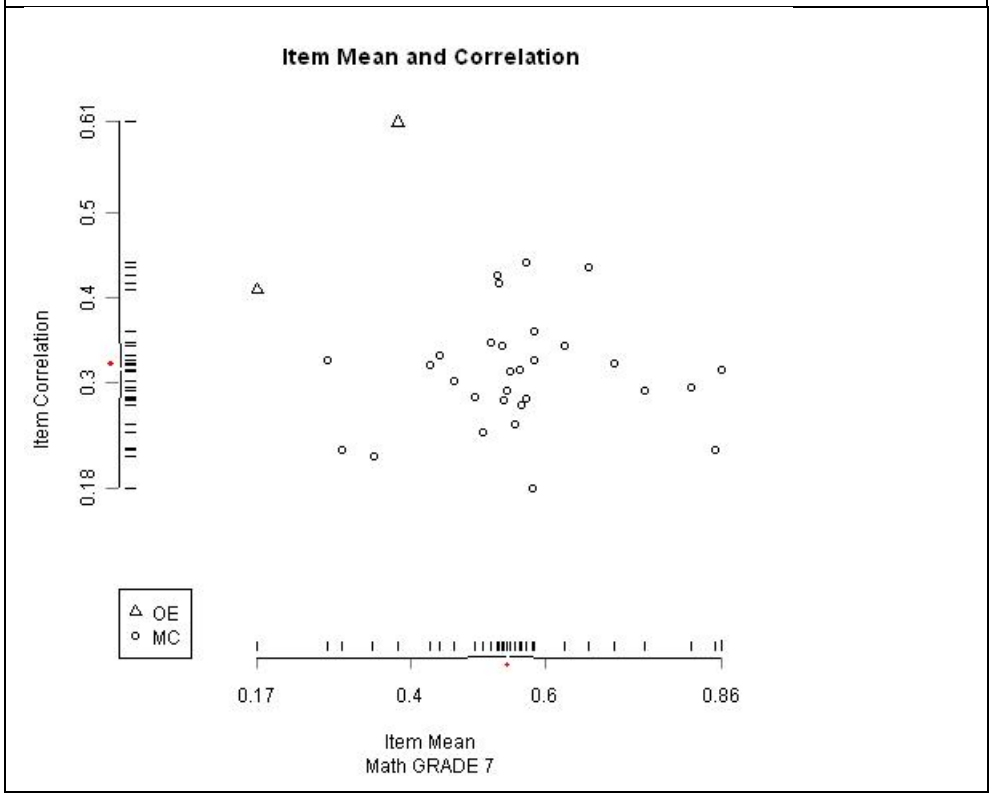
A small number of items had lower item discriminations (e.g., below 0.20). Some of these were observed on items that were very easy or hard. The mean point-biserial correlations ranged from 0.28–0.32 and 0.48–0.53 for the MC and OE items, respectively. While these values are somewhat lower than those observed on the general PSSA tests (which is not surprising given the PSSA is a longer, more reliable test), most would probably consider these values acceptable. The OE correlations tended to be higher than the MC correlations, which again is not surprising because the OE items include more score points.

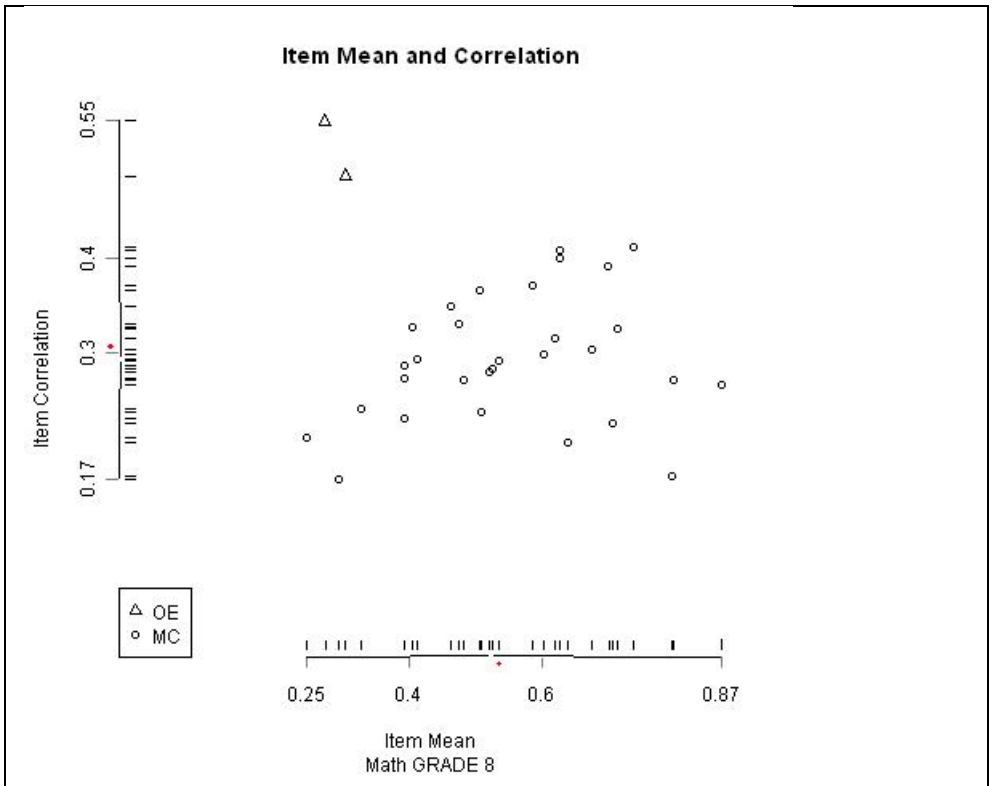
It is difficult to make global conclusions about overall test quality from the item statistics alone. With that caveat in mind, the results presented in this chapter suggest overall adequacy with respect to the PSSA-M items' difficulty and discrimination. This in turn implies that the items generally functioned as expected for the population of students who took the PSSA-M.

Figure 11–1. Discrimination on Difficulty Scatterplots




Item Mean and Correlation
Item Correlation
0.11
0.2
0.3
0.4
0.50
OE
MC
0.22
0.4
0.6
0.8
0.94
Item Mean
Math GRADE 6


Item Mean and Correlation
Item Correlation
0.18
0.3
0.4
0.5
0.61
OE
MC
0.17
0.4
0.6
0.86
Item Mean
Math GRADE 7


Item Mean and Correlation
Item Correlation
0.17
0.3
0.4
0.55
OE
MC
0.25
0.4
0.6
0.87
Item Mean
Math GRADE 8

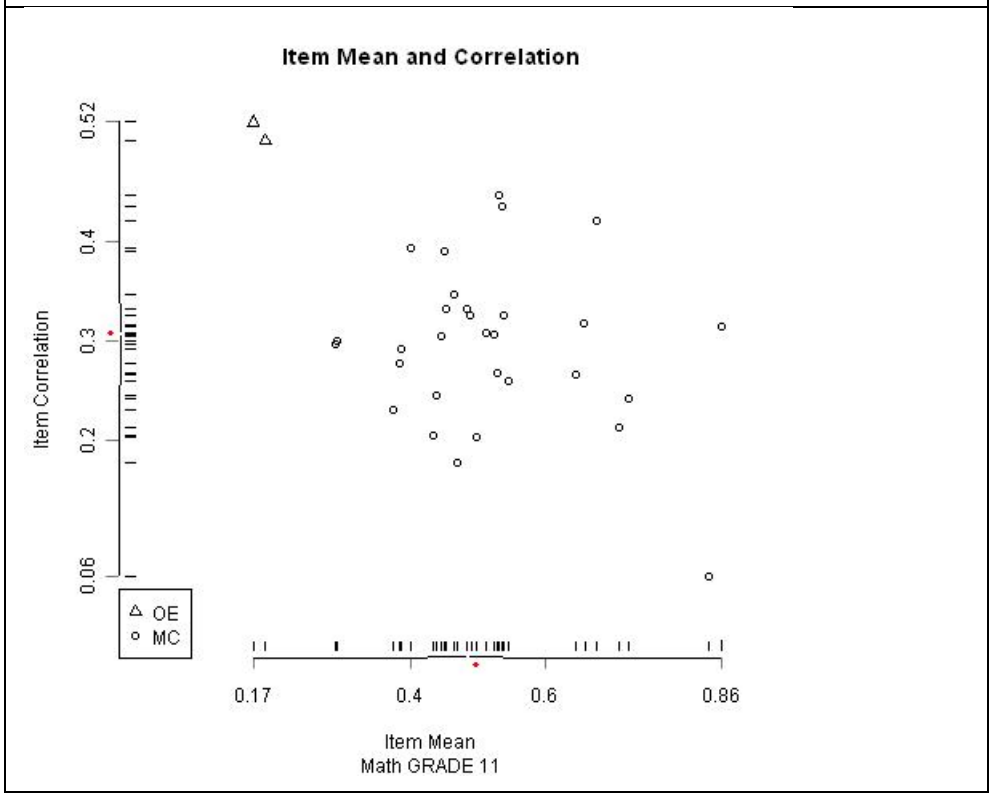

Item Mean and Correlation
Item Correlation
0.06
0.2
0.3
0.4
0.52
OE
MC
0.17
0.4
0.6
0.86
Item Mean
Math GRADE 11

Table 11–1. Sum and Mean Statistics for MC and OE Items

Subject	Grade	Multiple-Choice Items				Open-Ended Items			
		Points	Sum	Mean (%/100)	Mean I-T Corr.[5]	Points	Sum	Mean (%/100)	Mean I-T Corr.
Mathematics	4	30	20.236	0.675	0.315	8	2.654	0.332	0.530
	5	30	19.521	0.651	0.298	8	2.367	0.296	0.484
	6	30	16.199	0.540	0.281	8	2.073	0.259	0.488
	7	30	16.840	0.561	0.310	8	2.221	0.278	0.508
	8	30	16.594	0.553	0.293	8	2.325	0.291	0.516
	11	30	15.564	0.519	0.294	8	1.398	0.175	0.512

Table 11–2. Additional Summary Statistics for MC Items Only

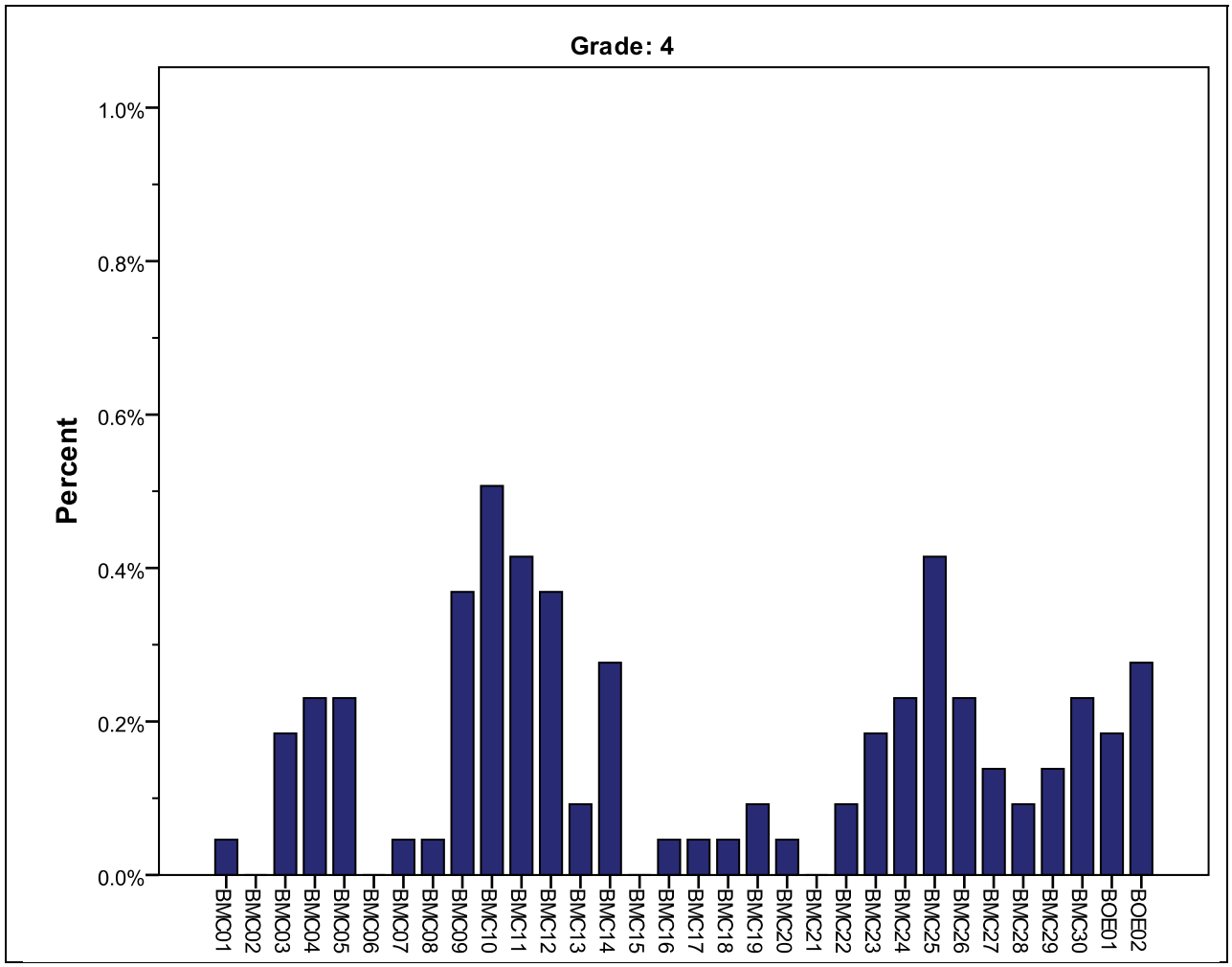
		P-Value				Point Biserial			
		Min	Max	Mean	Med	Min	Max	Mean	Med
Mathematics	4	0.39	0.93	0.68	0.67	0.17	0.44	0.32	0.32
	5	0.28	0.98	0.65	0.66	0.09	0.49	0.30	0.30
	6	0.30	0.94	0.54	0.50	0.11	0.47	0.28	0.28
	7	0.28	0.86	0.56	0.55	0.18	0.44	0.31	0.31
	8	0.25	0.87	0.55	0.53	0.17	0.41	0.29	0.29
	11	0.29	0.86	0.52	0.49	0.06	0.45	0.29	0.30

[5] The means for the I-T correlations were not computed using Fisher's Z transformation (which, strictly speaking, would have been more appropriate). However, this is not expected to affect the conclusions based on this data.

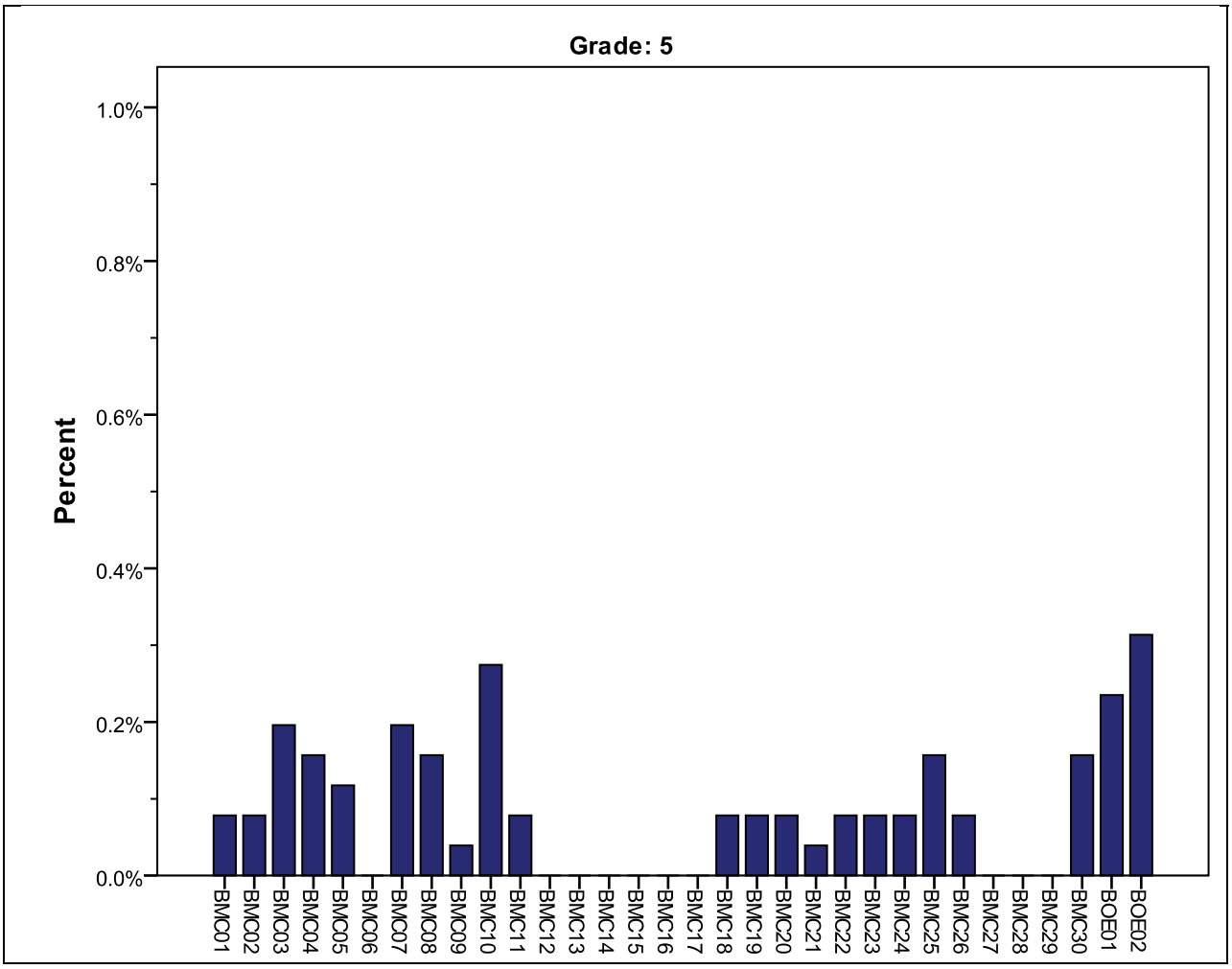
ITEM OMIT RATES

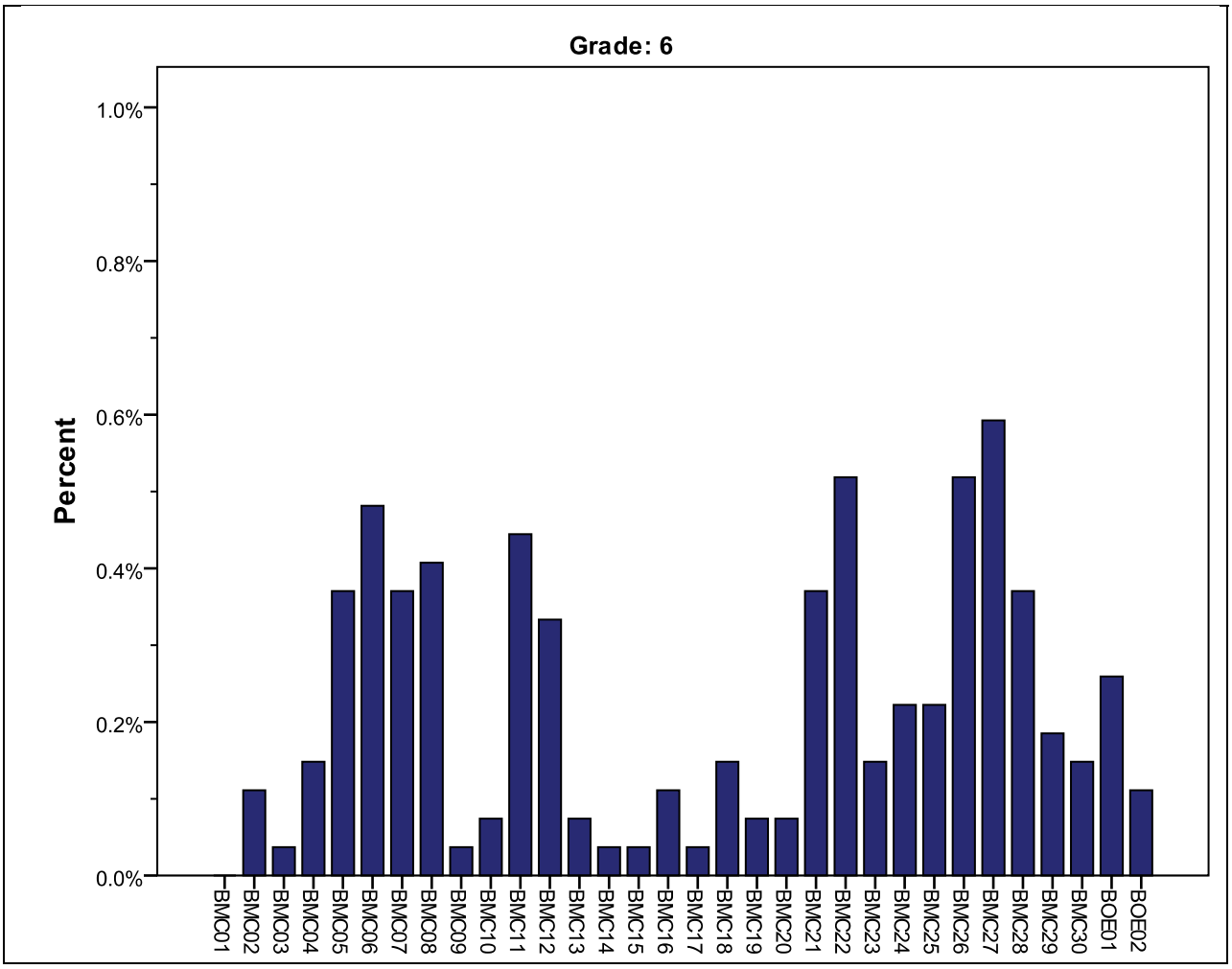
Omit rates are analyzed at the item level (Figure 11–2) and test level (Figure 11–3). High omit rates might be observed for a number of reasons (frustration on very hard items, evidence of fatigue, or speededness). Only students who meet the PSSA-M attemptedness criteria are included in these analyses (answering four or more items in both sections of the PSSA-M Mathematics test). Items are presented in their sequential administration order along the x-axis. With the exception of the two OE items at Grade 11, (OE items are located on the far right of the x-axis in all Figure 11–2 bar graphs), all item omit rates were less than 1.0%. At higher grades OE items were omitted more frequently than at lower grades. Not surprisingly, the majority of test takers at all grades answered all test items (i.e., had zero item omits in Figure 11–3). Grade 11 had the lowest percent of zero item omits at just under 90%. The PSSA-M would not be considered speeded based on the Swineford (1956) criteria.[6]

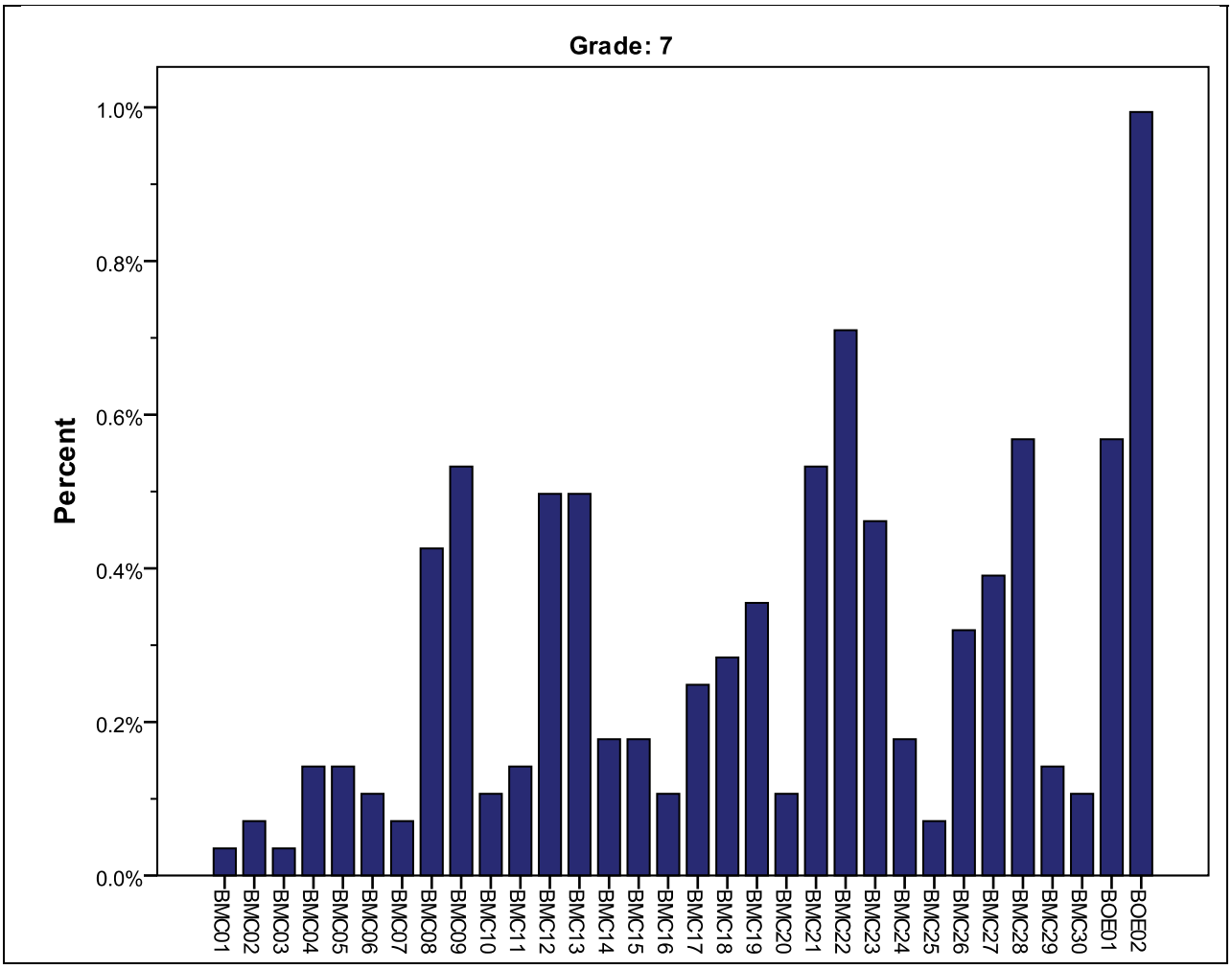
Figure 11–2. Omit Rates for Individual Test Items

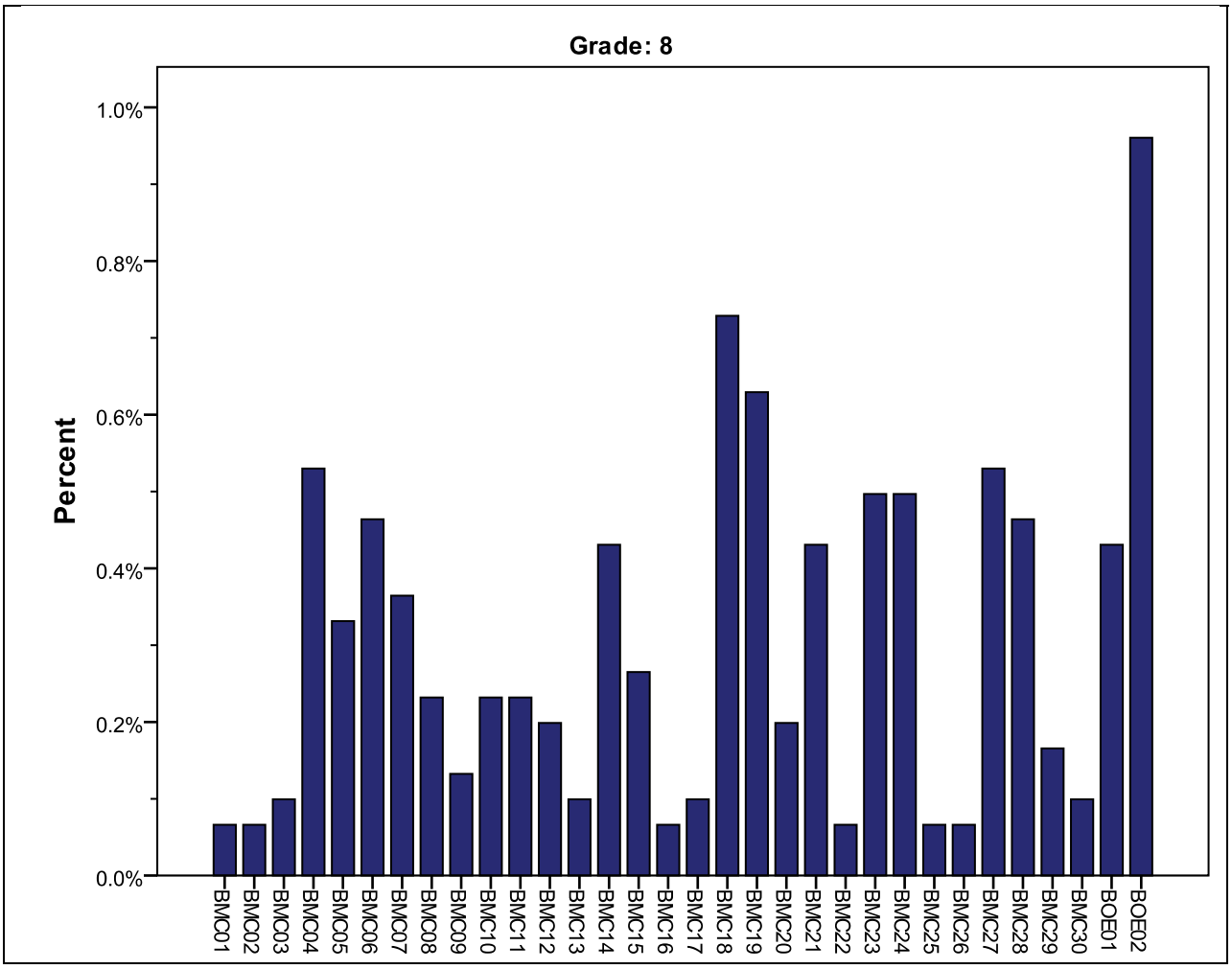


[6] If 99% of the examinees attempt 75% of the items, and if all items are attempted by 80% of examinees.


Grade: 5
Percent
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
BMC01
BMC02
BMC03
BMC04
BMC05
BMC06
BMC07
BMC08
BMC09
BMC10
BMC11
BMC12
BMC13
BMC14
BMC15
BMC16
BMC17
BMC18
BMC19
BMC20
BMC21
BMC22
BMC23
BMC24
BMC25
BMC26
BMC27
BMC28
BMC29
BMC30
BOE01
BOE02


Grade: 6
Percent
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
BMC01
BMC02
BMC03
BMC04
BMC05
BMC06
BMC07
BMC08
BMC09
BMC10
BMC11
BMC12
BMC13
BMC14
BMC15
BMC16
BMC17
BMC18
BMC19
BMC20
BMC21
BMC22
BMC23
BMC24
BMC25
BMC26
BMC27
BMC28
BMC29
BMC30
BOE01
BOE02


Grade: 7
Percent
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
BMC01
BMC02
BMC03
BMC04
BMC05
BMC06
BMC07
BMC08
BMC09
BMC10
BMC11
BMC12
BMC13
BMC14
BMC15
BMC16
BMC17
BMC18
BMC19
BMC20
BMC21
BMC22
BMC23
BMC24
BMC25
BMC26
BMC27
BMC28
BMC29
BMC30
BOE01
BOE02


Grade: 8
Percent
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
BMC01
BMC02
BMC03
BMC04
BMC05
BMC06
BMC07
BMC08
BMC09
BMC10
BMC11
BMC12
BMC13
BMC14
BMC15
BMC16
BMC17
BMC18
BMC19
BMC20
BMC21
BMC22
BMC23
BMC24
BMC25
BMC26
BMC27
BMC28
BMC29
BMC30
BOE01
BOE02

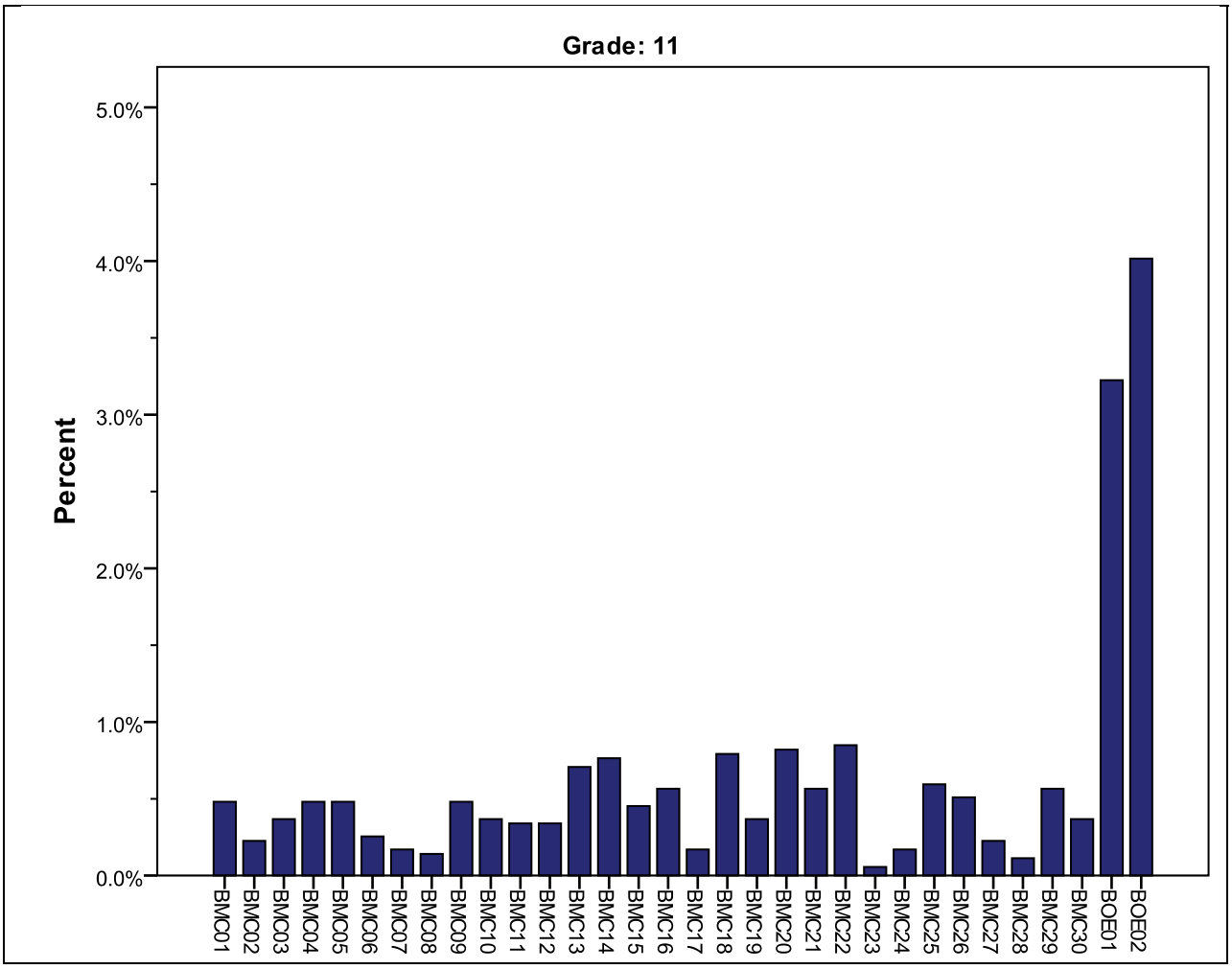
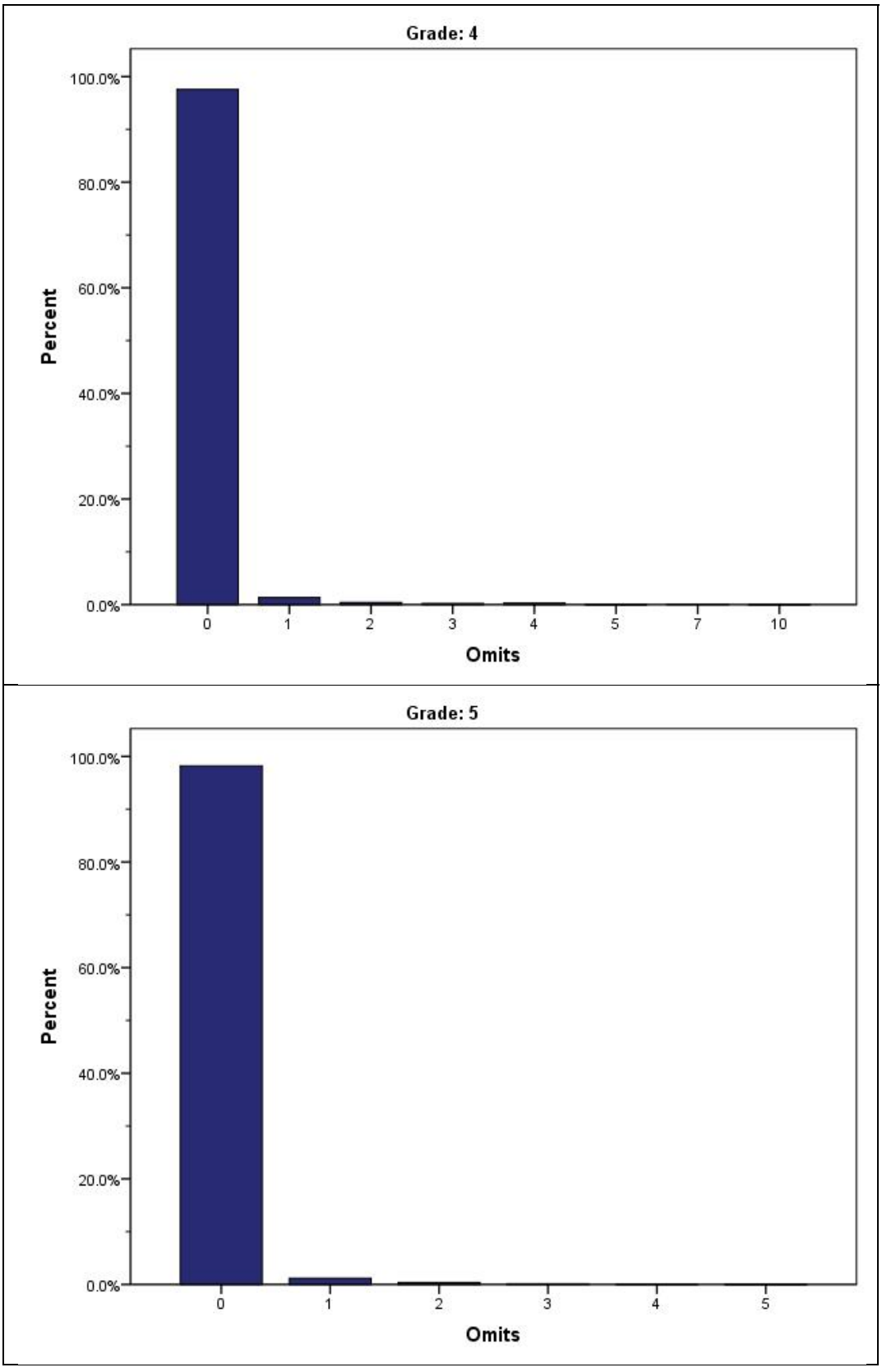
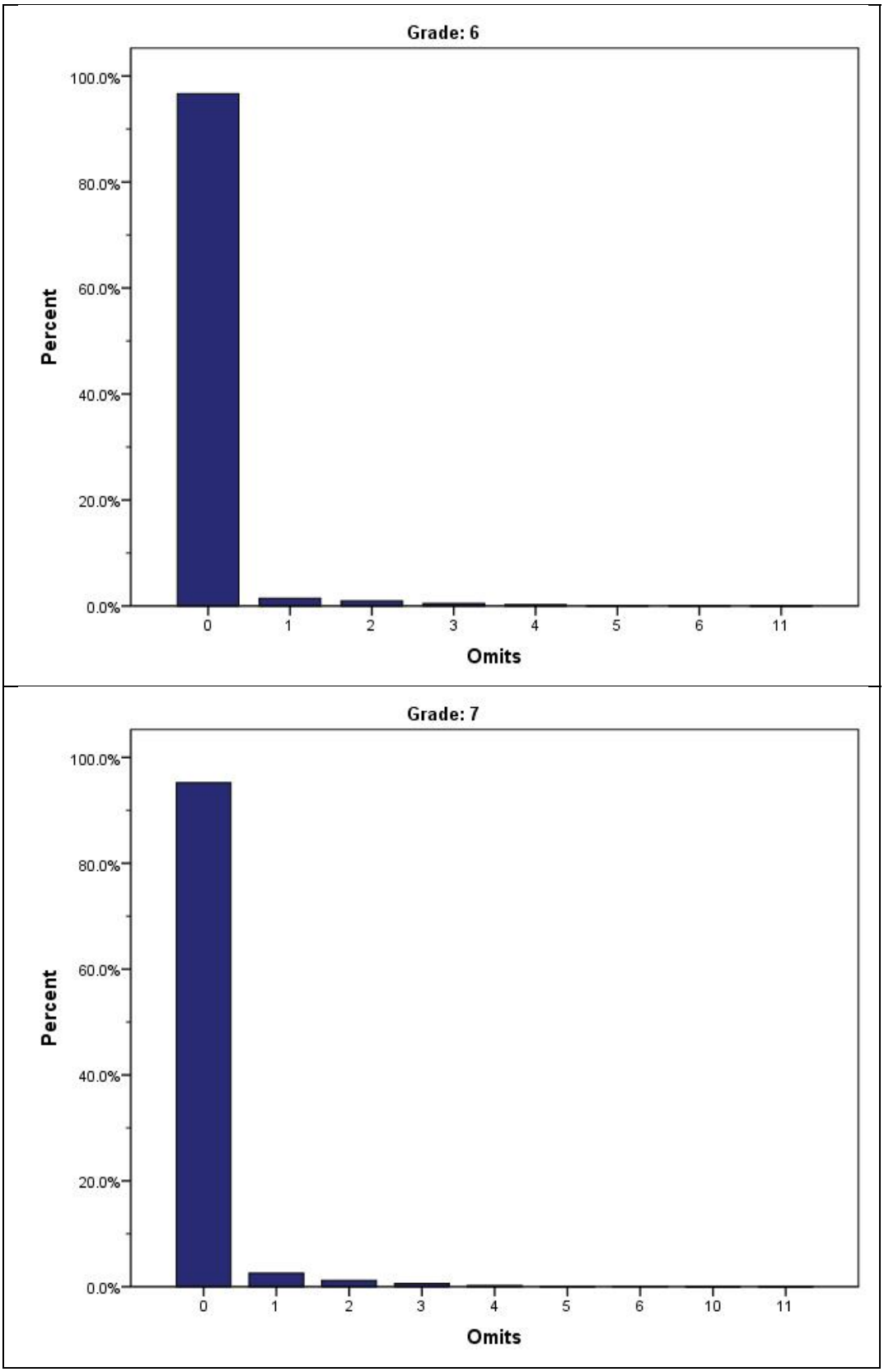
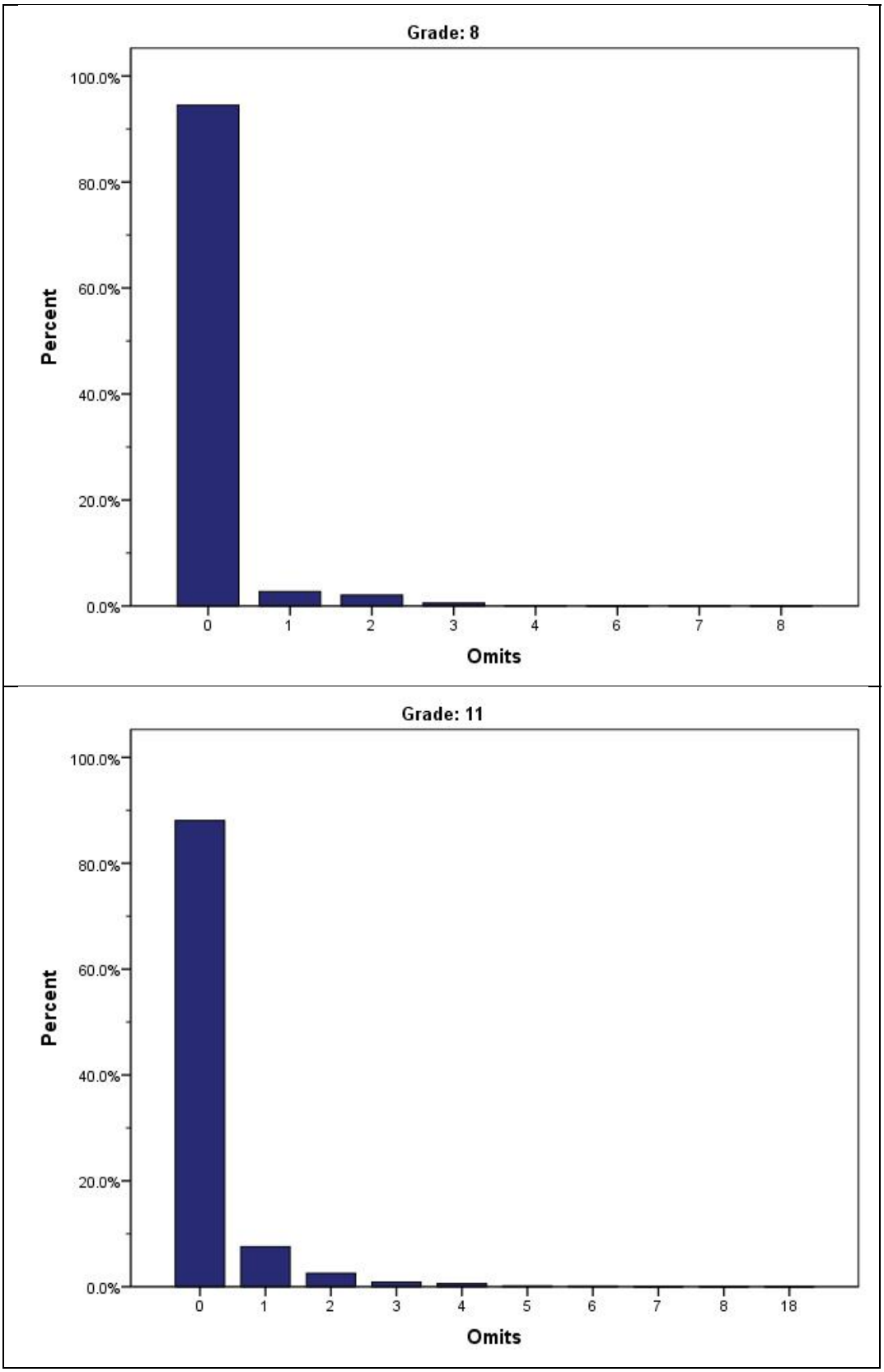

Grade: 11
Percent
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
BMC01
BMC02
BMC03
BMC04
BMC05
BMC06
BMC07
BMC08
BMC09
BMC10
BMC11
BMC12
BMC13
BMC14
BMC15
BMC16
BMC17
BMC18
BMC19
BMC20
BMC21
BMC22
BMC23
BMC24
BMC25
BMC26
BMC27
BMC28
BMC29
BMC30
BOE01
BOE02

Figure 11–3. Total Number of Items Omitted on Test




Grade: 6
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
Percent
0
1
2
3
4
5
6
11
Omits
Grade: 7
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
Percent
0
1
2
3
4
5
6
10
11
Omits


Grade: 8
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
Percent
0
1
2
3
4
6
7
8
Omits
Grade: 11
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
Percent
0
1
2
3
4
5
6
7
8
18
Omits

Chapter Twelve: Rasch Item Calibration

The particular Item Response Theory (IRT) model used for the PSSA-M is based on the work of Georg Rasch. Rasch models have had a long-standing presence in applied testing programs and it has been the methodology continually used to calibrate PSSA-M items in recent history. IRT has several advantages over classical test theory, so it has become the standard procedure for analyzing item response data in large-scale assessments. However, IRT models make a number of strong assumptions related to dimensionality, local independence, and model-data fit. Resulting inferences derived from any application of IRT rests strongly on the degree to which the underlying assumptions are met.

This chapter outlines the procedures used for calibrating the operational PSSA-M items. Generally, item calibration is the process of assigning a difficulty-parameter estimate to each item on an assessment so that they are placed onto a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics for the PSSA-M mathematics tests. Additional Rasch procedures are discussed with respect to scale linking in Chapter Fifteen.

DESCRIPTION OF THE RASCH MODEL

The Rasch partial credit model (RPCM; Wright and Masters, 1982) was used to calibrate PSSA-M items because both multiple-choice (MC) and open-ended (OE) items were part of the assessment. The RPCM extends the Rasch model (Rasch, 1960) for dichotomous (0, 1) items so that it accommodates the polytomous OE item data. Under the RPCM, for a given item i with m_i score categories, the probability of person n scoring x ($x = 0, 1, 2, \ldots m_i$) is given by:

$$P_{ni}(X = x) = \frac{\exp\sum_{j=0}^{x}(\theta_n - D_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(\theta_n - D_{ij})},$$

where θ_n represents a student's proficiency (ability) level, and D_{ij} is the step difficulty of the j^{th} step on item i. For dichotomous MC items, the RPCM reduces to the standard Rasch model and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person n getting item i correct as follows:

$$P_{ni}(X = 1) = \frac{\exp(\theta_n - D_{ij})}{1 + \exp(\theta_n - D_{ij})}.$$

The Rasch model places both student ability and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, it also provides person ability estimates that are independent of the items employed in the assessment, and conversely, estimates item difficulty independently of the sample of examinees. (As noted in Chapter Eleven, interpretation of item p-values confounds item difficulty and student ability.)

Software and Estimation Algorithm

Item calibration was implemented via WINSTEPS 3.54 computer program (Wright and Linacre, 2003), which employs unconditional (UCON), joint-maximum-likelihood estimation (JMLE).

Sample Characteristics

The characteristics of calibration samples are reported in Chapter Nine. These samples only include the students who attempted the tests. All omits (no response) and multiple responses (more than one response selected) were scored as incorrect answers (coded as 0s) for calibration.

CHECKING RASCH ASSUMPTIONS

Because the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the PSSA-M, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. Though a variety of methods are available for assessing these issues, the Rasch analyses and criteria available from WINSTEPS were used here. It should be noted that only operational items were analyzed since they are the basis of student scores.

Unidimensionality

Rasch models assume that one dominant dimension determines the difference among students' performance. WINSTEPS provides results from a Principal Components Analysis (PCA) that can be used to assess the unidimensionality assumption. Different from standard applications of PCA, WINSTEPS conducts its PCA on the response residuals, not the original observations. That is, the primary dimension from the Rasch model is removed first and then the residual variance is analyzed. The purpose of the analysis is to verify whether any other dominant component(s) exist among the residuals (i.e., whether they account for a practically significant amount of residual variance). If any other dimensions are found, the unidimensionality assumption would be violated.

WINSTEPS provides three PCA residuals: raw, standardized, and logit. All three should yield similar results. The mixed residual setting was used for the PCA because: 1) previous research has demonstrated that raw residuals (PRCOMP=R) give a more realistic estimate of explained variance than standardized residuals (PRCOMP=S), and 2) standardized residuals are better for decomposing the unexplained variance into contrasts (Linacre, 2009).

Table 12–1 presents the PCA results for the mathematics tests. The results include the total raw variance, raw variance explained by the model, unexplained total variance, and unexplained variance in the first component (both eigenvalue units and percent values are tabled). In addition, the modeled column provides variance components that would be explained if the data complied with the Rasch definition of unidimensionality.

As can been seen from Table 12–1, for PSSA-M mathematics the primary dimension in the Rasch model explained about 36–50 percent of the total variance across Grades 4 to 8 and 11. If the data fit the model in such a way that only random noise was present, about 35–49 percent of the variance would be explained. The empirical and model-based percentages were quite close, suggesting that the estimation of a primary Rasch dimension was successful. The unexplained variance ranged from approximately 50–64 percent for all the mathematics tests. This included the Rasch-predicted randomness and any departures in the data from the Rasch model (e.g., departure from unidimensionality).

The most important variance for evaluating dimensionality is in the row named "unexplained variance in 1st contrast." The eigenvalues of unexplained total variance were 32 for all the mathematics tests, which equals the total number of the operational items on each test. The eigenvalues of the first contrast (again, this is the second dimension beyond the first Rasch model dimension in WINSTEPS PCA) ranged from 1.5 to 1.8. This indicates that the second dimension accounted for only 1.5 to 1.8 units out of 32 units of item residual variance. Regarding the percentage, we can see that the second dimension represented less than 5.6 percent of the residual variance for each mathematics test (these percentages are shown in Column 4, which is unnamed in the column headings). Overall, WINSTEPS PCA suggests that there is one clearly dominant dimension for all mathematics tests.

Table 12–1. Results from PCA of Residuals in WINSTEPS - Mathematics

Mathematics Grade 4				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	61.0	100.0%		100.0%
Raw variance explained by measures	29.0	47.5%		46.9%
Raw unexplained variance (total)	32.0	52.5%	100.0%	53.1%
Unexplained variance in 1st contrast	1.6	2.7%	5.1%	
Mathematics Grade 5				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	63.5	100.0%		100.0%
Raw variance explained by measures	31.5	49.6%		48.9%
Raw unexplained variance (total)	32.0	50.4%	100.0%	51.1%
Unexplained variance in 1st contrast	1.5	2.4%	4.8%	
Mathematics Grade 6				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	51.2	100.0%		100.0%
Raw variance explained by measures	19.2	37.4%		36.3%
Raw unexplained variance (total)	32.0	62.6%	100.0%	63.7%
Unexplained variance in 1st contrast	1.5	2.9%	4.6%	
Mathematics Grade 7				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	53.3	100.0%		100.0%
Raw variance explained by measures	21.3	39.9%		40.1%
Raw unexplained variance (total)	32.0	60.1%	100.0%	59.9%
Unexplained variance in 1st contrast	1.8	3.3%	5.5%	
Mathematics Grade 8				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	49.9	100.0%		100.0%
Raw variance explained by measures	17.3	35.8%		34.8%
Raw unexplained variance (total)	32.0	64.2%	100.0%	65.2%
Unexplained variance in 1st contrast	1.8	3.6%	5.6%	
Mathematics Grade 11				
	Eigenvalue	**Empirical**		**Modeled**
Total raw variance in observations	53.8	100.0%		100.0%
Raw variance explained by measures	21.8	40.6%		39.1%
Raw unexplained variance (total)	32.0	59.4%	100.0%	60.9%
Unexplained variance in 1st contrast	1.6	3.0%	5.1%	

Local Independence

Local independence (LI) is a fundamental assumption of IRT. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. In formal statistical terms, a test X that is comprised of items X_1, X_2,…X_n is locally independent with respect the latent variable θ if, for all $x = (x_1, x_2,\ldots x_n)$ and θ,

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta).$$

This formula essentially states that the probability of any pattern of responses across all items (x), after conditioning on the abilities (θ) measured by the test, should be equal to the product of the conditional probabilities across each item (cf. the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of local independence. A weak form of local independence (WLI) was proposed by McDonald (1979). The distinction is important as many indicators of local dependency are actually framed by WLI. The requirement here would be for the conditional covariances of all pairs of item responses, conditioned on the abilities, to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on abilities, is the product of the probabilities of responses to these two items, as show below. (This is a weaker form because higher-order dependencies among items are allowed.) Based on the WLI, the following expression can be derived:

$$P(X_i = x_i, X_j = x_j \mid \theta) = P(X_i = x_i \mid \theta)P(X_j = x_j \mid \theta).$$

Marais and Andrich (2008) pointed out that local item dependence in the Rasch model can occur in two ways that some may not distinguish. The first way occurs when the assumption of unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension also determine students' performance (this can be called trait dependence). The second violation occurs when responses to an item depend on responses to another. This is a violation of statistical independence and can be called response dependence. Many people treat the assumptions of unidimensionality and local independence as one phenomenon and believe that once unidimensionality holds, that local independence also holds. By distinguishing the two sources of local dependence, one can see that while local independence can be related to unidimensionality, the two are different assumptions, and therefore, require different tests.

Residual item correlations provided in WINSTEPS for each item pair were used to assess the local dependence among the PSSA-M items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using ability and item parameter estimates. Next, deviations (residuals) between the examinees' expected and observed performance is determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

As mentioned before, three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It should be noted that the raw score residual correlation essentially corresponds to Yen's Q_3 index, a popular LI statistic. The expected value for the Q_3 statistic is approximately -1/(k-1) when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected Q_3 values should be approximately -0.03 for the PSSA-M tests (since most of the PSSA-M tests had 32 core items). Index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default standardized residual correlation in WINSTEPS was used for these analyses. Table 12–2 shows the summary statistics—mean, SD, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90}) — for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with the residual correlations greater than 0.20 are also reported in this table. The mean residual correlations were slightly negative and the values were close to -0.03. The vast majority of the correlations were very small, suggesting local item independence generally holds for the PSSA-M mathematics tests.

Table 12–2. Summary of Item Residual Correlations for PSSA-M Mathematics

	GRADE					
Statistic	**4**	**5**	**6**	**7**	**8**	**11**
	MATHEMATICS					
N	496	496	496	496	496	496
Mean	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
SD	0.04	0.03	0.03	0.04	0.04	0.03
Minimum	-0.13	-0.14	-0.12	-0.12	-0.12	-0.11
P_{10}	-0.07	-0.07	-0.07	-0.07	-0.08	-0.06
P_{25}	-0.05	-0.05	-0.05	-0.05	-0.07	-0.05
P_{50}	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
P_{75}	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
P_{90}	0.02	0.01	0.01	0.01	0.01	0.01
Maximum	0.22	0.12	0.14	0.33	0.48	0.14
>0.20	1	0	0	2	2	0

Item Fit

WINSTEPS provides two item-fit statistics (infit and outfit) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (Zstd with mean = 0 and variance = 1). MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. Though both are informative, the Zstd values are very likely too sensitive to the large sample sizes observed on the PSSA-M. In this situation it is recommended that the Zstd values be ignored if the MnSq values are acceptable (Linacre, 2009).

Both infit and outfit MnSq are the average of standardized residual variance (the difference between the observed score and the Rasch estimated score divided by the square root of the Rasch model variance). The difference is that the outfit statistic gives all examinees equal weight in computing the fit and tends to be affected more by unexpected responses far from the person, item, or rating scale category measure (i.e., it is more sensitive to outlying, off-target, low information responses). The infit statistic is weighted by the examinee locations relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses). Some feel that extreme infit values are a greater threat to the measurement process than extreme outfit since most tests intend to measure the on-target population rather than extreme outliers.

The expected MnSq value is 1.0, and can range from 0 to infinity. Deviation in excess of the expected value can be interpreted as noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable, too much noise). Rules of thumb regarding practically significant MnSq values vary. More conservative users might prefer items with MnSq values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values from 0.5 to 1.5. In the results below, values outside of 0.7 to 1.3 are given practical importance.

Table 12–3 presents the summary statistics of infit and outfit mean square statistics for the PSSA-M mathematics tests including the mean, SD, and minimum and maximum values. The number of items within the range of (0.7, 1.3) is also reported in Table 12–3. As can been seen, the mean values for both fit statistics were close to 1.00 for all tests. All the items had infit values falling in the range of (0.7, 1.3). Though more outfit values fell outside this range than infit values, most of the extreme values were just barely above 1.3 or below 0.7. Overall, these results indicate that the Rasch model fits the PSSA-M item data well.

Table 12–3. Summary of Infit and Outfit Mean Square Statistics for PSSA-M Mathematics

	Infit Mean Square					Outfit Mean Square				
Test	**Mean**	**SD**	**Min**	**Max**	**[0.7,1.3]**	**Mean**	**SD**	**Min**	**Max**	**[0.7,1.3]**
M4	0.98	0.07	0.83	1.12	32/32	0.97	0.15	0.56	1.23	31/32
M5	0.98	0.06	0.86	1.17	32/32	0.98	0.16	0.52	1.47	30/32
M6	0.99	0.07	0.87	1.16	32/32	0.99	0.11	0.66	1.22	31/32
M7	1.00	0.07	0.88	1.12	32/32	1.01	0.12	0.81	1.15	31/32
M8	1.00	0.09	0.81	1.21	32/32	1.00	0.09	0.81	1.21	32/32
M11	0.99	0.07	0.85	1.12	32/32	1.00	0.11	0.72	1.33	31/32

RASCH ITEM STATISTICS

As noted earlier, the Rasch model expresses item difficulty (and student ability) in units referred to as *logits*, rather than on the percent-correct metric. In the simplest case, a logit is a transformed *p*-value with the average *p*-value becoming a logit of zero. In this form, logits resemble *z*-scores or standard normal deviates; a very difficult item might have a logit of +4.0 and a very easy item might have a logit of –4.0. However, they have no formal relationship to the normal distribution.

The logit metric has several mathematical advantages over *p*-values. Logits have an interval scale, meaning that two items with logits of 0.0 and +1.0 (respectively) are the same distance apart as two items with logits of +3.0 and +4.0. Logits are not dependent on the ability level of the students. For example, a test form can have a mean logit of zero, whether the average item *p*-value for the student sample is 0.8 or 0.3.

The standard Rasch calibration procedure arbitrarily fixes the mean difficulty of the items on any form at zero. Under normal circumstances where all students are administered the same set of items, any item with a *p*-value lower than the average item on the form receives a positive logit difficulty and any item with a *p*-value higher than the average receives a negative logit. Consequently, the logits for any calibration relate to an arbitrary origin defined by the center of items on that form. Logits for both item difficulties and student abilities are placed on the same scale and relate to the same mean item difficulty.

There are a number of other arbitrary choices that could be made for centering the item difficulties. Rather than using all the items, the origin could be defined by a subset. For the PSSA-M, all test forms in a particular grade and content area share the same operational item set. All items on each form can then be easily adjusted to a single (but still arbitrary) origin by defining the origin as the mean of the operational items. With this done, the origins for all the forms will be statistically equal. For example, items on any two forms that are equally difficult will now have statistically equal logit difficulties. This is partly how PSSA-M items can be placed on the same logit difficulty scale across years. Chapter Fifteen has more detailed information about the PSSA-M scale linking procedures.

Appendix I reports the item statistics including classical and Rasch logit difficulties for all the operational items. Table 12–4 summarizes the Rasch logit difficulties of the operational items on each test. Within each content area, most grades had similar mean logits. The spread of the item mean difficulties for PSSA-M mathematics tests was a little more extreme. Here, Grade 7 had the largest mean logit value (0.29), whereas Grade 8 had the lowest mean logit value (0.08). The minimum and maximum values and standard deviations suggest the PSSA-M items covered a relative wide range of difficulties.

Table 12–4. Summary of Rasch Item Difficulties for PSSA-M Mathematics

Grade	N	Minimum	Maximum	Mean*	SD
		Mathematics			
4	32	-1.86	1.79	0.10	1.11
5	32	-3.48	2.46	0.11	1.16
6	32	-2.73	1.46	0.09	0.93
7	32	-1.74	1.43	0.29	0.70
8	32	-1.84	1.55	0.08	0.82
11	32	-1.90	2.04	0.11	0.80

*The mean logit values are not necessarily 0.0 because the item have been placed on a scale that was developed in prior years.

VISUALIZING THE *P*-VALUE-LOGIT RELATIONSHIP

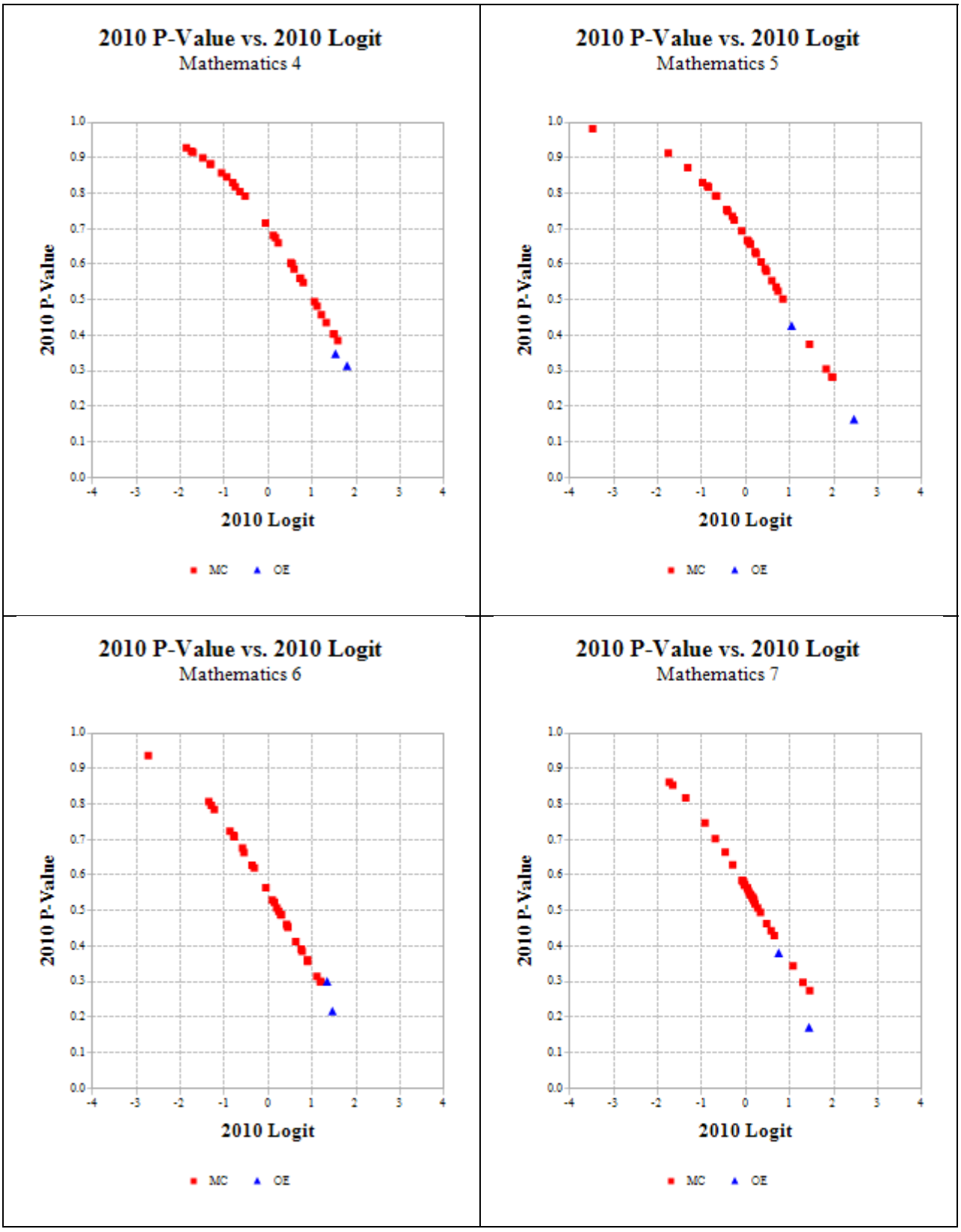
During the PSSA-M administration, test forms were spiraled within classrooms. In effect, students were administered the same set of common items but different nonoperational items (e.g., field test item sets). Cross checks can be made to ensure the calibration and linking processes are reasonable across forms. The goal of spiraling is to achieve randomly-equivalent samples of students across forms with equal standard deviations and arbitrary means. Any differences in performance observed among the groups should only be due to differences in form difficulty. After linking, the mean of the logit (Rasch student) abilities should be statistically equal for each sample of students. Because of the equivalent samples, common items should have the same *p*-values regardless of which form and sample is being considered. Also, for all items (operational and nonoperational) a plot of the relationship between the item *p*-values and item logits (Rasch item difficulty estimates) should fall along a single, curved line.

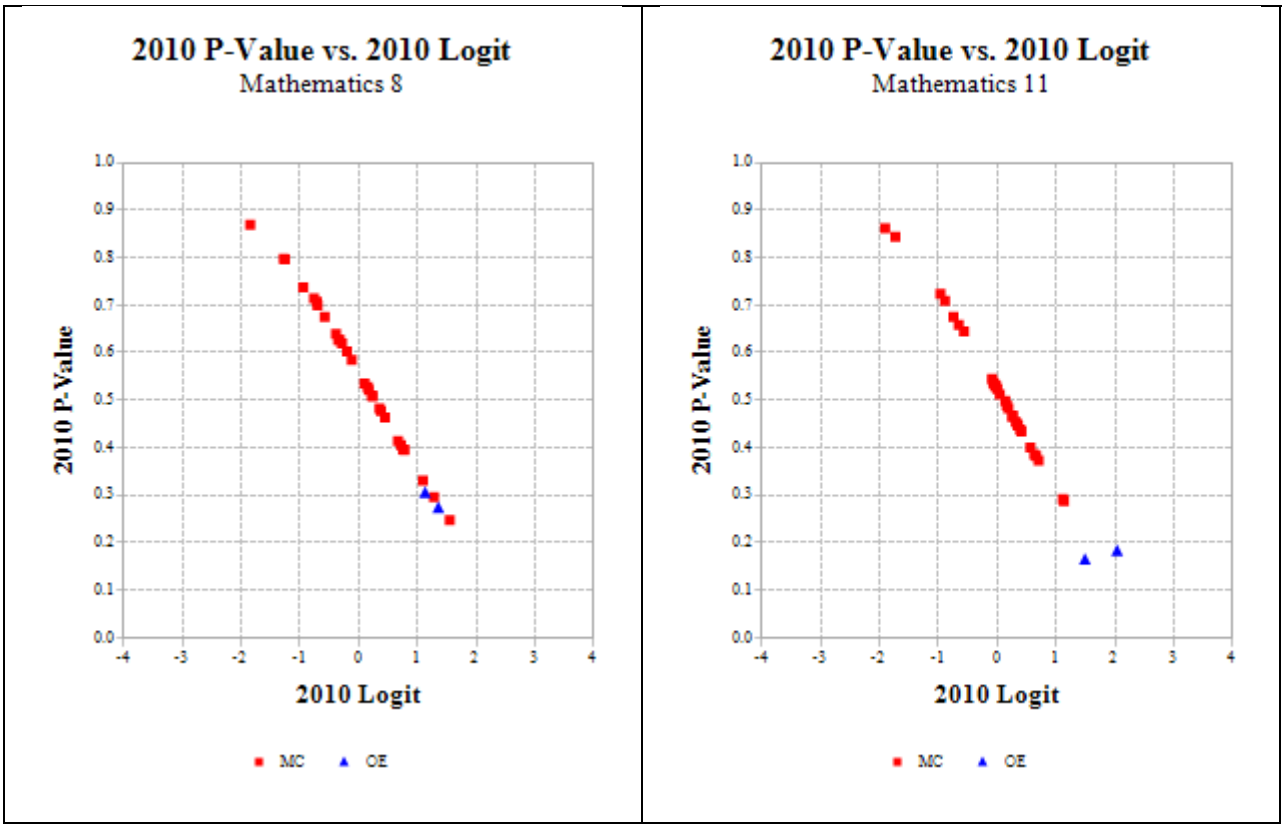
Figure 12–1 shows plots of the *p*-value-logit relationship for the operational items. The curves are nearly linear in the center, but curve towards asymptotes of one and zero, respectively, on the left and right. The graphs show that items with lower *p*-values (indicating a more difficult item that fewer students answered correctly) also had higher logit difficulties, and that items with higher *p*-values had lower logit difficulties (i.e., the *p*-value and logit scales are inversely related).

The spread of the graph points is indicative of the dispersion of item difficulties in the operational items. The dispersion and coordinates of items are roughly similar across grades for mathematics.

Common OE items are also graphed in Figure 12–1. These items appear with triangular markers. The OE items generally fall on the same curve as the MC items but subtle differences can occur. The OE items were placed on the MC item difficulty (*p*-value) scale—which ranges from 0.00 to 1.00—by dividing the mean OE item score by the maximum OE score possible. Also, the MC items were calibrated concurrently. The OE items were placed on the MC scale in a separate step (i.e., MC items were concurrently calibrated, then anchored by programmatically fixing their values when the difficulties of OE items were estimated). More information about the scale linking procedure is provided in Chapter Fifteen.

Figure 12–1. 2010 *P*-Values on 2010 Logit Values



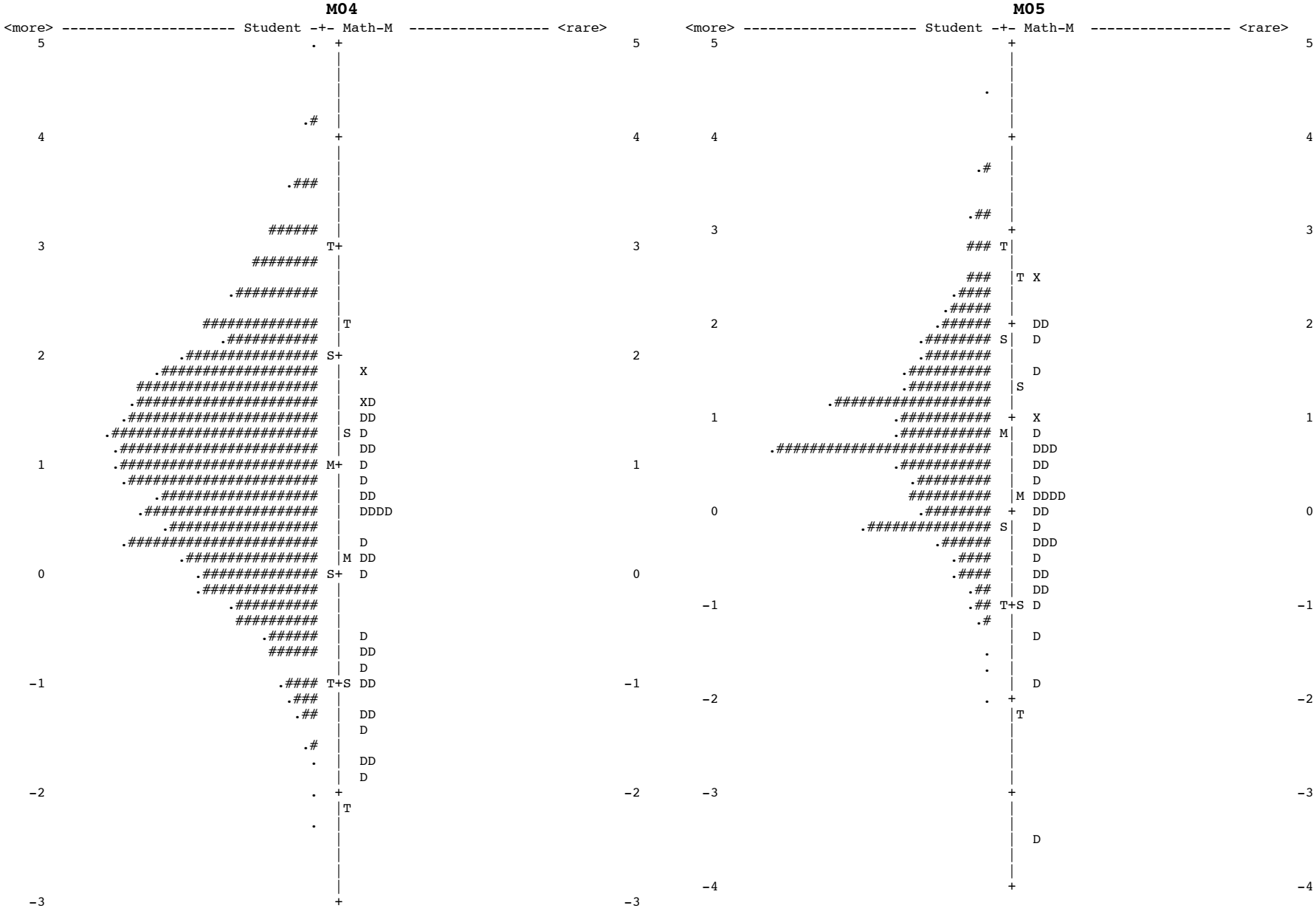


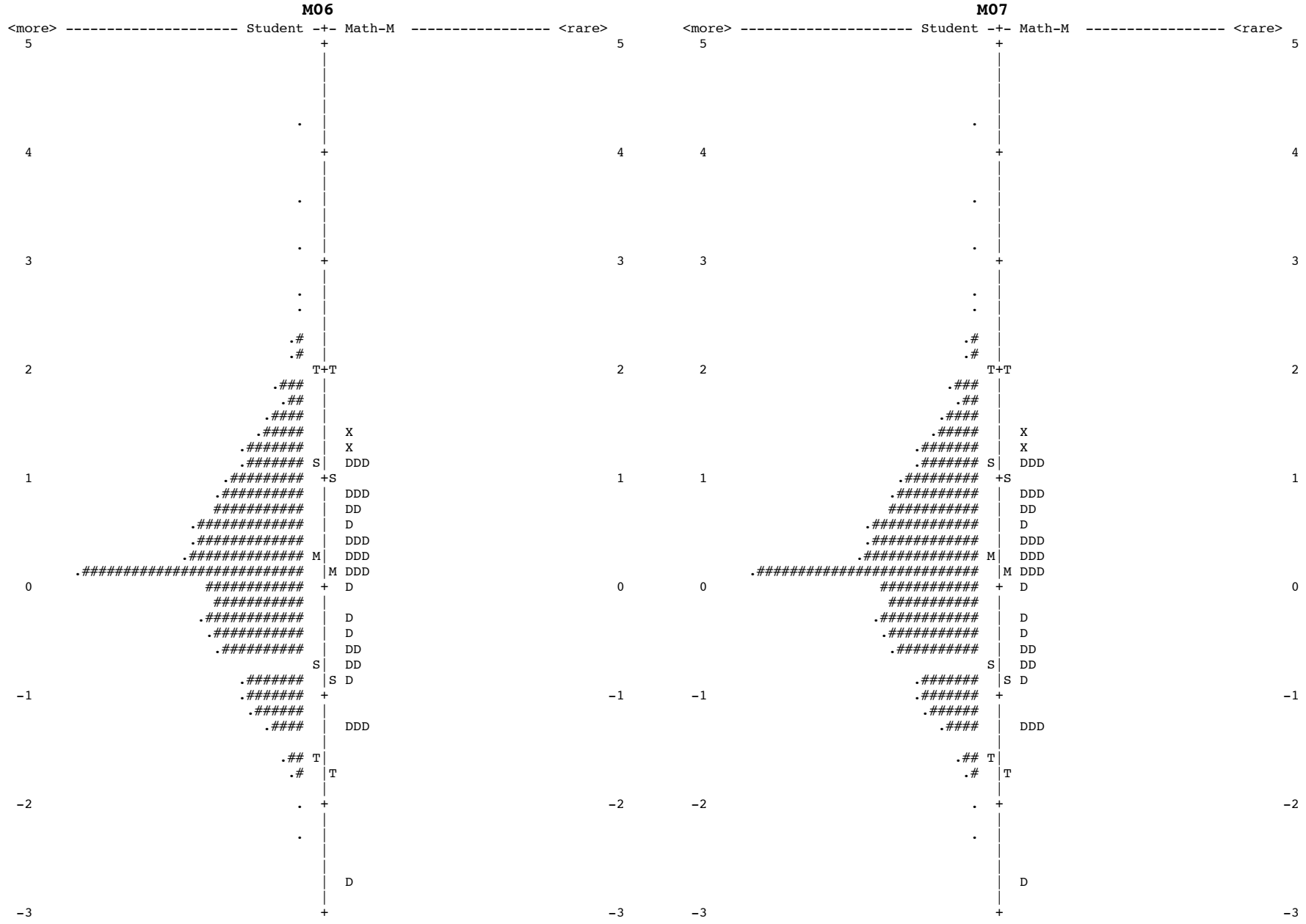
Item Difficulty-Student Ability Maps

The distributions of the Rasch item logits (item difficulty estimates) are shown on the item difficulty-student ability maps presented in Figure 12–2. In each item-student map, markers on the left-hand side represent student ability values, whereas markers on the right-hand side represent item difficulty parameter estimates. As noted earlier, the Rasch model enables placement of both items and students on the same scale. Consequently, one can easily visualize information about how the difficulty of the test items related to the ability distribution of students who took the test. The students located in the upper left quadrant of any given plot have relatively more ability. Items in the lower right quadrant are relatively easier. High ability students have higher probabilities of correctly answering easier items. Similarly, low ability students (in lower left quadrant of any given plot) have lower probabilities of answering harder items (in upper right quadrant).

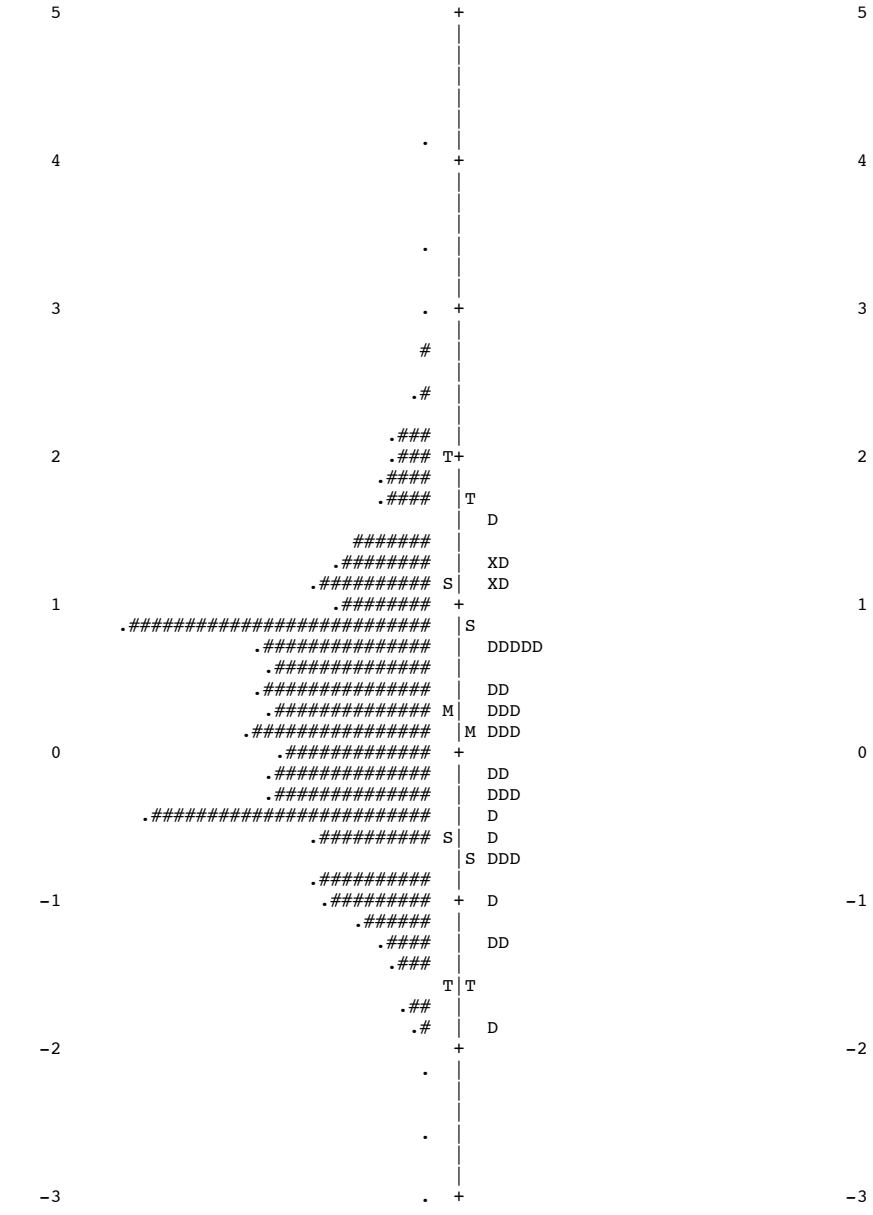
Overall, the distribution of student ability was roughly comparable to the distribution of item difficulties. The mean ability of the students was comparable to the mean item difficulty. The range of student ability and item logit was also comparable. It is also important to understand where the items are providing more accurate measurement (e.g., near the cut scores or away from the cut scores). This issue is addressed more fully in Chapter Eighteen (see Figure 18–2). The OE items ("X's") were relatively more difficult than the MC items ("D's"). However, the OEs provide more information for higher ability students.

Figure 12–2. Item-Student Maps

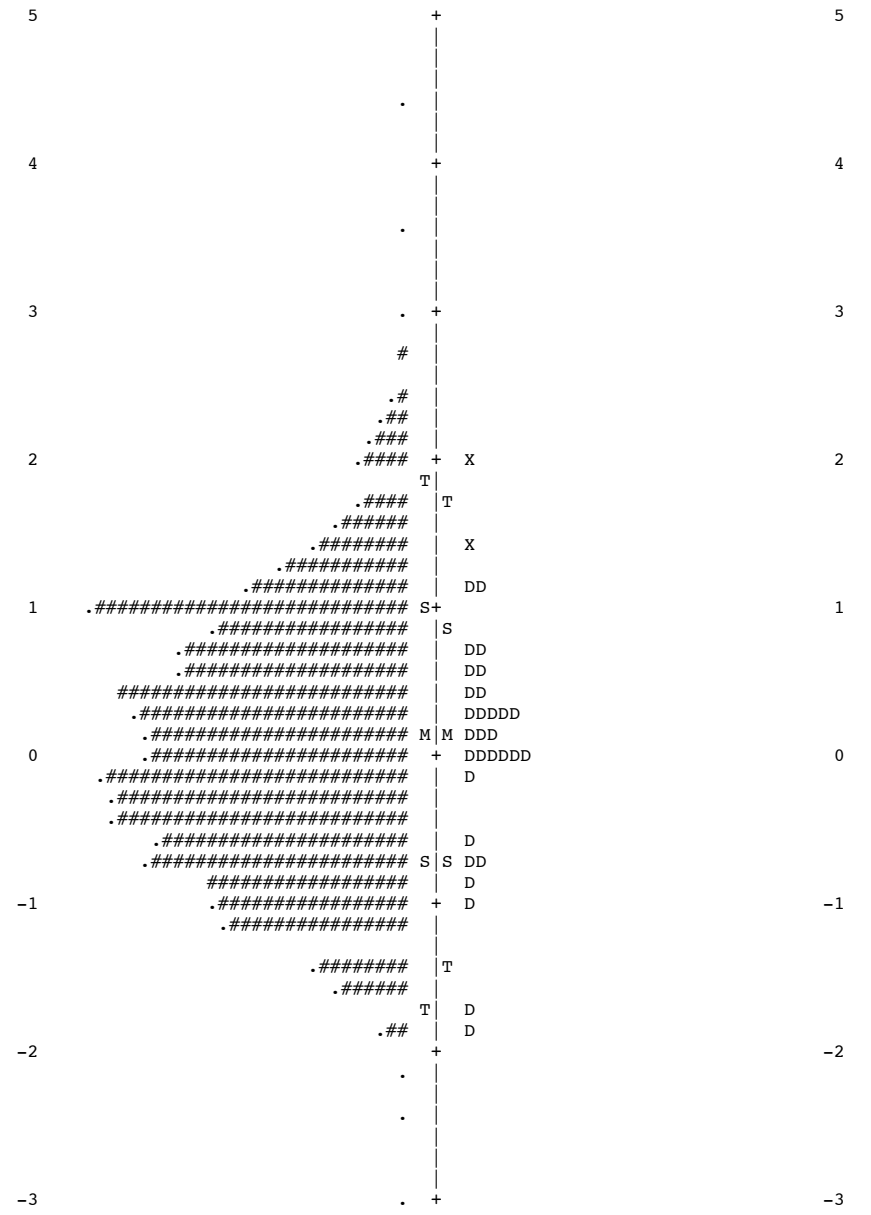



M06
<more> Student -+- Math-M <rare>
M07
<more> Student -+- Math-M <rare>

M08


<more> Student -+- Math-M <rare>
5 + 5
. |
4 + 4
. |
3 . + 3
|
.# |
.### |
2 .### T+ 2
.#### |
.#### |T
| D
####### |
.######## | XD
.########## S| XD
1 .######## + 1
.########################## |S
.############### | DDDDD
.############## |
.############### | DD
.############## M| DDD
.################ |M DDD
0 .############## + 0
.############## | DD
.############## | DDD
.########################## | D
.########## S| D
|S DDD
.########## |
-1 .######### + D -1
.###### |
.#### | DD
.### |
T|T
.## |
.# | D
-2 + -2
. |
. |
-3 . + -3

M11


<more> Student -+- Math-M <rare>
5 + 5
. |
4 + 4
. |
3 . + 3
|
.# |
.## |
.### |
2 .#### + X 2
T|
.#### |T
.###### |
.######## | X
.########## |
.############## | DD
1 .############################ S+ 1
.################## |S
.#################### | DD
.#################### | DD
########################## | DD
.######################## | DDDDD
.######################## M|M DDD
0 .######################## + DDDDDD 0
.########################## | D
.########################## |
.########################## |
.####################### | D
.####################### S|S DD
################## | D
-1 .################ + D -1
.################ |
|
.######## |T
.###### |
T| D
.## | D
-2 + -2
. |
. |
-3 . + -3

Chapter Thirteen: Performance Level Setting

The performance level setting for PSSA-M Mathematics tests was conducted by Data Recognition Corporation (DRC) using the Ordered-Item Booklet (OIB) Angoff (Yes/No) method during a workshop held in Lancaster, Pennsylvania on June 22–24, 2010. A brief summary of the methodology and results is provided below. Full details of the performance level setting event can be found in the following technical report:

> *Standard Setting Technical Report for the 2010 Pennsylvania System of School Assessment-Modified Mathematics Assessment (Data Recognition, 2010)*[7]

SUMMARY

Figure 13–1 presents general information about this event. The purpose of the event was to establish three cut scores for the modified Mathematics assessments at Grades 4–8 and 11. The three cut scores place students into four performance levels: Below Basic-M, Basic-M, Proficient-M, and Advanced-M. The Pennsylvania Department of Education (PDE) recruited panel members from across the state and targeted the educators who have experience in Mathematics at a specific grade level, and are knowledgeable about modified content standards. A total of 34 participants were selected and 11 of them were special educators with experience with IEP students. These 34 panelists were assigned into one or two grade-level panels based on their experience. The percent of special educators was at least 40% in each panel.

The entire PSSA-M Mathematics performance level event included:

1. Training session.
2. Phase I: three rounds of the OIB Angoff Yes/No procedure for each grade level.

   Panelists were asked to review the items, one by one, in order of their difficulty. For each item, the panelists were instructed to think about whether the borderline student at each given performance level would answer an item correctly. Only a dichotomous "Yes" or "No" response was required for each item. Since the items were rank ordered in difficulty from the easiest to hardest, panelists were expected to make more "Yes" judgments at the beginning of the OIB (easier items) and more "No" judgments at the end of OIB (more difficult items). Ideally, each panelist might have come to a point where all their "Yes" answers would change to "No" answers. However, the actual pattern of ratings could differ as the item ordering only provided useful information to the panelists, not an absolute rule about answering "Yes" or "No."

   An Evaluation Form was used to collect validity evidence from the panelists. The results were positive and suggested that the event processes went efficiently. In addition, the panelists had high confidence about the final recommended cut scores.
3. Phase II: vertical articulation across 6 grades (Grades 4–8 and 11).

Table 13–1 summarizes the final recommended raw, theta, and scale score cuts and their associated conditional standard errors of measurement (CSEMs) for each grade. On July 1, 2010 the Pennsylvania State Board of Education approved the panelists' recommendations at all grades.

[7] This report is available upon request from PDE at 1-717-705-2343.

Figure 13–1. General Information about PSSA-M Mathematics Performance Level Setting

Official Title	Performance Levels Setting for PSSA-M Mathematics		
Event Dates	June 22–24, 2010		
Methodology	Ordered-Item Booklet Angoff (Yes/No)		
Number of Performance Levels	Four	**Performance Level Names**	Below Basic-M Basic-M Proficient-M Advanced-M
Content Area(s)	Modified Math	**Grades**	4, 5, 6, 7, 8, and 11
Panelists	34 Total at least 8 per Group Note: Percent of Special Ed. was at least 40% in each group	**Tables**	1 Table per Group
Rounds	Three (3) plus Articulation	Impact Groups	Total Group

Table 13–1. Final Cut Scores and Conditional Standard Errors of Measurement (CSEM) for Scale Score Cuts

Grade	Below Basic-M/Basic-M				Basic-M/Proficient-M				Proficient-M/Advanced-M			
	Raw	Theta	Scale	CSEM	Raw	Theta	Scale	CSEM	Raw	Theta	Scale	CSEM
4	12	-0.5891	1150	34	22	0.8935	1275	31	29	1.8540	1356	32
5	12	-0.5352	1150	36	22	0.8640	1275	32	30	1.9734	1374	37
6	10	-0.9606	1150	39	19	0.3441	1275	35	27	1.4543	1381	37
7	10	-0.8442	1150	35	21	0.5878	1275	29	30	1.6086	1364	32
8	11	-0.7858	1150	36	21	0.5407	1275	34	30	1.8139	1395	39
11	12	-0.5492	1150	44	20	0.5317	1275	42	28	1.6389	1403	44

Chapter Fourteen: Scaling

The purpose of a scaling analysis is to create a score scale. Scaling is used to transform test score values onto a scale that can be more easily interpreted by users. For the PSSA-M, the resulting scale scores will be used for score reporting and performance level classification. The PSSA-M classifies students into four achievement levels: Below Basic-M, Basic-M, Proficient-M, and Advanced-M.

SCALED SCORES

Individual student scores are reported as scaled scores. However, they are initially estimated as Rasch abilities (more information on the Rasch model is given in Chapter Twelve). Generally, scaled scores are preferred over Rasch ability values for reporting purposes. One issue is that Rasch ability values are on a scale that includes negative and decimal values. By transforming the Rasch ability values to scaled scores, all reported values can become positive integers. Scaled scores are usually obtained through some linear transformation of the Rasch ability values. The linear transformations used for the PSSA-M produces numeric values with three or four digits that are unit interval scaled scores. Each grade and subject has its own unique PSSA-M scaled score. Having positive scores with no decimals makes more sense to parents and students. Since Rasch ability values are comparative after linking to the base year, the transformed scaled scores have a common scale across years, even though the corresponding raw scores may differ. (Linking is discussed further in Chapter Fifteen.)

Essentially, PSSA-M scaled scores are derived through a two step process. First, there is a nonlinear transformation that converts number correct scores to Rasch ability logits. Next, a linear transformation is used to convert logits to scaled scores. These, and some additional considerations (e.g., rounding rules) are discussed further below.

Definition of Scoreability

Answer documents are considered scoreable if they meet the criteria for inclusion in the data files (see Chapter Nine). All omit (i.e., no response) and multiple marks (i.e., more than one response selected, without machine-discernable erasures) are scored as zeroes.

WINSTEPS Scaling

Parameter estimates are derived using the WINSTEPS 3.54 computer program (Wright and Linacre, 2003), which employs unconditional (UCON), joint-maximum-likelihood estimation (JMLE). WINSTEPS provides a conversion table that maps raw scores to logits (Rasch ability estimates). The logits are transformed to scaled scores as discussed below. Every year, each test is scaled separately – then linked (see Chapter 15).

ZERO AND PERFECT SCORES

WINSTEPS does not provide a direct ability estimate for zero (i.e., no points earned) or perfect (i.e., all points earned) raw scores. However, WINSTEPS has a default procedure for estimating such extreme scores and this was used for the PSSA-M. Essentially, a fractional raw score (i.e., a value less than one) was added to zero scores and subtracted from perfect scores to determine the corresponding logit values for these extreme scores.

Linear Transformation Formulas

PSSA-M scaled scores are obtained through a linear transformation of the Rasch ability estimates ($\hat{\theta}$). Specifically,

$$SS = m\hat{\theta} + b,$$

where m is the slope and b is the intercept. The slopes and intercepts for deriving PSSA-M scaled scores are provided in Table 14–1. For reference purposes, the PSSA-M theta cut scores have been reproduced in this table as well.

Rounding

The linearly transformed scaled scores are generally rounded to the nearest integer value for reporting purposes. Values greater than or equal to 0.50 are rounded up. Values less than 0.50 are rounded down. However, at each performance level cut point, scores are rounded up (even if less than 0.50) if this action would put the rounded score into a higher performance level. As an example, the Grade 4 mathematics Proficient-M cut score (in scaled score units) is 1275. If there had been a raw score that converted to an unrounded scaled score of 1274.20, this scaled score would have been rounded up to 1275 for reporting purposes.

Lowest Obtainable Scaled Scores

All PSSA-M mathematics tests have a lowest obtainable scale score (LOSS) of 1075. LOSS values are documented in Table 14–2. See tables in Appendix I for LOSS *n*-counts.

Highest Obtainable Scaled Scores

A highest obtainable scale score (HOSS) is not set for the PSSA-M. Thus, the maximum possible scaled score value is allowed to float for each subject and grade. The upper bound varies from year to year, depending on the difficulty of the test form. Table 14–2 shows the maximum possible observed score for the current year's test. (Note: It may be that no student actually earned the maximum possible.) See tables in Appendix I for HOSS *n*-counts.

RAW-SCORE TO SCALED-SCORE TABLES

Raw-to-scaled-score tables can be found in Appendix I.

Table 14–1. PSSA-M Cut Scores (on θ metric), Intercept, and Slope by Grade and Subject Area

		θ Cuts				
Grade	**Content**	**BB/B**	**B/P**	**P/A**	**Intercept**	**Slope**
4	Mathematics	-0.5891	0.8935	1.8540	1199.67	84.31
5		-0.5352	0.8640	1.9734	1197.81	89.34
6		-0.9606	0.3441	1.4543	1242.03	95.81
7		-0.8442	0.5878	1.6086	1223.69	87.29
8		-0.7858	0.5407	1.8139	1224.05	94.23
11		-0.5492	0.5317	1.6389	1213.51	115.64

Notes. Linear Transformation Intercepts and Slopes are used to derive the Scaled Scores.
BB = Below Basic-M; B = Basic-M; P = Proficient-M, and A = Advanced-M

Table 14–2. PSSA-M Scaled Score Cuts for each Performance Level by Grade and Subject Area

Grade	Content	Min	Scaled Score Cuts[1]			Max[2]
			BB/B	B/P	P/A	
4	Mathematics	1075	1150	1275	1356	1655
5		1075	1150	1275	1374	1712
6		1075	1150	1275	1381	1772
7		1075	1150	1275	1364	1663
8		1075	1150	1275	1395	1730
11		1075	1150	1275	1403	1890

Notes. [1] BB = Below Basic-M; B = Basic-M; P = Proficient-M, and A = Advanced-M.
[2] Scaled Score Maximum Values are unique for the current year's test.

DOMAIN SCORE STRENGTH PROFILE

The following process was followed to derive the domain score strength profile:

- The items for each domain were identified.
- WINSTEPS runs were undertaken that anchored the logit values for each domain's items to get the raw-to-logit score table for each domain. This is sometimes referred to as fixed item parameter scaling.
- The appropriate linear transformation (based on content and grade from Table 14–1) were applied to the logit values to derive domain scaled scores.

The domain scale scores were categorized as follows: L=Low (equivalent to Below Basic-M and Basic-M); M=Medium (equivalent to Proficient-M); H=High (equivalent to Advanced-M). The maximum possible domain score was converted to H in cases where no domain scaled score equaled or exceeded the Advanced-M scaled score cut. See Chapter 16 for more information on domain scores and how they are used in score reports.

Chapter Fifteen: Linking

FORWARD

This was the first administration of the PSSA-M mathematics test. As such, no linking/equating was required; therefore, Tables 15–2, 15–3, and 15–4 show data as TBD (To Be Determined). Linking/equating data will be determined in 2011. The purpose of this chapter is to preview various linking issues and procedures that will be encountered during the second administration of PSSA-M mathematics in 2011.

INTRODUCTION

In large-scale testing programs, it is a common practice to have different item sets appear in test forms within and/or across years. Linking operational scores from the different test forms ensures that all forms for a given grade and subject area provide comparable scores. Consequently, students are not given an unfair advantage or disadvantage because the particular test form they took is easier or harder than a test form taken by other students.

When multiple forms are administered, students who have the same ability could obtain different raw (i.e., number-correct) scores over the different test forms. As discussed further in Chapter Sixteen, raw scores can only be interpreted relative to the particular set of items used. This is because item difficulty distributions are nearly always different across different item sets.

Just like raw scores are not necessarily interchangeable across forms, Item Response Theory (IRT) item parameters and ability estimates are not necessarily interchangeable across separate calibration runs either. Application of an IRT scale linking methodology is usually required to place the item parameters and student ability estimates on the same scale as other forms. (As cautioned earlier, the success of these methods depends on how well the IRT assumptions are met.) The IRT model used for the PSSA-M is the Rasch Partial Credit Model (RPCM; Masters, 1982). Further descriptions of the RPCM are given in Chapter Twelve.

A chained linking design will be utilized for the PSSA-M operational scores in mathematics. Scores from the new test (2011) form will be linked to the scale of the old test form (2010). The chain originates from the test's base form, which is used as the reference for calibrating all items in the item pool. The base form is usually the form upon which the cut scores were established (see Chapter Thirteen). When the item parameters from the new test are placed on the bank's scale, the resulting scaled scores for the new test form will be the same as the scaled scores of the base form. In order to compare students' PSSA-M scaled scores across different years, the new operational items need to be placed on the bank scale via scale linking.

This chapter begins with a brief summary of the expected PSSA-M linking procedures. This is followed by a more detailed explanation of selected design elements and processes.

BRIEF SUMMARY OF THE PSSA-M LINKING PROCEDURE

The first two steps concern calibration of the MC and OE items, which is considered as within-year linking in this chapter.

1. Calibrate selected MC items in an unanchored run:
    - Include all MC items in the core operational section (OP MCs).

2. Calibrate selected open-ended (OE) items in an anchored run by putting them on the MC item scale from Step 1:

- Include all OE items in the Core section (OP OE).
- Fix all MC items from Step 1.

3. Evaluate the stability of the linking items using Robust *Z*:

- Include all core linking (LK) items – LK MC.
- Calculate Robust *Z* for each item in the linking.

Once the above calculations were made, the following guidelines are used in determining possible sets of linking items used for the equating:

- Items with an absolute value of Robust *Z* exceeding 1.645 may be considered for exclusion.
- No more than 20 percent of the pool of linking items may be considered for exclusion.
- The ratio of the standard deviations of previous year and current Rasch difficulties should be in the 90 to 110 percent range.
- The correlation of previous year and current year Rasch difficulties is greater than 0.95.

Final decisions about the linking items follow these rules:

- Drop items that DRC identified as having a large Robust *Z* and were out of sequence because they were pulled from a separate FT form.
- If an item has been changed in any way from the previous year, it may no longer be used for linking.

Scatterplots of the linking item difficulties (logits) are constructed (i.e., the current year values are plotted against those from the prior year). Ideally, these plots should have a strong linear trend. Items straying from the trend line did not perform in the same way in both years. As noted above, items that departed significantly from this are further evaluated. An example is shown in Figure 15–1.

4. Calculate the mean shift over MC linking items using item difficulties:

- Include all core linking (LK) items – LK MC.

5. Apply the mean shift to the item parameters calibrated in Steps 1 and 2:

- All OP items (OP MC + OP OE).

6. Scale the operational test by fixing all operational (OP) items obtained in Step (5):

- The result from this step is a Raw-to-Logit (Rasch Ability) table.

7. Apply the appropriate linear transformation to the logit values to derive the scaled scores and SEMs:

- The result from this step is a Raw-to-Scale table.

PSSA-M MATHEMATICS

Data Collection Design

The item status codes used in the IDEAS item banking system are given in Table 15–1. For brevity, these codes are used for the remainder of this chapter.

The link between years will be based on the core linking (LK). These items will have been used in previous administrations. The LK items are used in approximately the same context. The *same context* in this situation means the items are not altered in any way, they appeared in about the same position in the booklet, and they are administered at about the same time of year.

The equivalence of student samples across years cannot be assumed. Further, the same item can have different properties in different years because of changes in the item's position or changes in the students' experiences. Consequently, between-year linking requires considerable scrutiny.

The linking design employed for PSSA-M is often referred to as a common-item nonequivalent groups (CINEG) design. Test forms will contain a set of common items, called core linking (LK) items, which serve as anchors for comparison of test forms across years. LK items are internal anchor items (i.e., contribute to student test scores).

All LK items are common between years since all came from the prior year's administration. The proportion of the LK items may be different depending on the subject and grade. These will be summarized in a manner similar to Table 15–2.

Table 15–1. Item Status Codes in IDEAS

Item	Comments	Code in IDEAS
Core	Include core linking (i.e., anchor) items and unique core items	OP
Core linking	Linking items in the core section which include MC and OE items.	LK

Table 15–2. 2011 PSSA-M Linking Designs: MATHEMATICS

Grade	Core		Core Links
	MCs (1 pt)	OEs (4 pts)	MC(1)
4	30	2	TBD
5	30	2	TBD
6	30	2	TBD
7	30	2	TBD
8	30	2	TBD
11	30	2	TBD

Note. This was the first administration of the PSSA-M mathematics test. As such no linking/equating was required; therefore, data shows as TBD.

LINKING METHOD FOR PSSA-M MATHEMATICS

The overall linking procedure was summarized at the start of this chapter. In review, the first step is to conduct a within-year linking to place all item parameters on the same scale. This is accomplished by first concurrently calibrating all OP (including LK) MC items. Next, the resulting MC item parameters are anchored in WINSTEPS while all OE items in the operational section (including OP LKs) items are calibrated. At this point all item parameters are on a unique scale. Between-year linking is required to place the items on the bank scale.

Between-year linking will utilize the current year's LK item parameters and their banked counterparts. The scale transformation methodology used for PSSA-M is known as the mean-shift procedure. After evaluating the robustness of the link by identifying items that do not maintain their relative difficulty across years, the difference between the new and banked parameters will be determined. The mean of the differences is then used to statistically adjust the new parameters to the bank scale. The final (linking) item parameters are then used to estimate student abilities, which are, in turn, transformed to scaled scores. (Transformation formulas are provided in Chapter Fourteen.)

RESULTS SUMMARY

Table 15–3 will show the numbers of linking items started with and ended with the shift parameters associated with those over the two years, and the correlation of item difficulties across years for each grade/content area.

Dropped Items

Table 15–4 will illustrate a summary of linking items dropped from final equating. In some cases items may be dropped. A change in their location may affect their performance or changes may be discovered in the items that have the potential to affect item difficulty. Table 15–4 shows the item ID, item type, source year (where the item was pulled from), item sequence in previous form and in 2009, item difficulty (i.e., logits) in previous form and in 2009, and the corresponding Robust *Z* values.

Table 15–3. Summary Data for Linking Items

		Initial Counts		Final Counts				
Subject	**Grade**	**MC**	**OE**	**MC**	**OE**	**Initial Shift**	**Final Shift**	**Final Correlation**
Math	**4**	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	5	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	6	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	7	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	8	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	11	TBD	TBD	TBD	TBD	TBD	TBD	TBD

Note. This was the first administration of the PSSA-M mathematics test. As such no linking/equating was required; therefore, data shows as TBD. Starting in the 2011 PSSA-M Technical Report, an appendix tracking the historical performance levels will be provided.

Table 15–4. Summary of Linking Items Dropped from Final Equating

Subject	Grade	Item Type	Source Year	Seq Prev	Seq 2009	Logit Prev	Logit 2009	Robust Z Value
Math	4	TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
	5	TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
	6	TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
	7	TBD	TBD	TBD	TBD	TBD	TBD	TBD
	8	TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
		TBD	TBD	TBD	TBD	TBD	TBD	TBD
	11	TBD	TBD	TBD	TBD	TBD	TBD	TBD

Note. This was the first administration of the PSSA-M mathematics test. As such no linking/equating was required; therefore, data shows as TBD.

The previous and current values for item sequence, *p*-values, and logits are provided. An appendix will provide the mean raw and scaled score points across years. Together, the appendices and tables will provide a summary of how the items and tests change across years.

VISUALIZATION SUPPLEMENT

As noted earlier, between-year linking requires considerable scrutiny. This is partly because student samples are not equivalent across years. Additionally, identical items can have different properties in different years because of changes in any given item's context or changes in the students' experiences. Since the linking process forces the logit difficulties for the linking items to have the same mean in the new year as they did in the old year, the current-year logit item difficulties will be displaced from the estimates they would have received from an independent calibration. The size of the displacements reflects the difference, if any, in the origins. The variation among the displacements corresponds to the approximate size of the standard errors for the items. The graphs in Figure 15–1 should help visualize this information.

Graphs

Future Technical Reports will emphasize figures to help visualize the across-year differences in linking items at each grade. This section presents four types of figures, three of which illustrate the stability between the old (i.e., banked) and new item data:

1. Scatterplot of new-year *p*-values on old-year *p*-values.
2. Scatterplot of new-year logits on old-year logits.
3. Scatterplot of old and new *p*-values on new logits.
4. Test Characteristic Curves (TCCs) for the linked score distribution.

All four plots will be presented for each test. Each plot is described further below.

NEW-YEAR *P*-VALUES ON OLD-YEAR *P*-VALUES

The top left-hand plot in Figure 15–1 describes the relationship between the item *p*-values for the two years. The data points in these plots should have a clear trend where the vertical axis values rise as the horizontal axis values increase (i.e, as one moves from left to right). If the *p*-values for both years were correlated at 1.0, one would expect the relationship to fall on a straight line. Generally, linking items are not perfectly stable across years, so some scatter is expected. The extent to which the trend does not pass through the origin indicates a change in student performance.

Many test score users are familiar with the *p*-value metric, which is why these charts are provided. However, the logit charts discussed below have advantages for visualizing this trend data.

NEW-YEAR LOGITS ON OLD-YEAR LOGITS

The top right-hand plot in Figure 15–1 focuses on the logit difficulties. It shows more clearly the relationship between new and old-year item difficulties. Logit plots often provide more defined trends, but still can present varying degrees of scatter and in some instances reveal outlier data points.

OLD- AND NEW-YEAR *P*-VALUES ON NEW-YEAR LOGITS

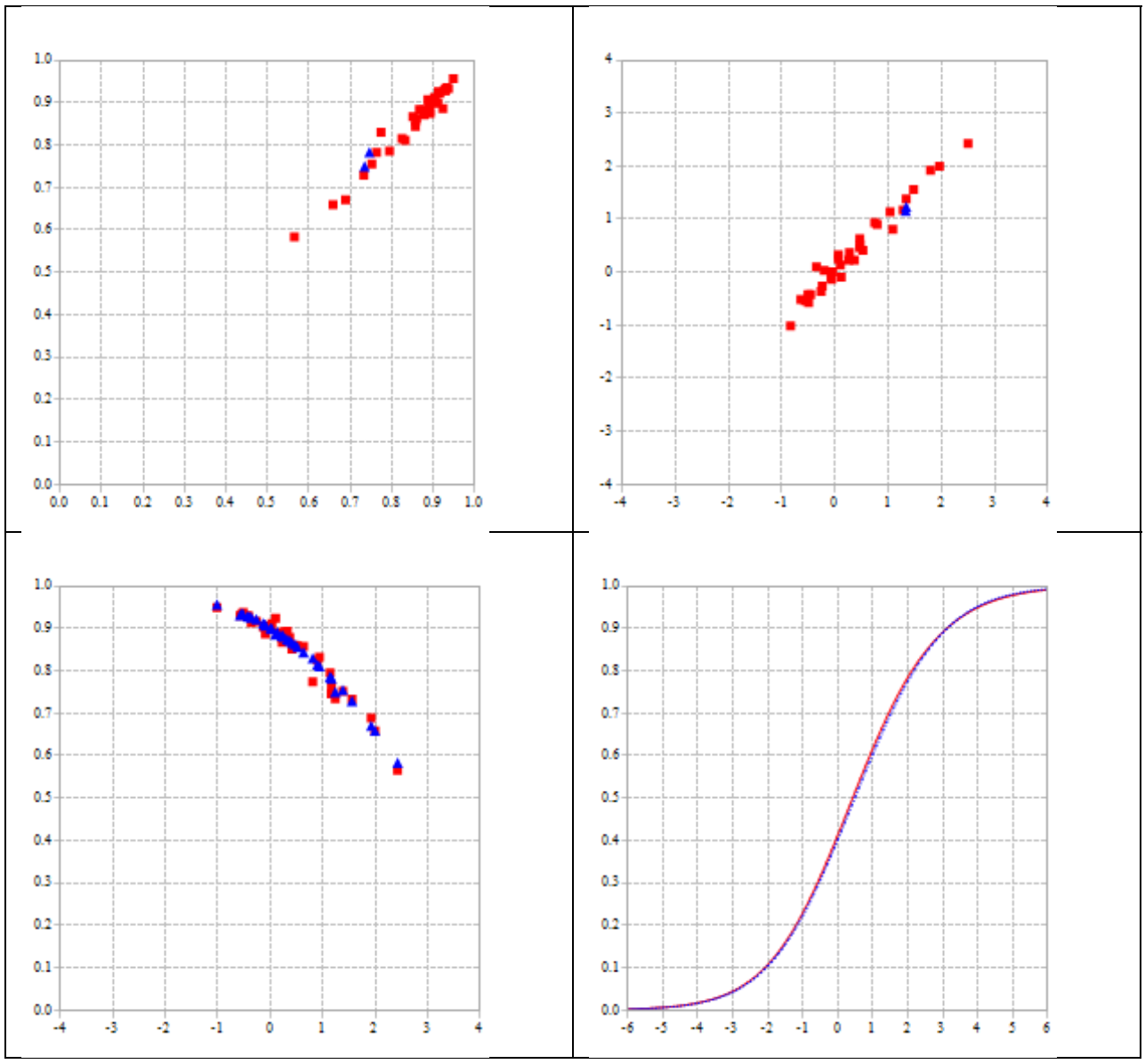
Plotting *p*-values against logit difficulties across years is not as reliable as it is within a year. Within year, the *p*-values-on-logit plot should be a single curved line. (See plots in Chapter Twelve as examples.) The corresponding between-year plots could have separate lines for each year. The difference between the two lines is a reflection of the adjustment (positive or negative) that is required to link the two item sets.

In Figure 15–1, the two lines sloping downward toward the right relate item *p*-values for the two years to the new-year logit difficulties. Again, these graphs have some similarity with the set of graphs that were part Chapter Twelve. Both show the *p*-value-on-logit relationship, with the Chapter Twelve plots showing the current year *p*-values for operational items while Figure 15–1 shows the *p*-values for linking items from the current year and the prior year. Both illustrate the curvilinear relationship required by the model, with low *p*-values being translated into high logit difficulties and high *p*-values being converted into low logit difficulties.

TEST CHARACTERISTIC CURVES

The old and new-year Test Characteristic Curves (TCCs) by grade and subject are shown in the figures[8]. TCCs show the similarity between the new- and old-year tests in terms of difficulty in the logit metric (new-year results are for the final, linked values). Assuming equal numbers of items for the two years, curves that are close to being coincident will translate into similar raw-score cut points. With extreme differences in test difficulties, some loss of precision and reliability may result.

Figure 15–1. Item Stability Plots and Test Characteristic Curves



Note. For Illustrative Purposes Only.

[8] In the TCC figures, the Y-Axis Probability represents total test raw score expressed on a proportion-correct metric.

Chapter Sixteen: Scores and Score Reports

This chapter provides information about the scores provided for the PSSA-M (e.g., scaled scores, performance levels, and strand scores), how they are presented on score reports, and appropriate and inappropriate uses of the scores.

SCORING THE PSSA-M

PSSA-M items are comprised of multiple-choice (MC) and open-ended (OE) items. Each correct response to a MC item receives a score of 1. Incorrect responses receive a score of 0. Scores on OE items range from 0–4, depending on the grade and subject area. Table 16–1 summarizes the types of items used on each subject-area test. More detailed information about the various item types is provided in Chapter Three.

Table 16–1. Item Types Used by Subject Area

	Subject
Item Type	**Mathematics**
Multiple-Choice (1 point)	■
Open-Ended (2 point)	
Open-Ended (3 point)	
Open-Ended (4 point)	■
Prompts (4 point)	

DESCRIPTION OF TOTAL TEST SCORES

Different types of scores have been developed for PSSA-M reporting. Since the underlying properties of these scores are not necessarily the same, the particular scores used depend on the purposes for which the test has been given. The following types of scores are provided for reporting overall performance on each PSSA-M subject-area test:

- Raw Scores
- Scaled Scores
- Performance Levels

Raw Scores

A raw score is the number of points a student earned over the operational MC and OE items. By itself, the raw score has very limited utility. One limitation is that it can only be interpreted with reference to the total number of items on a subject-area test (e.g., a raw score of 15 on a 20 item test is different from a raw score of 15 on a 30 item test). In addition, raw scores depend on the difficulty of test items across test forms (e.g., a raw score of 15 on a test with 20 easy items is different from a raw score of 15 on a test with 20 difficult items). Because the difficulty of the items on a test can change from year to year, raw scores should not be compared across tests or administrations.

Scaled Scores

Scaled scores were introduced in Chapter Fourteen and additional information is provided there including information on the development of the PSSA-M scaled score system. In the simplest sense, a scaled score is a transformed number-correct score. The specifics of the transformation processes for the PSSA-M were also discussed in Chapter Fourteen. When all students take the same items, as with the operational items on the PSSA-M, the more points the student earns, the higher the associated scaled score will be.

The value of switching to the more abstract scaled score metric lies in the fact that it produces more general and equitable results. As noted above, a raw score of 30 is meaningless unless the total points possible is known. The difficulty of the test items was also mentioned as an additional challenge with interpreting raw scores. Number-correct scores are transformed to scaled scores to remove the effects of test length and item difficulty. (Strictly speaking, transformation of number-correct scores to percent-correct scores would also remove the effect of test length, but it would do nothing to adjust for the difficulty of the items.)

Another advantage of scaled scores is that they lend themselves to interpretations at what is referred to as an interval level, while raw scores do not. Interval-level scales allow an interpretation of a scaled score difference of 5 points to be the same whether the scores are 1295 vs. 1300 or 1445 vs. 1450. Raw score differences, in this context, cannot be interpreted in this manner and are thus neither generalizable nor equitable.

A scaled score of 1300—or any other value for a particular grade and content area test, like Grade 4 mathematics—should have the same absolute meaning in the current year as it had in previous years, when test scores are properly linked across years. More importantly, an increase in the scaled score for Grade 4 mathematics from last year to the current year means that student performance improved;[9] it does not say anything about whether this year's test is easier or harder than last year's test. To make these interpretations requires no information about the length or the difficulty of the test in either year, although these variables are essential for the process of deriving the scaled scores.

There is considerable auxiliary information presented in this report that might aid in further contextualizing PSSA-M scaled scores. Refer to the following information:

- Chapter Fourteen provides information on the development of the PSSA-M scaled score system, including transformation formulas, rounding rules, and general scale characteristics (e.g., minimum values).
- Chapter Seventeen provides total test score statistics. In particular, Table 17–1 lists the scaled score means and standard deviations for this year's test results.

Performance Levels

PSSA-M results are also reported using four Performance Levels: Below Basic-M, Basic-M, Proficient-M, and Advanced-M. The cut scores on the scaled score metric (i.e., the lowest possible scaled score to enter the Basic-M, Proficient-M, and Advanced-M levels) were presented earlier in this report. However, the information is repeated in Table 16–2 for convenience.

[9] This example is not an endorsement of conducting a trend analysis with just two years of results. Further, small differences may not be statistically or practically significant.

Table 16–2. PSSA-M Scaled Score Cuts for each Performance Level by Grade and Subject Area

Grade	Content	Min	Scale Score Cuts[1]			Max[2]
			BB/B	B/P	P/A	
4	Mathematics	1075	1150	1275	1356	1655
5		1075	1150	1275	1374	1712
6		1075	1150	1275	1381	1772
7		1075	1150	1275	1364	1663
8		1075	1150	1275	1395	1730
11		1075	1150	1275	1403	1890

Notes. [1] BB = Below Basic-M; B = Basic-M; P = Proficient-M; and A = Advanced-M.
[2] Scaled Score Maximum Values are unique for the current year's test.

Performance levels descriptors (PLDs) are another way to attach meaning to the scaled score metric. They associate precise quantitative ranges of scaled scores with verbal, qualitative descriptions of student status. While much less precise, the qualitative description of the levels is one way for parents and teachers to interpret the student scores. They are also useful in assessing the status of the school. The Pennsylvania General Performance Level Descriptors (PLDs), as developed by PDE and teacher panels are given below. These are also included on student score reports.

- **Advanced-M:** More than satisfactory academic performance on grade level standards as measured on an assessment with modifications to the general assessment. Advanced-M work indicates a more than adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.
- **Proficient-M:** Satisfactory academic performance on grade level standards as measured on an assessment with modifications to the general assessment. Proficient-M work indicates an adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.
- **Basic-M:** Academic performance approaching satisfactory on grade level standards as measured on an assessment with modifications to the general assessment. Basic-M work indicates a less than adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.
- **Below Basic-M:** Unsatisfactory academic performance on grade level skills as measured on an assessment with modifications to the general assessment. Below Basic-M work indicates little understanding of the skills included in the Pennsylvania Assessment Anchor Content Standards.

DESCRIPTION OF REPORTING CATEGORY SCORES

The following types of scores are provided for PSSA-M reporting category scores:

- Reporting Category Scores (i.e., Strand Scores)
- Strength Profile

Reporting Category Scores (Strand Scores)

A reporting category score describes a student's or school/district's performance on a particular reporting category (i.e., content standard defined in the test). For the PSSA-M, reporting category scores are raw scores, indicating the points a student or a school/district earned for that reporting category. (Attributes of raw scores are described earlier in this chapter.)

Reporting category scores cannot be compared across years because they are not statistically linked. Also, it is not advisable to compare reporting category raw scores even within the same form because some reporting categories may contain items that are easier or more difficult than other reporting categories; the strength profile, discussed below, mitigates this problem to some degree. A greater concern is the low reliability of many of these scores, especially for strand scores based on a small number of possible points. Chapter Eighteen provides more information about strand-score reliability.

When compared to other results from the same year, reporting category scores can be somewhat helpful in identifying a group's strengths and weaknesses on the test. For example, it can be informative to compare average reporting category scores of a school against that of another reference's group (e.g., the state average). Hence, reporting category scores can suggest group strengths and weaknesses relative to another reference group. (Challenges pertaining to interpreting results for individual students are discussed below.)

Strength Profile

The strength profile provides another indication of a student's performance within each of the reporting categories. This profile can be used to identify areas in which a student needs to improve and areas in which a student has performed more successfully. Unlike reporting category scores that are reported as raw scores, strength profile scores categorize students into one of three levels: Low, Medium, and High. These categories take into account the difficulty of the items and are based on the same scaling techniques used to derive the PSSA-M scaled scores. (Details regarding the creation of the strength profile are provided in Chapter Fourteen. These scaled scores are not printed on score reports. They only exist to determine whether performance in the reporting categories was Low, Medium, or High.) A Low score on the strength profile indicates performance that is below Proficient-M on the overall PSSA-M scale. A Medium score on the strength profile indicates performance that is comparable to the Proficient-M level on the PSSA-M. A High strength profile indicates performance that is comparable to the Advanced-M level.

APPROPRIATE SCORE USES

Individual Students

Scaled scores on the PSSA-M indicate a student's achievement over the PSSA-M Assessment Anchors and Eligible Content. Scaled scores are primarily used to determine student performance level classifications (i.e., a criterion-referenced inference). Scaled scores that are based on IRT models are typically assumed to be of the interval type; so comparisons may be made on differences in scaled scores. If this assumption holds, then it would be safe to infer for Grade 4 mathematics that the ability difference between an 1110 and 1120 represents the same ability difference that separates 1250 and 1260. Scaled scores can also be used to compare the performance of an individual student to the performance of a similar demographic or subgroup at a school or district. Test score standard errors (discussed in Chapter Eighteen) should be considered.

Groups of Students

Test results can be used to evaluate the performance over time. Mean scaled scores can be compared across administrations within the same grade and subject area to indicate whether student performance is improving across years. Generally, such trend analyses benefit from using mean results from as many test administration years as possible. Different cohorts of students are used (i.e., the same student or students are not tracked across grade levels). All scores can be analyzed within the same subject and grade for any single administration to determine which demographic or program group had, for example, the highest average performance or the highest percent of students beyond the Proficient-M standard.

Reporting category scores can help evaluate academic areas for relative strengths or weaknesses. These category scores provide information to identify areas where further diagnosis is warranted. Generalizations from test results may be made to the specific content domain represented by the academic standards measured in the PSSA-M. However, all instruction and program evaluations should include as much information from other sources as possible to provide a more complete picture of performance.

CAUTIONS FOR SCORE USE

Extreme Error for Extreme Scores

Student scores toward the minimum or maximum ends of the score range will have very large standard errors of measurement and such scores should be viewed very cautiously. The maximum scaled score only provides a very rough estimate of a student's ability. For instance, if the maximum score for the PSSA-M Grade 6 mathematics test were 1800 (it's not, at least for this year) and a student achieves this score, it cannot be determined whether the student could have achieved an even higher scaled score. If the test were 10 items longer a different estimate might have been obtained. Similarly, if the items in a new test were more difficult than the items on a previous administration, the maximum scaled score would likely be higher on the new test because it would take a greater level of achievement to answer the items correctly. In this manner, extreme scaled scores may vary from one administration to the next even if the number of test items does not change. The fluctuation of extreme scaled scores complicates the comparisons of students with scaled scores at the extreme ends of the score distribution. To minimize confusion and potential misinterpretation, the minimum scaled scores possible on the PSSA-M tests have been fixed (see Table 16–2) so they do not change between administrations.

However, the maximum scaled score values have not been fixed. Therefore, caution must be taken when comparing scores at the maximum end of the scale.

Each Test has a Unique Scale

Scaling was conducted for each grade and subject area test separately. Therefore, PSSA-M scale scores should be interpreted only within each content area. PSSA-M scaled scores are not status indicators in the same sense as percentile ranks (or scales that are essentially transformations of percentile ranks) and therefore cannot be used to profile relative strengths and weaknesses across subject areas. As an example, a student with scaled scores of 1300 in Grade 4 reading and 1250 in Grade 4 mathematics does not necessarily imply that the student performed better in reading than in mathematics. The PSSA-M scaled scores do not represent a developmental or vertical scale either. This means that no across-grade comparisons or growth statements for a student are appropriate. For example, a 1250 in Grade 4 mathematics and a 1250 in Grade 5 mathematics does not indicate a student had no achievement growth from Grade 4 to Grade 5 in mathematics.

Strength Profile Caveats

The category labels of Low, Medium, and High were deliberately used instead of any of the PSSA-M performance level names—Below Basic-M, Basic-M, Proficient-M, and Advanced-M—to acknowledge that the PSSA-M cut scores were established on the basis of the total test score. Therefore, the domain categories should not be interpreted the same way as PSSA-M performance levels because they likely do not carry the same meaning.

While the strength profile might facilitate comparisons of a student's strengths and weaknesses across reporting categories in some cases, several factors merit caution. As noted earlier, many of the strand scores are very unreliable. The scaling underlying the strength profile does not mitigate this problem.

Additionally, the categories reflect more absolute comparisons. Relative comparisons are more difficult to make. As an example, if one scored High in both strand A and B, we know the student did very well in both strands compared to overall performance in the state (i.e., absolute status). However, we don't know whether the student's performance in strand A was better or worse relative to the performance in strand B (relative status).

Finally, some seemingly unusual results might occur that may be difficult for users to understand. As one example, it may be possible for a student to earn Medium in all reporting categories but have an Advanced-M performance level. This can happen because the strand scores are correlated, meaning the distributional properties of the total score depends not just on the variances of the strand scores, but the covariances among the strand scores as well. (An analogy would be when a school track team places first overall in a competition although they did not win a single event.)

Using PSSA-M Results for Other Purposes

Should PSSA-M results be used for placement decisions or for other special programs or services? Frequently asked questions about the PSSA-M pertain to the maximum possible PSSA-M scaled scores for various subjects, or what PSSA-M score represents the 90th percentile. The motivation behind many of these questions may be associated with special program eligibility.

Other uses or inferences based on PSSA-M results may or may not be valid as the validity evidence and arguments provided in Chapter Nineteen may not necessarily support other score uses and interpretations. According to the *Standards* (i.e., Standard 1.4) if a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary. Finally, a universal caveat for any test's result is that it not be used for placement and educational planning alone. Instead, other information about the student (e.g., other test performance data) should be included.

REPORTS

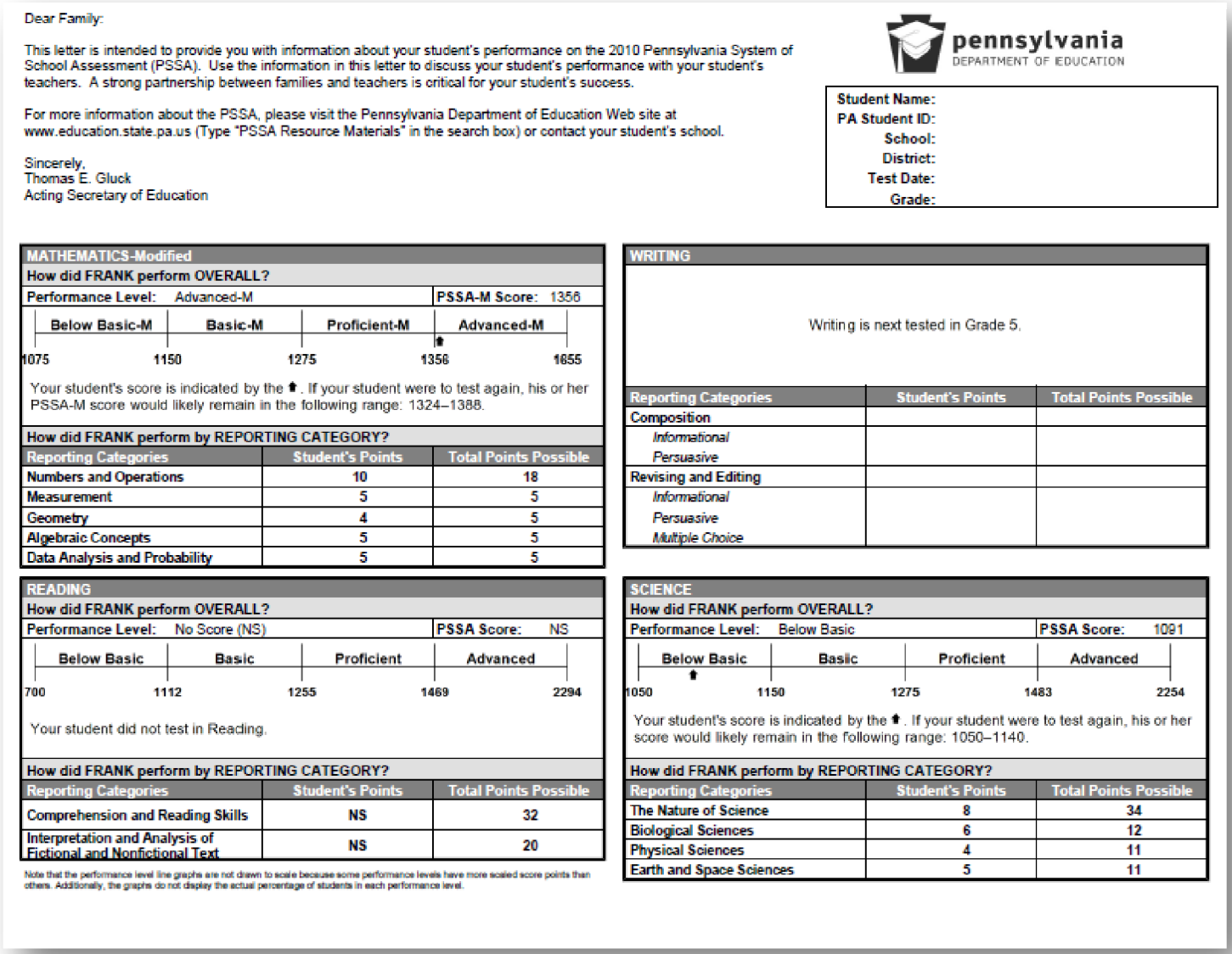
These following score reports are provided to students, parents, schools, and districts for the PSSA-M tests in mathematics:

- Parent Letter
- Individual Student Report
- School Summary Report
- District Summary Report
- Interpretive Guide

Parent Letter

Parent letters were delivered to Pennsylvania districts on August 2, 2010. This score report provides parents and students their first glimpse of performance on the spring 2010 PSSA-M tests. This report provides results at the student level. A sample of the report is provided in Figure 16–1.

Figure 16–1. Sample of Parent Letter



Dear Family:

This letter is intended to provide you with information about your student's performance on the 2010 Pennsylvania System of School Assessment (PSSA). Use the information in this letter to discuss your student's performance with your student's teachers. A strong partnership between families and teachers is critical for your student's success.

For more information about the PSSA, please visit the Pennsylvania Department of Education Web site at www.education.state.pa.us (Type "PSSA Resource Materials" in the search box) or contact your student's school.

Sincerely,
Thomas E. Gluck
Acting Secretary of Education

pennsylvania
DEPARTMENT OF EDUCATION

Student Name:
PA Student ID:
School:
District:
Test Date:
Grade:

MATHEMATICS-Modified

How did FRANK perform OVERALL?

Performance Level: Advanced-M **PSSA-M Score:** 1356

Below Basic-M | Basic-M | Proficient-M | Advanced-M

1075 | 1150 | 1275 | 1356 | 1655

Your student's score is indicated by the ⬆. If your student were to test again, his or her PSSA-M score would likely remain in the following range: 1324–1388.

How did FRANK perform by REPORTING CATEGORY?

Reporting Categories	Student's Points	Total Points Possible
Numbers and Operations	10	18
Measurement	5	5
Geometry	4	5
Algebraic Concepts	5	5
Data Analysis and Probability	5	5

WRITING

Writing is next tested in Grade 5.

Reporting Categories	Student's Points	Total Points Possible
Composition		
Informational		
Persuasive		
Revising and Editing		
Informational		
Persuasive		
Multiple Choice		

READING

How did FRANK perform OVERALL?

Performance Level: No Score (NS) **PSSA Score:** NS

Below Basic | Basic | Proficient | Advanced

700 | 1112 | 1255 | 1469 | 2294

Your student did not test in Reading.

How did FRANK perform by REPORTING CATEGORY?

Reporting Categories	Student's Points	Total Points Possible
Comprehension and Reading Skills	NS	32
Interpretation and Analysis of Fictional and Nonfictional Text	NS	20

Note that the performance level line graphs are not drawn to scale because some performance levels have more scaled score points than others. Additionally, the graphs do not display the actual percentage of students in each performance level.

SCIENCE

How did FRANK perform OVERALL?

Performance Level: Below Basic **PSSA Score:** 1091

Below Basic | Basic | Proficient | Advanced

1050 | 1150 | 1275 | 1483 | 2254

Your student's score is indicated by the ⬆. If your student were to test again, his or her score would likely remain in the following range: 1050–1140.

How did FRANK perform by REPORTING CATEGORY?

Reporting Categories	Student's Points	Total Points Possible
The Nature of Science	8	34
Biological Sciences	6	12
Physical Sciences	4	11
Earth and Space Sciences	5	11

Individual Student Report

A student report is provided for all students who took the PSSA-M. This report was delivered to Pennsylvania school districts on September 1, 2010. Districts are responsible for sending them home to the individual students. This report is a four-page color document that provides the types of scores explained earlier in this chapter. Screen shots of the four pages from a sample individual student report are provided in Figures 16–2 to 16–5.

Figure 16–2. Page 1 of the Individual Student Report

PENNSYLVANIA

Student Report

Dear Family:

This report is designed to provide you with specific information about your student's strengths and needs as measured by the 2010 Grade 4 Pennsylvania System of School Assessment (PSSA). I encourage you to use the information in this report to discuss with your student's teacher(s) ways to enhance your student's education. A strong partnership between families and teachers is critical for every child's success. Working together, we can help all children succeed in school.

For more information about the PSSA, please visit the Pennsylvania Department of Education Web site at www.education.state.pa.us (Type "PSSA Resource Materials" in the search box) or contact your student's school.

Sincerely,

Thomas E. Gluck
Acting Secretary of Education

Table of Contents

Page 1 ..... General Overview
Page 2 ..... Math, Reading, and Science Detailed Results
Page 3 ..... Writing Detailed Results
Page 4 ..... Helping Your Student Achieve Success

An Interpretation Guide for this report is available at www.education.state.pa.us (Type "student report guide" in the search box).

Student Name:
PA Student ID:
School:
District:
Test Date: Spring 2010
Grade: 4

Student's PSSA Results by Subject

Subject	Below Basic	Basic	Goal Range: Proficient	Goal Range: Advanced
Mathematics*			✓	
Reading	✓			
Science			✓	
Writing				

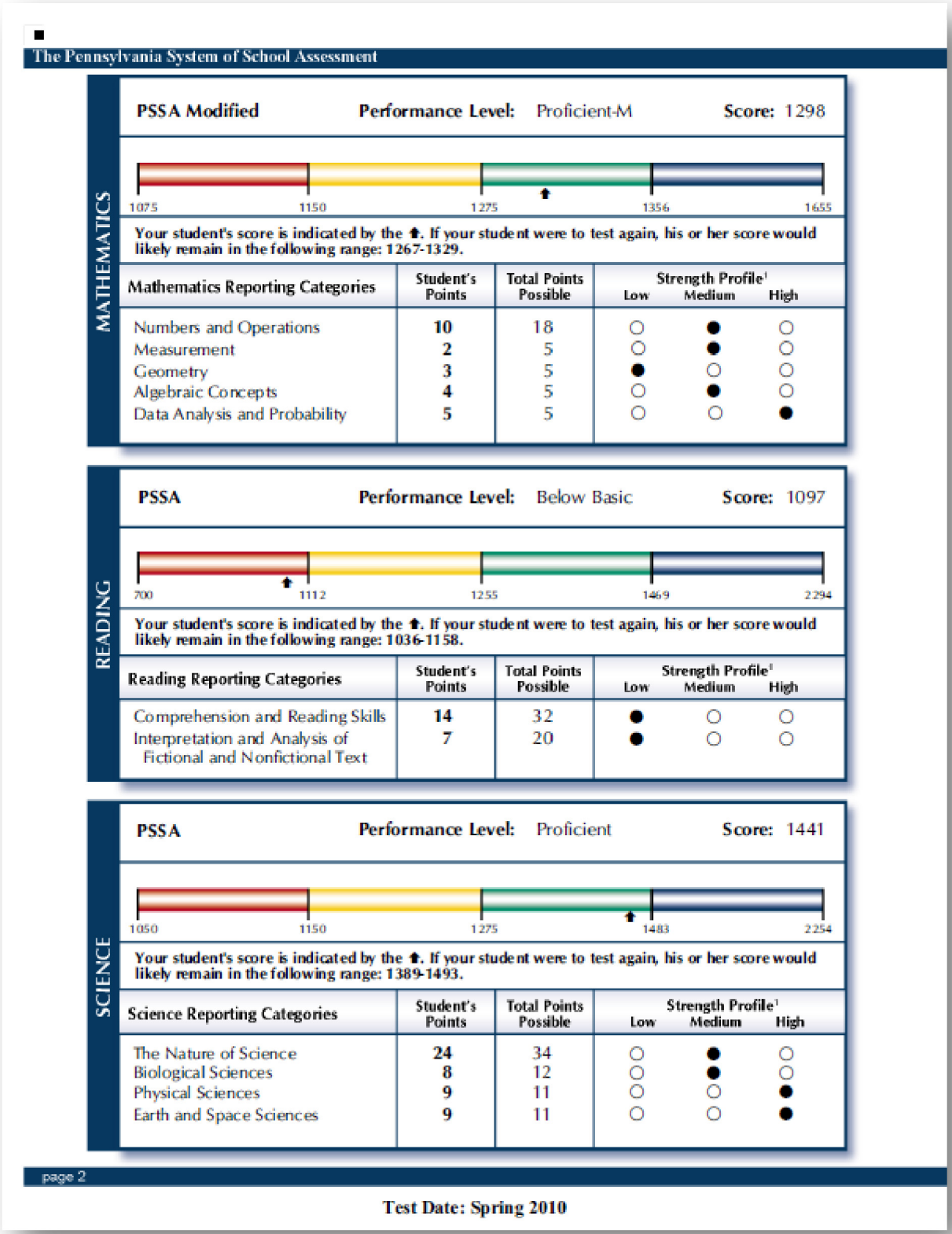
Writing is not assessed in Grade 4.
*Student participated in the PSSA Modified Assessment (PSSA-M). Contact your school for more information concerning the PSSA-M.

pennsylvania
DEPARTMENT OF EDUCATION

The Pennsylvania System of School Assessment page 1

www.education.state.pa.us

Figure 16–3. Page 2 of the Individual Student Report



The Pennsylvania System of School Assessment

MATHEMATICS

PSSA Modified Performance Level: Proficient-M Score: 1298

1075 1150 1275 1356 1655

Your student's score is indicated by the ⬆. If your student were to test again, his or her score would likely remain in the following range: 1267-1329.

Mathematics Reporting Categories	Student's Points	Total Points Possible	Strength Profile[1] Low	Medium	High
Numbers and Operations	10	18	○	●	○
Measurement	2	5	○	●	○
Geometry	3	5	●	○	○
Algebraic Concepts	4	5	○	●	○
Data Analysis and Probability	5	5	○	○	●

READING

PSSA Performance Level: Below Basic Score: 1097

700 1112 1255 1469 2294

Your student's score is indicated by the ⬆. If your student were to test again, his or her score would likely remain in the following range: 1036-1158.

Reading Reporting Categories	Student's Points	Total Points Possible	Strength Profile[1] Low	Medium	High
Comprehension and Reading Skills	14	32	●	○	○
Interpretation and Analysis of Fictional and Nonfictional Text	7	20	●	○	○

SCIENCE

PSSA Performance Level: Proficient Score: 1441

1050 1150 1275 1483 2254

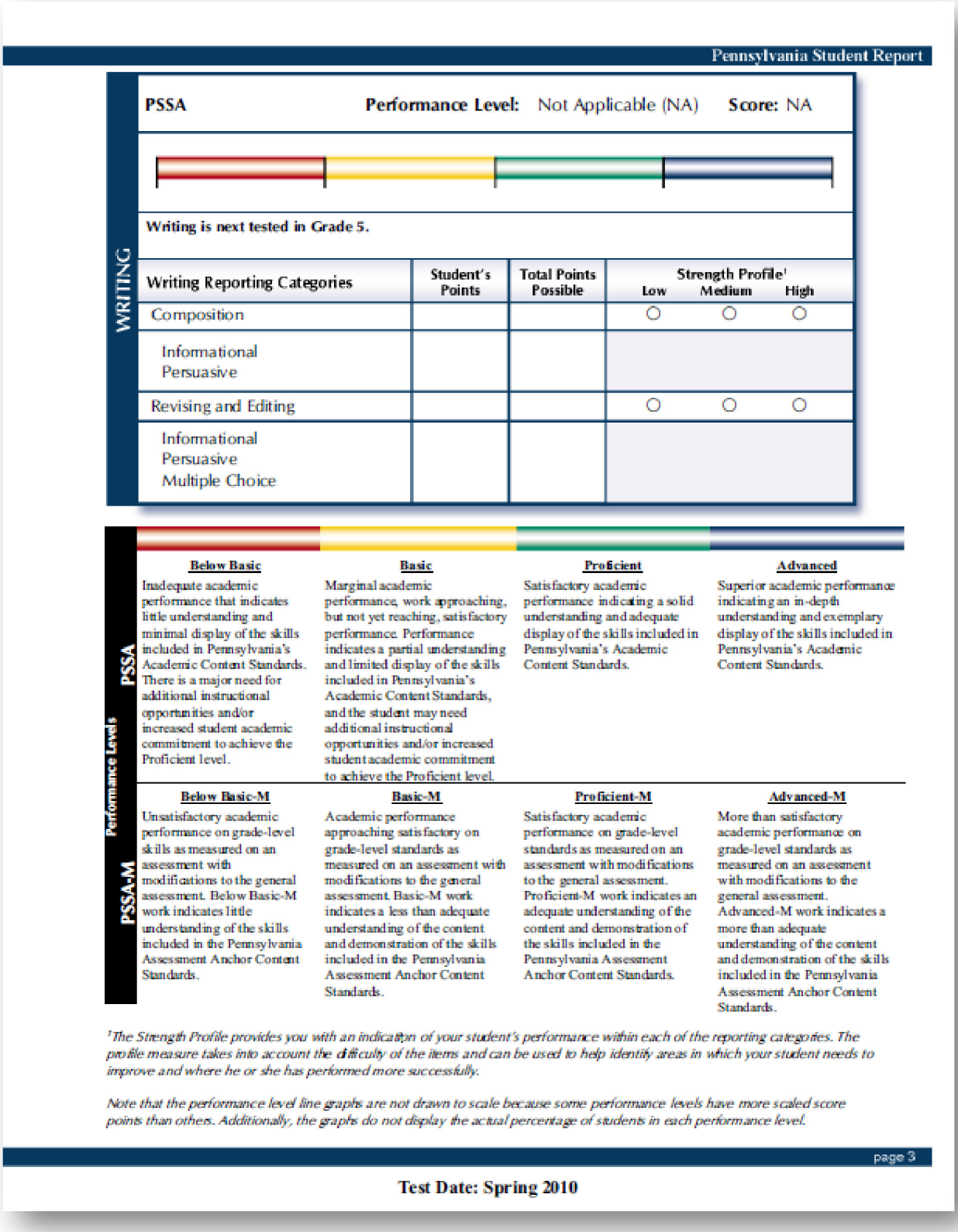
Your student's score is indicated by the ⬆. If your student were to test again, his or her score would likely remain in the following range: 1389-1493.

Science Reporting Categories	Student's Points	Total Points Possible	Strength Profile[1] Low	Medium	High
The Nature of Science	24	34	○	●	○
Biological Sciences	8	12	○	●	○
Physical Sciences	9	11	○	○	●
Earth and Space Sciences	9	11	○	○	●

page 2

Test Date: Spring 2010

Figure 16–4. Page 3 of the Individual Student Report



Pennsylvania Student Report

WRITING

PSSA **Performance Level:** Not Applicable (NA) **Score:** NA

Writing is next tested in Grade 5.

Writing Reporting Categories	Student's Points	Total Points Possible	Strength Profile[1] Low	Medium	High
Composition			○	○	○
Informational Persuasive					
Revising and Editing			○	○	○
Informational Persuasive Multiple Choice					

Performance Levels

PSSA

Below Basic
Inadequate academic performance that indicates little understanding and minimal display of the skills included in Pennsylvania's Academic Content Standards. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient level.

Basic
Marginal academic performance, work approaching, but not yet reaching, satisfactory performance. Performance indicates a partial understanding and limited display of the skills included in Pennsylvania's Academic Content Standards, and the student may need additional instructional opportunities and/or increased student academic commitment to achieve the Proficient level.

Proficient
Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in Pennsylvania's Academic Content Standards.

Advanced
Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in Pennsylvania's Academic Content Standards.

PSSA-M

Below Basic-M
Unsatisfactory academic performance on grade-level skills as measured on an assessment with modifications to the general assessment. Below Basic-M work indicates little understanding of the skills included in the Pennsylvania Assessment Anchor Content Standards.

Basic-M
Academic performance approaching satisfactory on grade-level standards as measured on an assessment with modifications to the general assessment. Basic-M work indicates a less than adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.

Proficient-M
Satisfactory academic performance on grade-level standards as measured on an assessment with modifications to the general assessment. Proficient-M work indicates an adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.

Advanced-M
More than satisfactory academic performance on grade-level standards as measured on an assessment with modifications to the general assessment. Advanced-M work indicates a more than adequate understanding of the content and demonstration of the skills included in the Pennsylvania Assessment Anchor Content Standards.

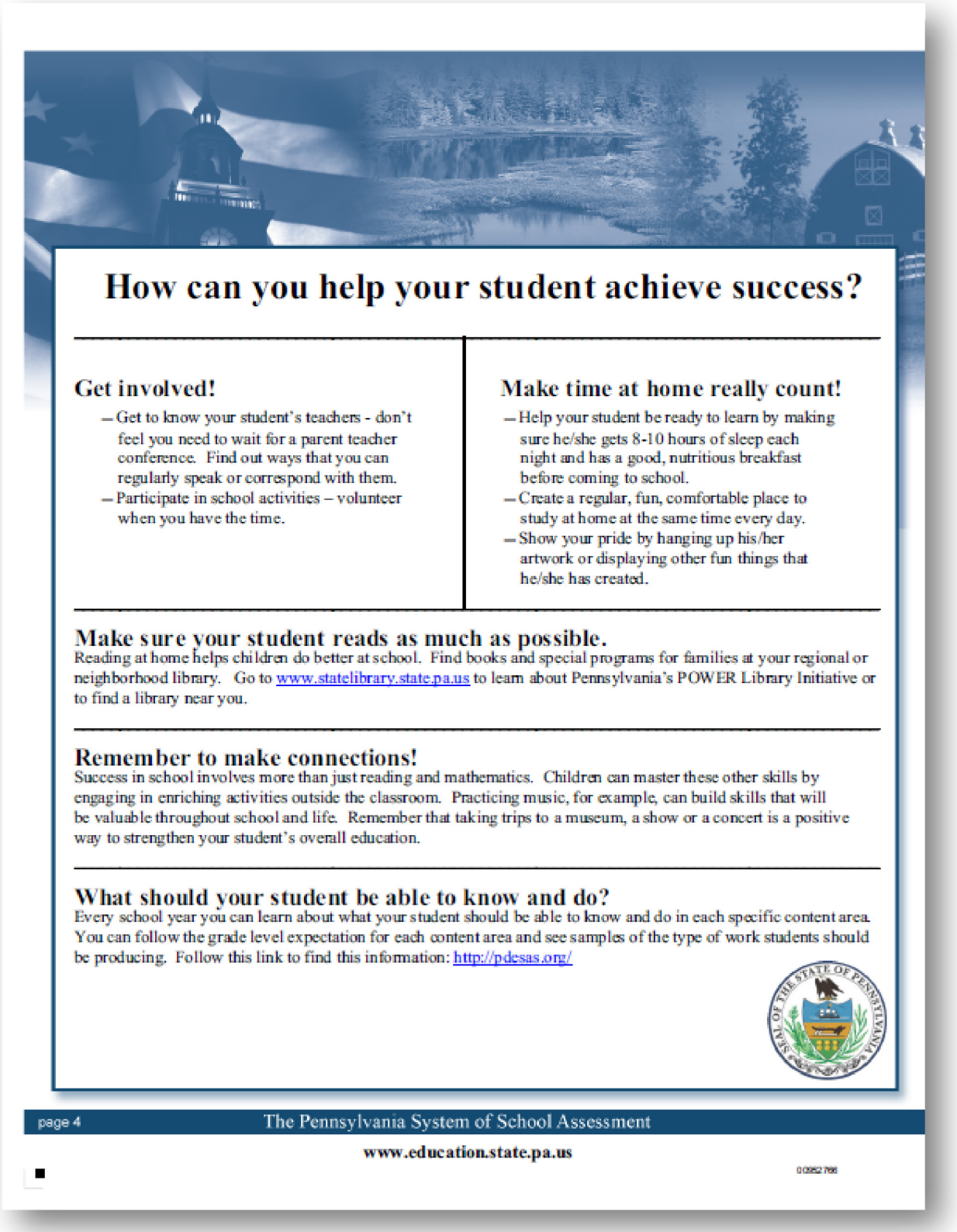
[1]The Strength Profile provides you with an indication of your student's performance within each of the reporting categories. The profile measure takes into account the difficulty of the items and can be used to help identify areas in which your student needs to improve and where he or she has performed more successfully.

Note that the performance level line graphs are not drawn to scale because some performance levels have more scaled score points than others. Additionally, the graphs do not display the actual percentage of students in each performance level.

page 3

Test Date: Spring 2010

Figure 16–5. Page 4 of the Individual Student Report



How can you help your student achieve success?

Get involved!

- Get to know your student's teachers - don't feel you need to wait for a parent teacher conference. Find out ways that you can regularly speak or correspond with them.
- Participate in school activities – volunteer when you have the time.

Make time at home really count!

- Help your student be ready to learn by making sure he/she gets 8-10 hours of sleep each night and has a good, nutritious breakfast before coming to school.
- Create a regular, fun, comfortable place to study at home at the same time every day.
- Show your pride by hanging up his/her artwork or displaying other fun things that he/she has created.

Make sure your student reads as much as possible.

Reading at home helps children do better at school. Find books and special programs for families at your regional or neighborhood library. Go to www.statelibrary.state.pa.us to learn about Pennsylvania's POWER Library Initiative or to find a library near you.

Remember to make connections!

Success in school involves more than just reading and mathematics. Children can master these other skills by engaging in enriching activities outside the classroom. Practicing music, for example, can build skills that will be valuable throughout school and life. Remember that taking trips to a museum, a show or a concert is a positive way to strengthen your student's overall education.

What should your student be able to know and do?

Every school year you can learn about what your student should be able to know and do in each specific content area. You can follow the grade level expectation for each content area and see samples of the type of work students should be producing. Follow this link to find this information: http://pdesas.org/

page 4 The Pennsylvania System of School Assessment

www.education.state.pa.us

School and District Summary Reports

Summary reports are provided at the school and district level. These reports contain summary information about the percentage of students in each of the four performance levels. Raw scores are also provided by assessment anchor to allow schools or districts to identify content strands of strength or weakness.

Interpretative Guide

An interpretative guide is provided to help parents and other PSSA-M stakeholders better understand test result information presented in the individual student report. The interpretative guide can be found on the PDE website.

Chapter Seventeen: Operational Test Statistics

This chapter presents various summary statistics for the PSSA-M total test scores based on the final data file described in Chapter Nine. Related information covered elsewhere in this report includes the item-level statistics that were presented in Chapters 11 (classical item statistics) and 12 (Rasch item statistics). Refer to those chapters for additional consideration as item difficulty distributions can affect total score distributions.

PERFORMANCE LEVEL STATISTICS

Table 17–1 presents performance level percentages by grade and content[10]. Starting in the 2011 PSSA-M Technical Report, an appendix tracking the historical performance levels will be provided.

Table 17–1. Performance Level Percentages for the 2010 PSSA-M

		Percentage in Each Performance Level			
Grade	**Content**	**Below Basic-M**	**Basic-M**	**Proficient-M**	**Advanced-M**
4	Math	4.8	35.8	38.3	21.2
5	Math	5.5	43.4	38.0	13.0
6	Math	7.3	44.6	38.9	9.2
7	Math	8.3	50.4	33.2	8.1
8	Math	10.0	49.2	35.2	5.6
11	Math	21.5	45.3	27.6	5.6

SCALED SCORES

Summary Statistics

Table 17–2 provides the scaled score means and standard deviations[10].

Table 17–2. Means and Standard Deviations for the 2010 PSSA-M Scaled Scores

	Mathematics	
Grade	**Mean**	**SD**
4	1286.2	83.8
5	1275.5	84.7
6	1264.6	83.5
7	1252.0	79.2
8	1250.3	83.5
11	1228.4	99.6

[10] These are not the final official results as the appeals process was still ongoing at the time this report was written. Official results will be posted on PDE's website. See section Every Test has a Unique Scale in Chapter Sixteen for some caveats regarding interpretation of scale scores.

Scaled Score Distributions

Scaled scores are based on a linear transformation of the Rasch ability estimates. Distributions of the Rasch abilities were provided at the end of Chapter Twelve.

RAW SCORES

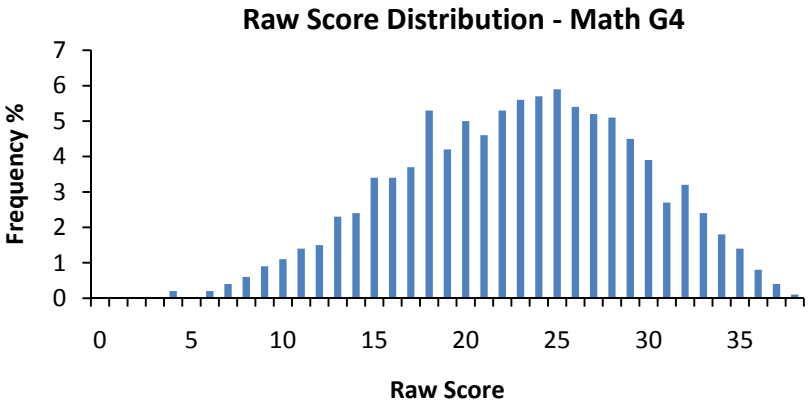
Summary Statistics

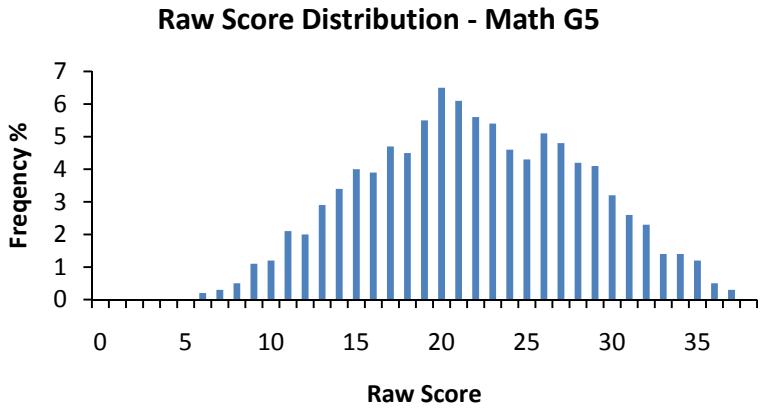
The reader is referred to Appendix H for summary statistics for the operational raw scores. The statistics reported include the: number of points possible (Pts.), number of items (Len.), number of students tested (N), mean number of score points received (Mean), standard deviation of test scores (SD), reliability (r), traditional standard error of measurement (SEM), and item types (Items) used to determine each score. These statistics are based on the total test using both multiple-choice (MC) and open-ended (OE) items for the operational sections of each form. (For those interested in information disaggregated by item type, Chapter Eleven provides breakout statistics for MC and OE items.)

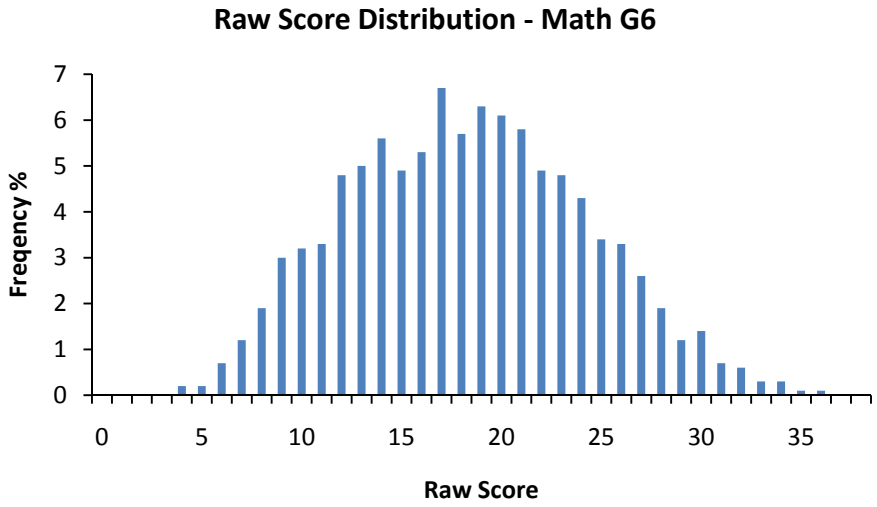
Score Distributions

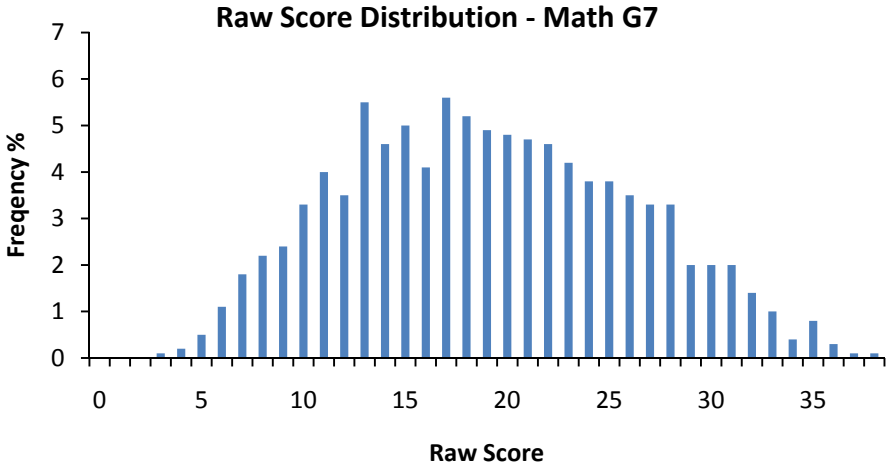
Raw score relative-frequency (rf) distributions are provide in Figure 17–1. Most distributions are unimodal and slightly positively skewed (with the exception of Grade 4, which was negatively skewed).

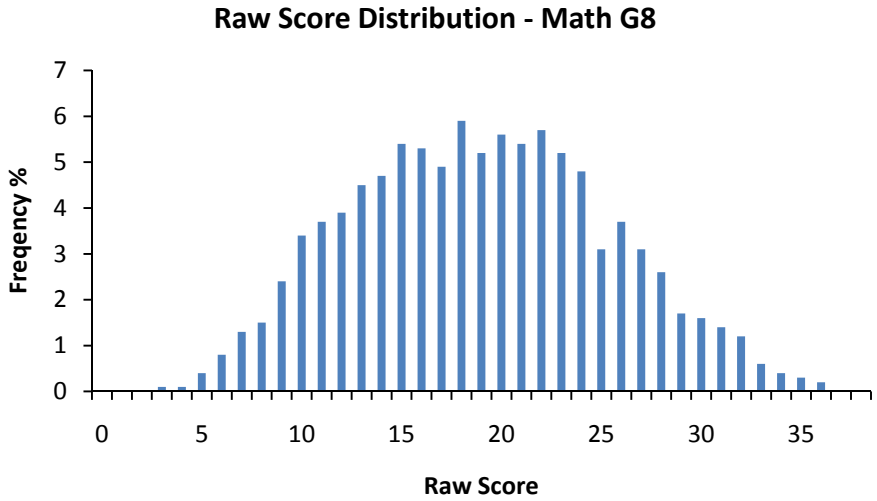
Figure 17–1. Raw Score Distributions

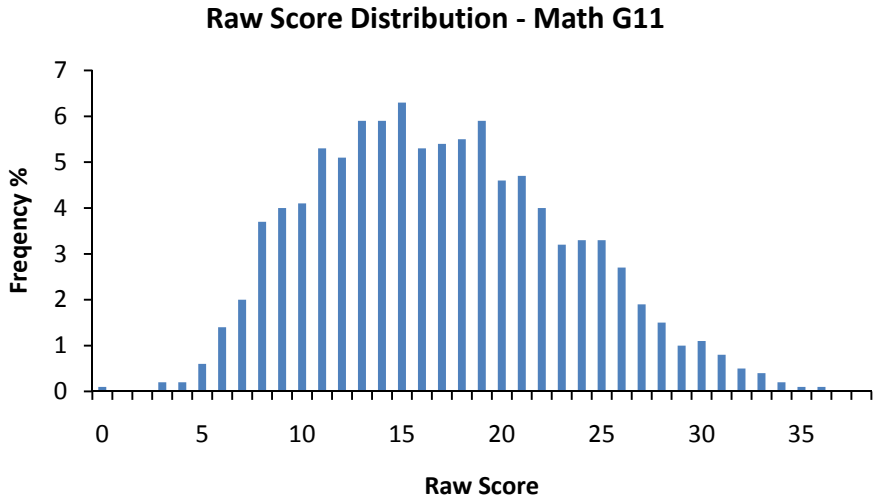









Raw Score Distribution - Math G8
Freqency %
0
1
2
3
4
5
6
7
0
5
10
15
20
25
30
35
Raw Score


Raw Score Distribution - Math G11
Freqency %
0
1
2
3
4
5
6
7
0
5
10
15
20
25
30
35
Raw Score

Chapter Eighteen: Reliability

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), reliability refers to:

> the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group (p. 25).

Frisbie (2005) highlighted several elements of this definition. First, reliability is a property of the test scores, not of the test itself. Many may appreciate this distinction, but in casual usage, individuals frequently make reference to a reliable test. While reliability concerns test scores (and not the test specifically), it's important to appreciate the fact that test scores can be affected by characteristics of the instrument. For example, all other things being equal, tests with more items/points tend to be more reliable than tests with fewer items/points. Second, reliability coefficients are group specific. Reliabilities tend to be higher in populations that are more heterogeneous and lower in populations that are more homogeneous. Consequently, both test length and population heterogeneity should be considered when evaluating reliability.

There are other reliability considerations that may be less evident from the *Standard's* definition, yet are still important for test users to understand. While freedom from measurement error is highlighted in the definition above, reliability is specifically concerned with random sources of error. Indeed, the degree of inconsistency due to random error sources is what determines reliability: less consistency is associated with lower reliability and more consistency is associated with higher reliability. Of course, systematic error sources also exist. These can artificially increase reliability and decrease validity. (Validity is further discussed in Chapter Nineteen.)

Another noteworthy issue is that multiple sources of error exist (e.g., the day of testing, the items used, the raters who score the items). However, most widely-used reliability indices only reflect a single type of error. Consequently, it is important for test users to understand what specific type of error is being considered in a reliability study; and equally, if not more important, what types are not.

Understanding the distinction between relative error and absolute error is also important as many reliability indices only reflect relative error. Relative error is of interest whenever the relative ordering of individuals respective to their test performance is of interest. Understanding examinee rank-order stability is important; however, such stability might be well achieved even when the specific score values are considerably different. When specific score values are considered important (e.g., if cut scores are used), then one should be interested in absolute error too. Generally, there is more error variance when considering the absolute scores of examinees, which, in turn, suggests lower reliability.

As the above suggests, reliability is a complex, non-unitary notion that cannot be adequately represented by a single number. There are several reliability indices available and these may not provide the same results (Frisbie, 2005). The remainder of this chapter covers the following:

- Reliability coefficients and their interpretation
- Unconditional and conditional standard errors of measurement (SEMs and CSEMs)
- Decision consistency
- Rater agreement

RELIABILITY INDICES

As shown below, the reliability coefficient expresses the consistency of test scores as the ratio of true score variance to total score variance. The total variance contains two components: 1) the variance in true scores, and 2) the variance due to the imperfections in the measurement process. Put differently, total variance equals true score variance plus error variance[11].

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Reliability coefficients indicate the degree to which differences in test scores reflect true differences in the attribute being tested rather than random fluctuations. Total test score variance (i.e., individual differences) is partly due to real differences in the attribute (i.e., true variance) and partly due to random error in the measurement process (i.e., error variance).

Reliability coefficients range from 0.0 to 1.0. If all test score variance were true, the index would equal 1.0. The index will be 0.0 if none of the test score variance were true. Such scores would be pure random noise (i.e., all measurement error). If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in practice, it is clear that larger coefficients are more desirable as that indicates that test scores are less influenced by random error. (How big is big enough and how small is too small are issues considered in a later section.)

As noted in the introduction, there are several different indices that can be used to estimate this ratio. One approach is referred to as internal consistency, which is derived from analyzing the performance consistency of individuals over the items within a test. As discussed below, these internal consistency indices do not take into account other sources of error—for example, variations due to random errors associated with the linking process; day-to-day variations (student health, testing environment, etc.); or rater inconsistency.

COEFFICIENT ALPHA

Although a number of reliability indices exist, perhaps the most frequently reported for achievement tests is Coefficient Alpha. Consequently, this index is the one reported for the PSSA-M. Alpha indicates the internal consistency over the responses to a set of items measuring an underlying trait, in this case, academic achievement in subject areas such as mathematics.

[11] A covariance term is not required as true scores and error are assumed to be uncorrelated in classical test theory.

Alpha is an internal consistency index. It can be conceptualized as the extent to which an exchangeable set of items from the same domain would result in a similar rank ordering of students. Note that relative error is reflected in this index. Variation in student performance from one sample of items to the next should be of particular concern for any achievement test user. Consider two hypothetical vocabulary tests intended for the same group of students. Each test contains different sets of unique words that are believed to be randomly equivalent, perhaps like the ones shown below:

Table 18–1. Two Hypothetical Vocabulary Tests

Test One	Test Two
Abase	Abate
Boon	Bilk
Capricious	Circuitous
Deface	Debase
....	....
Zealous	Zenith

If a representative group of students could take both of these tests, and the correlation between the scores obtained, then that result would represent the parallel forms reliability of the test scores. However, such data-collection designs are impractical in large-scale settings and experimental confounds like fatigue and practice effects are likely to affect the results. Internal-consistency reliability indices arose in part to provide reliability measures using the data from just a single test administration. So, if students only took Test One and the Coefficient Alpha index for those test scores were high, then this would suggest that Test Two would provide a very similar rank ordering of the students if they had taken it instead. If Coefficient Alpha were low, dissimilar rank orderings would likely be observed—again, relative-error variance is reflected in Alpha. (It should also be noted that Coefficient Alpha is algebraically identical to a $p \times I$ design under Generalizability Theory when relative error variance is assumed.)

Formula

Consider the following data matrix representing the scores of persons (rows) on items (columns):

Table 18–2. Person × Item Score (X_{pi}) Infinite (Population-Universe) Matrix

	Item			
Person	1	2	. . . *I*	. . . *k*
1	*Y*11	*Y*12	. . . *Y*1*i*	. . . *X*1*k*
2	*Y*21	*Y*22	. . . *Y*2*i*	. . . *X*2*k*
. . . . .				
. . . . .				
P	*Yp*1	*Yp*2	. . . *Ypi*	. . . *Xpk*
. . . . .				
. . . . .				
N	*YN*1	*YN*2	. . . *YNi*	. . . *XNk*

Note. Adapted from Cronbach and Shavelson (2004).

Then, a general computational formula for Alpha is as follows:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right),$$

where N is the number of parts (items or testlets), σ_X^2 is the variance of the observed total test scores, and $\sigma_{Y_i}^2$ is the variance of part i.

FURTHER INTERPRETATIONS

Rules of Thumb

What reliability value is considered high enough? What values are considered too low? Although frequently asked for, any rules of thumb for interpreting the magnitude of reliability indices are mostly arbitrary. Another approach is to research the reliabilities from similar testing instruments to see what values are commonly observed. For the PSSA-M, comparisons to tests of similar lengths that were administered to similar student populations from other large-scale assessment programs would be relevant. For many other state assessments programs, reliabilities in the low 0.90s are usually the highest ever observed and reliabilities in the high 0.80s are very common.

The lower a given reliability coefficient, the greater the potential for over-interpretation of the associated results. As suggested above, there is no firm guideline regarding how low is too low. However, as an informative point of reference, a reliability coefficient of 0.50 would suggest that there is as much error variance as true-score variance in the scores.

Is Alpha a Lower Limit to Reliability?

According to Brennan (1998), "the conventional wisdom that Coefficient Alpha is a lower limit to reliability is based largely on a misunderstanding." In reflecting on the 50th anniversary of his seminal 1951 article, Cronbach—in Cronbach and Shavelson (2004)—expressed similar misgivings about this conventional wisdom:

> one could argue that alpha was almost an unbiased estimate of the desired reliability….the almost in the preceding sentence refers to a small mathematical detail that causes the alpha coefficient to run a trifle lower than the desired value. This detail is of no consequence and does not support the statement made frequently in textbooks or in articles that alpha is a lower value to the reliability coefficient. That statement is justified by reasoning that starts with the definition of the desired coefficient as the expected consistency among measurements that had a higher degree of parallelism than the random parallel concept implied.

The assumptions for three common parallelism models are presented in Table 18–3. Alpha's assumptions come from the Essentially-Tau Equivalent model, which does not require equal means or equal variances across test parts. Based on this, Brennan (1998) asserts that the lower-limit issue, as conceptualized by many, provides an answer to a question that is of minimal importance. Reframed differently, the goal of selecting a reliability coefficient is not to find the one that provides the highest coefficient, but the one that most accurately reflects the test data under study.

It is important to note that there are factors encountered in practice that may legitimately make Coefficient Alpha an underestimate of reliability. However, there are also factors that might make Coefficient Alpha an overestimate of reliability. Both possibilities are discussed further below and generally arise when the Essentially-Tau Equivalent assumptions are strained.

Table 18–3. Summary of Expectations/Observable Relationships for Different Parallelism Models

Relationship	Degree of Measurement Parallelism*		
	Classically Parallel	Essentially-Tau Equivalent	Congeneric
Content Similarity	Yes	Yes	Yes
Equal Means across Parts	Yes	No	No
Equal Variances across Parts	Yes	No	No
Equal Covariances across Parts	Yes	Yes	No
Equal Covariances with other Variables	Yes	Yes	No

*Other models exist, but are not considered here due to their limited application in practice.

Biases that Might make Alpha an Underestimate of Reliability

There are factors that might negatively bias Coefficent Alpha, making the apparent reliability lower than it may actually be. Two situations frequently encountered in practice that might cause this include: 1) tests that are comprised of mixed item types (e.g., MC and OE items); and 2) tests that include a planned stratification of the test items according to topics or subdomains.

Although both situations strictly violate the assumptions on which Coefficient Alpha was derived (i.e., the tests are not based on equal part lengths in the former case and are not randomly parallel in the latter case), neither necessarily guarantees that the reliability will be markedly lower. In the latter case, reliability will be underestimated only when strand items are homogeneous enough for the average covariance within strata to exceed the average covariance between strata. Although both are potential influences for the PSSA-M s, most of the total test score reliabilities reported in Appendix J are all close to or above 0.80, indicating fairly consistent test scores for these instruments.

Biases that Might make Alpha an Overestimate of Reliability

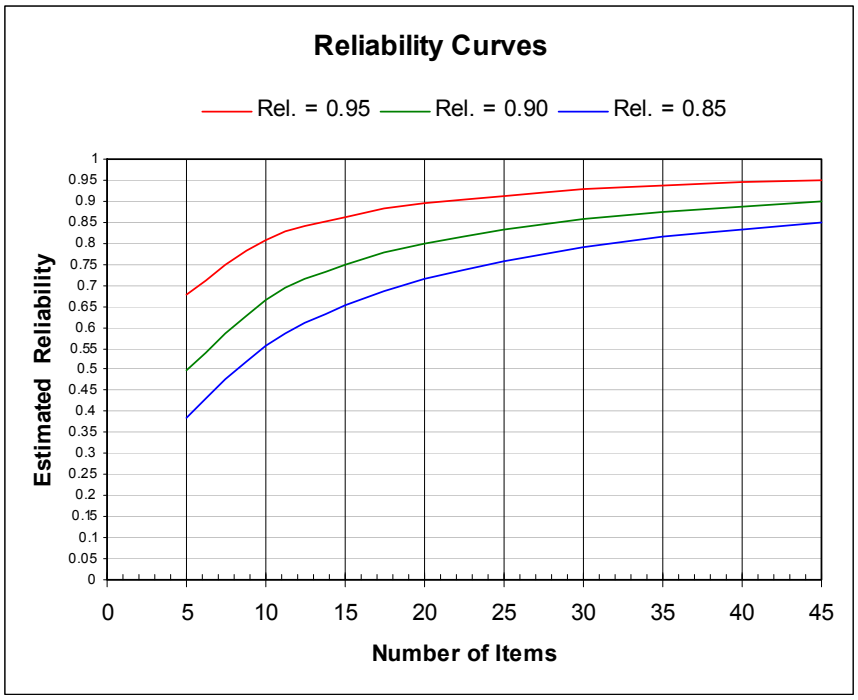
As emphasized in earlier sections, Coefficient Alpha only takes into account measurement error that arises from the selection of items used on a particular test form. There are other sources of random inaccuracy. One is due to the occasion of testing. Other various random conditions that might affect students on any particular testing occasions are: illness, fatigue, anxiety, etc. Also, when a test includes OE items, as the PSSA-M does, another source that can cause random fluctuation is from the OE item scorers. In a sense, Alpha may be positively biased because it does not take into account these other important sources of random error. Really, any internal consistency reliability index might understate the overall problem of measurement error because such sources of random error are ignored by them.

Another positive bias can occur when items are associated (clustered) with a common stimulus. Item bundles and testlets are other frequently used terms for this situation. One concrete example is when multiple reading comprehension items are associated with a common passage selection. Again, such a situation does not guarantee that the reliability estimate will be markedly affected, but the potential exists.

Strand Scores

As noted in the introduction, reliabilities tend to go up in value with an increase in test length and go down in value with a decrease in test length. Figure 18–1 illustrates this relationship for a hypothetical 45-point test with three total score reliabilities: 0.95, 0.90, and 0.85. As an example, the curve for reliability equal to 0.90 suggests that a 10-item strand would be expected to have a score reliability of just over 0.65. The use of the Spearman-Brown prophecy formula assumes all items are exchangeable, which in practice, they may not be. While such a chart may not perfectly model actual strand correlations, the intent is only to illustrate the substantial impact that limited numbers of strand items can have on strand-score reliability. It is not surprising that strand scores with more points tend to show higher reliability coefficients and those with fewer points tend to show lower reliability coefficients. Further, what is most important for PSSA-M users to note is that some strand score reliabilities may be too low to warrant interpretation at the individual student level.

Figure 18–1. Example of the Relationship between Test Length and Reliability



Note. Tabled values derived using the Spearman-Brown formula.

Individual-Level versus Group-Level Scores

The results presented in this chapter pertain to the reliability of individual scores. Group results (e.g., state and district levels) are also provided on PSSA-M score reports, but the reliability of those scores are not specifically calculated here. However, as a general rule it should be noted that the reliabilities of group mean scores are almost always higher (sometimes substantially) than the corresponding reliabilities for individual scores. This is especially important to remember for strand scores because those scores can be quite reliable at the group level, even though their individual reliabilities may be too low. Because the reliability of group mean scores (e.g., school or district means) tends to be higher than that of individual scores, the interpretation of strand scores at these aggregate levels is likely very reasonable in most instances. Even though the reliability for means scores based on only a few items might be adequate, the validity of those same scores might be suspect because use of only few items may not adequately cover the construct of interest. Validity is further discussed in Chapter Nineteen.

STANDARD ERROR OF MEASUREMENT (SEM)

The reliability coefficient is a unit-free indicator that reflects the degree to which scores are free of measurement error. It always ranges between 0.0 and 1.0 regardless of the test's scale. Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent in a group. However, they are not that useful for helping users interpret test scores. The standard error of measurement (SEM) is another indicator of test score precision that is better suited for determining the effect of measurement inconsistencies for the scores obtained by individual examinees. This is particularly so for Conditional SEMs (CSEM) discussed further below.

Traditional Standard Error of Measurement

A precise, theoretical interpretation of the SEM is somewhat unwieldy. A beginning point for understanding the concept is as follows. If everyone being tested had the same true score[12], there would still be some variation in observed scores due to imperfections in the measurement process, such as random differences in attention during instruction or concentration during testing, or the sampling of test items. The standard error is defined as the standard deviation[13] of the distribution of observed scores for students with identical true scores. Because the SEM is an index of the random variability in test scores in actual score units, it represents very important information for test score users.

The SEM formula is provided below:

$$SEM = SD\sqrt{1 - reliability}$$

and indicates that the value of the SEM depends on both the reliability coefficient and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value) the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value) the SEM would be 0.0. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). Additionally, the value of the SEM takes the group variation (i.e., score standard deviation) into account. Consider that an SEM of 3 on a 10-point test would be very different from an SEM of 3 on a 100-point test.

[12] True score is the score the person would receive if the measurement process were perfect.

[13] The standard deviation of a distribution is a measure of the dispersion of the observations. For the normal distribution about 16 percent of the observations are more than one standard deviation above the mean.

Traditional SEM Confidence Intervals

The SEM is an index of the random variability in test scores in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the precision of individual tests scores. SEMs help place 'reasonable limits' (Gulliksen, 1950) around observed scores through construction of an approximate score band. Often referred to as confidence intervals, these bands are constructed by taking the observed scores, *X*, and adding and subtracting a multiplicative factor of the SEM. As an example, students with a given true score will have observed scores that fall between +/- 1 SEM about two thirds of the time[14]. For +/- 2 SEM confidence intervals, this increases to about 95 percent.

Further Interpretations

ONE SEM FOR ALL TEST SCORES

The SEM approach described above only provides a single numerical estimate for constructing the confidence intervals for examinees regardless of their score level. In reality however, such confidence intervals vary according to one's score. Consequently, care should be taken using the SEM for students with extreme scores. (An alternate approach that conditions the SEM on one's score estimate is described in the next sections.)

GROUP SPECIFIC

As noted in the introduction, reliabilities are group specific. The same is true for SEMs because both score reliabilities and score standard deviations vary across groups.

RAW SCORE METRIC

The SEM approach is calculated using raw scores, and as such, the resulting confidence interval bands are on the raw score metric. Error bands on the scaled score metric are considered in the next section.

TYPE OF ERROR REFLECTED

The interpretation of the SEM should be driven by the type of score reliability that underpins it; so, the PSSA-M SEMs involve the same source of error relevant to internal consistency indices. As noted earlier, a precise technical explanation of the SEM (and resulting confidence intervals) can be unwieldy. Because of this, score users are often provided less complex interpretations.

One simpler description sometimes used is that a confidence interval represents the possible score range that one would observe if a student could be tested twice with the same instrument. Taking the same test on a different day implies the only source of random error being considered is related to the occasion of testing, such as a student might be sleepier one day than another, or may be sick, or did not get a good breakfast. There is a reliability index that captures this source of random error and it is referred to as the test-retest reliability coefficient. This is not the type of reliability computed for the PSSA-M. When internal consistency reliability estimates are used, such an explanation blurs the fact that random error based on the occasion of testing is not considered.

[14] Some prefer the following interpretation: if a student were tested an infinite number of times, the +/- 1 SEM confidence intervals constructed for each score would capture the student's true score 68 percent of the time.

When SEMs are derived from internal consistency reliability estimates, a better approach is to describe the confidence interval as providing reasonable bounds for the range of scores that a student might receive if he or she took an equivalent version of the test. (That is, the student took a test that covered exactly the same content, but included a different set of items.) As an example, if the PSSA-M score was 1750 and the SEM band was 1700 to 1800, then a student would be likely to receive a score somewhere between 1700 and 1800 if a different version of the test had been taken. (cf. "If an infinite number of tests with equivalent content were taken, the student's true score will lie within the constructed confidence intervals 68 percent of the time" the prior version may be more adequate for lay persons.)

Results and Observations

Coefficient Alpha results and associated (traditional) SEMs for various PSSA-M scores are documented in Appendix J. Values were derived using the PSSA-M final data file (see Chapter Nine). The results are organized by subject area and grade. Each table also breaks out the various reporting strands and groups of interest (i.e., the total student population); gender and ethnic groups; English language learners (ELL), students with individualized education plan (IEP), and the economically disadvantaged (ED). The statistics reported include the: number of points possible (Pts.), number of items (Len.), number of students tested (N), mean number of score points received (Mean), standard deviation of test scores (SD), reliability (r), traditional standard error of measurement (SEM), and item types (Items) used to determine each score.

Note that these tables report the standard deviations of observed scores. Assuming normally-distributed scores, one would expect about two-thirds of the observations to be within one standard deviation of the mean. An estimate of the standard deviation of the true scores can be computed as: $\hat{\sigma}_T = \sqrt{\hat{\sigma}_x^2 - \hat{\sigma}_x^2(1-\hat{\rho}_{xx})}$.

The overall test score reliability values are at what many would consider to be the lower end of the adequate range for making decisions about individual students (with many in the low 0.80s) for mathematics. Earlier it was noted that reliabilities tend to go up in value with an increase in test length[15] and population heterogeneity and go down in value with a decrease in test length and more homogeneous populations. Across the grades and subjects tabled in Appendix J, reliabilities for the sub-strands tended to follow these same trends; that is, strands with more items tended to show higher reliability coefficients. Also, groups exhibiting more variability in test scores tended to have higher reliability coefficients. Perhaps the most significant result pertains to an earlier caution (i.e., that some strand score reliabilities are too low to warrant interpretation at the individual student level[16]). Once again, there is no firm guideline regarding how low is too low. The lower a given reliability coefficient, the greater the potential for over-interpretation. As a point of reference, a reliability coefficient of 0.50 would suggest that there is as much error variance as true-score variance in the scores. It should be noted that the reliability of group mean scores (e.g., school or district means) tends to be higher than that of

[15] Using the Spearman-Brown formula, if the PSSA-M mathematics test was the same length as the general PSSA mathematics test, the projected reliability would be in the high 0.80s. Coefficient Alpha estimates from the PSSA mathematics test are generally in the low 0.90s. The reduced test length largely accounts for the difference. Homogeneity in the testing population may be responsible for the remainder.

[16] In fact, a few reliability values in the appendix are negative. Theoretically, reliability values should be non-negative. However, the computational formula for alpha can yield negative results on rare occasions (when sample sizes are small). This likely indicates that the true score variance is in reality extremely small and sampling error resulted in the negative alpha estimate.

individual scores, suggesting that interpretation of strand scores at these aggregate levels might be reasonable in some cases.

RASCH CONDITIONAL STANDARD ERRORS OF MEASUREMENT

The CSEM also indicates the degree of measurement error but does so in scaled-score units and varies as a function of one's actual scaled score. Therefore, the CSEM may be especially useful in characterizing measurement precision in the neighborhood of a score level used for decision-making, such as cut scores for identifying students who meet a performance standard.

Technically, when a Rasch model is applied, the CSEM at any given point on the ability continuum is defined as the reciprocal of the square root of the test information function derived from the Rasch scaling model.

$$CSEM\,(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

where, $\text{CSEM}\left(\hat{\theta}\right)$ = conditional standard error of measurement and $I(\theta)$ = test information function. Test information depends on the sum of the corresponding information functions for the test items. Item information depends on each item's difficulty and conditional item score variance. The formula above utilizes the Rasch ability (θ) metric. The conditional standard error on the scaled score (SS) metric is determined by simply multiplying the CSEM $\left(\hat{\theta}\right)$ by the slope (multiplicative constant, m) of the linear transformation equation used to convert the Rasch ability estimates to scaled scores:

$$\text{CSEM(SS)} = CSEM\left(\hat{\theta}\right) * m,$$

Chapter Fourteen provides the linear transformation formulas for each PSSA-M test.

Rasch CSEM Confidence Intervals

CSEMs also allow statements regarding the precision of individual tests scores. And like SEMs, they help place reasonable limits around observed scaled scores through construction of an approximate score band. The confidence intervals are constructed by adding and subtracting a multiplicative factor of the CSEM and may be interpreted as described in the earlier section.

Further Interpretations

DIFFERENT CSEMS FOR DIFFERENT TEST SCORES

The CSEM approach provides different numerical estimates for constructing the confidence intervals for examinees depending on their specific score level. The magnitude of the CSEM values is "U" shaped with larger CSEM values associated with lower and higher scores.

GROUP SPECIFIC

Assuming reasonable model-data fit—as explored in Chapter Twelve—the Rasch based CSEMs (conditioned on score level) should not vary across groups.

SCALED SCORE METRIC

The CSEM and associated confidence interval bands are on the scaled score metric.

TYPE OF ERROR REFLECTED

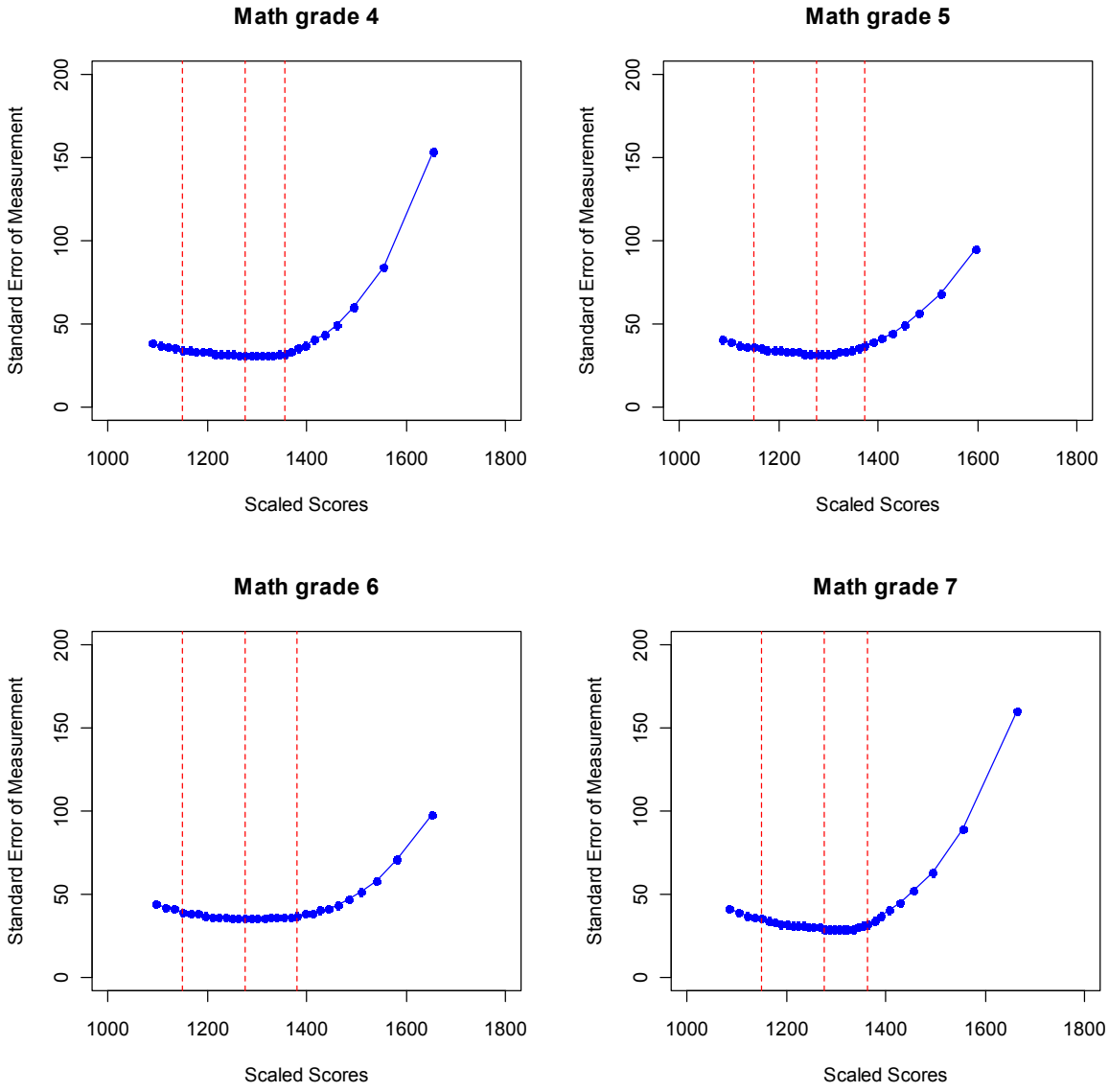
The SEMs documented on the PSSA-M score reports are the Rasch-based conditional standard errors of measurement described above. These are provided by the WINSTEPS scaling program described in Chapter Twelve. As noted earlier, these CSEMs are based on the concept of statistical information. For the purpose of providing a simpler explanation of SEMs to test score users, the earlier description of SEMs framed using the idea of internal consistency reliability was provided in the PSSA-M score report interpretive documents[17]. Score report content is considered in greater detail in Chapter Sixteen.

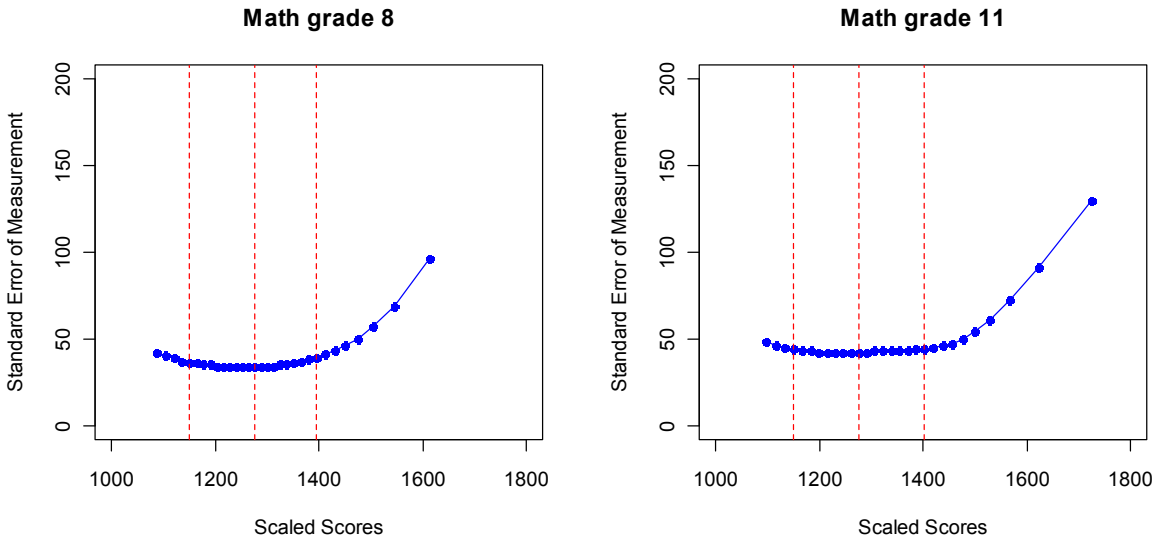
Results and Observations

Figure 18–2 shows the Rasch CSEMs associated with each scaled score level. (This information is also provided in tabular form in Appendix L.). Values were derived using the final data file described in Chapter Nine. The values are fairly consistent across a noticeably large range of the scaled scores, as demonstrated by the relatively flat bottoms of most plots. The values increase at both extremes (i.e., at smaller and larger scaled scores) giving these figures their typical u-shaped pattern. (Only the SEMs for scores greater than the lowest observable scaled scores (LOSS) are shown in the figures; consequently, the complete u-shape does not appear in most plots.) The three red-dashed lines represent the Basic-M, Proficient-M, and Advanced-M scaled score cuts, respectively, moving from lower to higher scaled score values. SEM values at the cut score lines were generally associated with smaller SEM values, indicating more precise measurement occurs at these cuts. This was particularly true for the Proficient-M and Advanced-M cuts.

[17] Because IRT CSEMs are based on statistical information, it is questionable if they account for error variance due to items. However, it seems difficulty to construct a simple explanation of IRT CSEMs for the general public.

Figure 18–2. Conditional Standard Error Plots for each Grade and Subject





DECISION CONSISTENCY

Classification consistency refers to the degree with which the achievement level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). In a standards-based testing program there should be great interest in knowing how accurately students are classified into performance categories. In contrast to Coefficient Alpha that is concerned with the relative rank-ordering of students, it is the absolute values of student scores that are important in decision consistency.

Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test? If two parallel forms of the test were given to the same students, the consistency of the measure would be reflected by the extent that the classification decisions made from the first set of test scores matched the decisions based on the second set of test scores. Consider Tables 18–4 and 18–5 below:

Table 18–4. Pseudo-Decision Table for Two Hypothetical Categories

		TEST ONE		
		LEVEL I	LEVEL II	MARGINAL
TEST TWO	LEVEL I	φ_{11}	φ_{12}	$\varphi_{1\bullet}$
	LEVEL II	φ_{21}	φ_{22}	$\varphi_{2\bullet}$
	MARGINAL	$\varphi_{\bullet 1}$	$\varphi_{\bullet 2}$	**1**

Table 18–5. Pseudo-Decision Table for Four Hypothetical Categories

		TEST ONE				
		LEVEL I	LEVEL II	LEVEL III	LEVEL IV	MARGINAL
TEST TWO	LEVEL I	φ_{11}	φ_{12}	φ_{13}	φ_{14}	$\varphi_{1\bullet}$
	LEVEL II	φ_{21}	φ_{22}	φ_{23}	φ_{24}	$\varphi_{2\bullet}$
	LEVEL III	φ_{31}	φ_{32}	φ_{33}	φ_{34}	$\varphi_{3\bullet}$
	LEVEL IV	φ_{41}	φ_{42}	φ_{43}	φ_{44}	$\varphi_{4\bullet}$
	MARGINAL	$\varphi_{\bullet 1}$	$\varphi_{\bullet 2}$	$\varphi_{\bullet 3}$	$\varphi_{\bullet 4}$	**1**

If a student is classified as being in one category based on Test One's score, how probable would it be that the student would be reclassified as being in the same category if he or she took Test Two (a non-overlapping, equally difficult form of the test)?

The proportions of correct decisions, φ for two and four categories are computed by the following two formulas, respectively:

$$\varphi = \varphi_{11} + \varphi_{22}$$
$$\varphi = \varphi_{11} + \varphi_{22} + \varphi_{33} + \varphi_{44.}$$

It is the sum of the diagonal entries—that is, the proportion of students classified by the two forms into exactly the same achievement level—that would signify the overall consistency.

Since it is not feasible to repeat PSSA-M testing in order to estimate the proportion of students who would be reclassified in the same performance levels, a statistical model needs to be imposed on the data in order to project the consistency of classifications solely using data from the available administration (Hambleton and Novick, 1973). Although a number of procedures are available, two well-known methods were developed by Hanson and Brennan (1990) and Livingston and Lewis (1995) utilizing specific True Score Models. These approaches are fairly complex and the cited sources contain details regarding the statistical models used to calculate decision consistency from the single PSSA-M administration.

Further Interpretations

Several factors might affect decision consistency. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications. Another factor is the location of the cut score in the score distribution. More consistent classifications are observed when the cut scores are located away from the mass of the score distribution. For example, when scores are close to being normally distributed, the mass is concentrated in the middle of the distribution, and thus, classifications tend to become more consistent when cut scores go up from 70 percent to 80 percent to 90 percent, or alternatively go down from 30 percent to 20 percent to 10 percent. The number of performance levels is also a consideration. Consistency indices for four performance levels should be lower than those based on two categories. This is not surprising since classification using four levels would allow more opportunity to change achievement levels; hence, there would be more classification errors with

four achievement levels, resulting in lower consistency indices. Lastly, some research has found that results from the Hanson and Brennan (1990) method on a dichotomized version of a complex assessment yields similar results to the Livingston and Lewis (1995) method (Stearns and Smith, 2007).

Results and Observations

The results for the overall consistency across all four performance levels as well as for the dichotomies created by the three cut scores are presented in Table 18–6. The tabled values—derived using the program *BB-Class* (Brennan, 2004)—showed that consistency values across the two methods were generally very similar. The Hanson and Brennan values were equal to or just slightly higher than the Livingston and Lewis values (by about 0.01) in most cases.

The overall decision consistency was generally in the mid 0.60s. It should be noted that the overall consistency indices (across all four performance levels) should logically be lower than those based on two categories (as discussed above).

Regarding dichotomous decisions, the Basic-M cuts generally had the highest consistency values at the lower grade levels where most exceeded 0.90. The Advanced-M cuts had the highest consistency values at the higher grade levels where most exceeded 0.90. Proficient-M cut decision consistency values were in the low to mid 0.80s at all grade levels.

As a point of comparison, recent general PSSA mathematics decision consistency values typically ranged from the Mid 0.90s to low 0.90s, with the Basic cut generally yielding the highest values and the Advanced cut yielding the lowest values. Overall consistency values were generally in the mid 0.70s. Thus, for the PSSA-M, some individual cut consistencies were as high as the generally PSSA, while the overall and Proficient-M cut consistencies were lower. The PSSA-M's shorter test length and lower reliabilities may have been contributing factors in these cases.

Table 18–6. Decision Consistency Results

Grade	Method	Overall	BBas/Bas	Bas/Prof	Prof/Adv
4	HB	0.63	0.95	0.82	0.86
	LL	0.63	0.95	0.81	0.86
5	HB	0.65	0.93	0.81	0.90
	LL	0.64	0.93	0.81	0.90
6	HB	0.65	0.92	0.80	0.92
	LL	0.65	0.91	0.80	0.92
7	HB	0.68	0.91	0.83	0.93
	LL	0.66	0.90	0.83	0.93
8	HB	0.67	0.90	0.82	0.95
	LL	0.67	0.90	0.82	0.95
11	HB	0.63	0.84	0.84	0.95
	LL	0.63	0.84	0.84	0.95

Note. Results derived using PSSA-M final data file (see Chapter Nine).

RATER AGREEMENT

Because open-ended items are included on the PSSA-M, another source of random error is related to the scorers of those items. Frisbie (2005) noted that "test score reliability differs from scorer reliability" and that "the need for one kind of estimate cannot be satisfied by the other." Additionally, the data most easily obtainable that captures this information comes from the "10 percent read behinds" collected during the scoring process (see Chapter Eight for a description). Partly because of the way this data is obtained and reported (i.e., it's not a ratio of true score variance over observed score variance), the term rater agreement is used here, not rater reliability or inter-rater reliability as these terms are somewhat misleading as explained above.

Further Interpretations

For the PSSA-M, only within-year consistency is available. In future administrations across-year rater consistency may be available for consideration as well.

Results and Observations

Within-year rater agreement information was provided in Chapter Eight. This information is reformatted in Table 18–7 for PSSA-M mathematics OE items. In addition, the percentages awarded to each score point are also presented in this table. The inter-rater agreement percentages (exact) generally ranged from the high 80s to high 90s. Validity indices generally ranged from the low 90s to high 90s. The tabled values are similar to results historically obtained for the general PSSA.

Table 18–7. Inter-Rater Agreement and Percentage Awarded for Each Score Point for OE Items—Mathematics

		Inter-Rater Agreement %			Percentage Awarded for Each Score Point %					
Grade	**Item**	**Exact**	**Adjacent**	**Validity**	**0**	**1**	**2**	**3**	**4**	**B/NS**
4	1	98	2	96	34	29	9	12	14	0
	2	100	0	92	31	32	21	9	6	0
5	1	96	4	89	61	21	8	7	2	1
	2	97	3	97	31	24	11	12	22	1
6	1	88	12	91	41	39	12	5	3	1
	2	99	1	91	16	54	21	7	1	1
7	1	96	4	96	26	30	16	17	11	1
	2	94	5	98	74	4	6	7	8	2
8	1	89	11	92	25	40	24	6	4	1
	2	87	13	98	33	37	16	10	3	2
11	1	93	7	94	49	35	7	2	3	4
	2	95	5	93	36	49	5	4	0	6

Note: B = blank; NS=non-scoreable. For more information regarding validity, see the section on Handscoring Validity Process in Chapter Eight.

Chapter Nineteen: Validity

As defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), validity refers to "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p.9). The *Standards* provides a framework for describing the sources of evidence that should be considered when evaluating validity. These sources include evidence based on: 1) test content, 2) response processes, 3) the internal structure of the test, 4) the relationships between test scores and other variables, and 5) the consequences of testing. In addition, when IRT models are used to analyze assessment data, validity considerations related to those processes should also be explored.

The validity process involves the collection of a variety of evidence to support the proposed test score interpretations and uses. The entire technical report describes the technical aspects of the PSSA-M tests in support of their score interpretations and uses. Each of the previous chapters contributes important evidence components that pertain to score validation: test development; test administration; test scoring; item analysis; Rasch calibration, scaling, and linking; score reporting; and reliability. This chapter is used to summarize and synthesize the evidence based on the *Standards'* framework. The purposes and intended uses of PSSA-M test scores are reviewed first and then each type of validity evidence is addressed in turn.

PURPOSES AND INTENDED USES OF THE PSSA-M

The *Standards* emphasize that validity pertains to how test scores are used. To help contextualize the evidence that will be presented below, the purposes of the PSSA-M will be reviewed first. As stated in Chapter One, the main purposes of the PSSA-M (as with the general PSSA) are to:

- Provide students, parents, educators, and citizens with an understanding of student and school performance.
- Determine the degree to which programs enable students to attain proficiency of academic standards.
- Provide results to school districts, including charter schools, and Career and Technical Centers (CTCs) for consideration in the development of strategic plans.
- Provide information to state policymakers, including the General Assembly, and the State Board, on how effective schools are in promoting and demonstrating student proficiency of the Academic Standards.
- Provide information to the general public on school performance.
- Provide results to school districts, including charter schools, and CTCs based on the aggregate performance of all students and for relevant subgroups, such as students with an IEP and for those without an IEP.

EVIDENCE BASED ON TEST CONTENT

Test content validity evidence for the PSSA-M rests greatly on establishing a link between each piece of the assessment (i.e., the items) and what the students should know and be able to do as required by the Assessment Anchors, Eligible Content, and/or the Academic Content Standards. The PSSA-M tests are intended to measure students' knowledge and skills described in the Assessment Anchors as defined by the Eligible Content for mathematics and thus the evidence supporting the alignment among the PSSA-M tasks and the Assessment Anchors as defined by the Eligible Content.

Lane (1999) suggests taking the following steps to support the validity of an assessment, such as the PSSA-M:

- Evaluate the degree to which the PSSA-M test specifications represent and align with the knowledge and skills described in the Assessment Anchors as defined by the Eligible Content for mathematics in terms of both content and cognitive processes.
- Evaluate the alignment between the PSSA-M items and test specifications to ensure representativeness.
- Evaluate the extent to which the curriculum aligns with the Assessment Anchors. If some contents are not included in the curriculum, then low scores on PSSA-M should not be interpreted as meaning that instruction was ineffective.
- Conduct content reviews of the PSSA-M items using a panel of content experts to see whether they measure the intended construct or they are the sources of construct-irrelevant variance.
- Conduct fairness reviews of the items to avoid issues related to a specific subpopulation.
- Evaluate procedures for administration and scoring such as the appropriateness of instructions to examinees, time limit for the assessment, and training of raters.
- Submit operational tests to third-party independent reviews.

Chapters 2–8 of this report present a considerable amount of evidence related to test content. As described in these chapters, all the PSSA-M items were developed and aligned with the Assessment Anchors and Eligible Content for mathematics following well-established procedures. After the items were developed, they underwent multiple rounds of content and bias reviews. After they were field tested, they were reviewed with respect to their statistical properties. Items selected for the operational assessment had to pass content, psychometric, and PDE reviews. Finally, the tests were administrated according to standardized procedures with allowable accommodations.

Some efforts made to ensure content validity are summarized below:

- DRC used Webb's (1999) DOK model to ensure the PSSA-M items aligned with the Assessment Anchors, as defined by the Eligible Content, and the Academic Content Standards in terms of both content and cognitive levels.
- DRC established detailed test and item/passage development specifications, and ensured the items were sufficient in number and adequately distributed across content, levels of cognitive complexity, and difficulty.

- DRC and WestEd selected qualified item writers and provided training to help ensure they wrote high-quality items.
- Each newly-developed item was first reviewed by content specialists and editors at DRC or/and WestEd to make sure that all items measured the intended Assessment Anchors, as defined by the Eligible Content for mathematics. Appropriateness for the intended grade was also considered, as well as depth of knowledge, graphics, grammar/punctuation, language demand, and distractor reasonableness.
- Prior to field testing, the test items were submitted to content committees (composed of Pennsylvania educators) for review using, but not limited to, the following categories:
  - Overall quality and clarity
  - Anchor, eligible content, and/or standard alignment
  - Grade-level appropriateness
  - Difficulty level
  - Depth of knowledge
  - Appropriate sources of challenge (e.g., unintended content and skills)
  - Correct answer
  - Quality of distractors
  - Graphics
  - Appropriate language demand
  - Freedom from bias
- The items were also submitted to a Bias, Fairness, and Sensitivity Committee for review. This committee reviewed items for issues related to diversity, gender, and other pertinent factors.
- Items passing all the prior hurdles were tried out in a field-test event. Several statistical analyses were conducted on the field test data including classical item analyses and distractor analyses. Items were once again carefully reviewed by DRC staff and a committee of Pennsylvania teachers with respect to their statistical characteristics.
- The PSSA-M tests were administrated according to standardized procedures with allowable accommodations. Students were given ample time to complete the tests (i.e., there were no speededness issues).
- As shown in Chapter Eight, the raters for OE items were carefully recruited and well trained. Their scoring was monitored throughout the scoring session to ensure that an acceptable level of scoring accuracy was maintained.

EVIDENCE BASED ON RESPONSE PROCESSES

Response-process evidence is used to examine the extent to which the cognitive skills and processes employed by students match that identified in the test developer's defined construct domains for all students and for each subgroup. Think-aloud procedures or Cognitive Interviews can be used to collect this type of evidence. In addition, when an assessment includes OE items, an examination of the extent to which the raters interpret and apply the scoring criteria accurately when assigning scores to students' responses on OE items also provides validity of the response-processes evidence.

Cognitive Interviews were conducted in Pennsylvania schools between May 11 and May 19, 2009. Information collected from these interviews was then used to aid decision-making in the strategies currently used to revise and/or enhance items for the PSSA-M to ensure that these enhancements would appropriately facilitate student access to the assessed content. See Chapter Three for information about the results of the Cognitive Interviews. For all the PSSA-M tests, well-organized scorer training and subsequent monitoring of rating accuracy helped ensure that raters strictly followed the scoring criteria and that no rubric-unrelated features significantly affected their scoring.

EVIDENCE BASED ON INTERNAL STRUCTURE

As described in the *Standards* (1999), internal-structure evidence refers to the degree to which the relationships among test items and test components conform to the construct on which the proposed test interpretations are based. For each PSSA-M test, one total test score as well as strand scores are reported (see Chapter Sixteen for more information about PSSA-M scores). Several dimensionality studies were conducted in order to provide internal-structure evidence relating to the use of both types of scores.

Item-Test Correlations

Item-test correlations were reviewed in Chapter Eleven. All values were positive. Although a few items had low correlations, the average correlation over all items appeared reasonable in magnitude.

IRT Dimensionality

Results from principle components analyses conducted using WINSTEPS were presented in Chapter Twelve. The PSSA-M mathematics tests were essentially unidimensional, providing evidence supporting interpretations based on the total scores for the respective PSSA-M tests.

Strand Correlations

Correlations and disattenuated correlations among strand scores within each subject area are presented below. Values were derived from the PSSA-M final data file (see Chapter Nine). This data can also provide information on score dimensionality that is part of internal-structure evidence. As noted in Chapter Three, the PSSA-M mathematics tests have five domains (denoted by M.A, M.B, M.C, M.D, and M.E).

For each grade, Pearson's correlation coefficients among these domains are reported in Tables 19–1a through 19–1f. The inter-correlations among the strands within the content areas were positive and generally moderate in value.

Table 19–1a. Correlations among Mathematics Strands for Grade 4

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.54*	-			
M.C	*0.39*	*0.28*	-		
M.D	*0.50*	*0.42*	*0.32*	-	
M.E	*0.47*	*0.39*	*0.38*	*0.45*	-

Table 19–1b. Correlations among Mathematics Strands for Grade 5

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.51*	-			
M.C	*0.38*	*0.38*	-		
M.D	*0.52*	*0.48*	*0.37*	-	
M.E	*0.43*	*0.38*	*0.38*	*0.44*	-

Table 19–1c. Correlations among Mathematics Strands for Grade 6

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.41*	-			
M.C	*0.44*	*0.36*	-		
M.D	*0.51*	*0.33*	*0.35*	-	
M.E	*0.50*	*0.32*	*0.39*	*0.41*	-

Table 19–1d. Correlations among Mathematics Strands for Grade 7

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.49*	-			
M.C	*0.53*	*0.45*	-		
M.D	*0.51*	*0.42*	*0.44*	-	
M.E	*0.46*	*0.38*	*0.43*	*0.40*	-

Table 19–1e. Correlations among Mathematics Strands for Grade 8

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.43*	-			
M.C	*0.37*	*0.34*	-		
M.D	*0.51*	*0.45*	*0.45*	-	
M.E	*0.44*	*0.39*	*0.39*	*0.52*	-

Table 19–1f. Correlations among Mathematics, Strands for Grade 11

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.35*	-			
M.C	*0.37*	*0.37*	-		
M.D	*0.48*	*0.49*	*0.43*	-	
M.E	*0.39*	*0.36*	*0.36*	*0.50*	-

The correlations in Tables 19–1a through 19–1f are for the observed strand scores. These observed-score correlations are weakened by existing measurement error contained within each strand. As a result, disattenuating the observed correlations can provide an estimate of the relationships among strands if there were no measurement error. (An important caveat is provided further below.) The disattenuated correlation coefficients (R_{xy}) can be computed by using the formula (Spearman 1904, 1910) below:

$$R_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}},$$

where r_{xy} is the observed correlation, and r_{xx} and r_{yy} are the reliabilities for Strand X and Strand Y. Tables 19–2a through 19–2f show the corresponding disattenuated correlations.

Disattenuated correlations very near 1.0 might suggest that the same or very similar constructs are being measured. Values somewhat less than 1.0 might suggest that different strands are measuring slightly different aspects of the same construct. Values markedly less than 1.0 might suggest the strands reflect different constructs.

Given that none of these strands have perfect reliabilities (see Chapter Eighteen), the disattenuated strand correlations are higher than their observed score counterparts. Within-subject strand correlations varied considerably in value. Some within-subject correlations were very high (e.g., above 0.95). As noted above, extremely high disattenuated correlations suggest that the within-subject strands might be measuring essentially the same construct. This, in turn, suggests that some strand scores might not provide unique information about the strengths or weakness of students.

On the other hand, there were some within-subject strand correlations that were somewhat lower than 1.0. For such strands, partial evidence is provided regarding the multidimensional structure of some tests and further supporting the validity of those specific strand scores.

Table 19–2a. Disattenuated Strand Correlations for Mathematics: Grade 4

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*1.26*	-			
M.C	*0.73*	*0.85*	-		
M.D	*0.90*	*1.23*	*0.75*	-	
M.E	*0.76*	*1.01*	*0.80*	*0.90*	-

Table 19–2b. Disattenuated Strand Correlations for Mathematics: Grade 5

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.98*	-			
M.C	*0.72*	*0.85*	-		
M.D	*1.17*	*1.29*	*0.97*	-	
M.E	*0.84*	*0.88*	*0.85*	*1.19*	-

Table 19–2c. Disattenuated Strand Correlations for Mathematics: Grade 6

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.97*	-			
M.C	*0.87*	*1.00*	-		
M.D	*0.99*	*0.90*	*0.82*	-	
M.E	*0.95*	*0.87*	*0.89*	*0.91*	-

Table 19–2d. Disattenuated Strand Correlations for Mathematics: Grade 7

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*1.00*	-			
M.C	*1.03*	*0.95*	-		
M.D	*0.98*	*0.89*	*0.87*	-	
M.E	*0.89*	*0.81*	*0.85*	*0.79*	-

Table 19–2e. Disattenuated Strand Correlations for Mathematics: Grade 8

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*0.99*	-			
M.C	*0.77*	*0.74*	-		
M.D	*1.03*	*0.95*	*0.85*	-	
M.E	*0.89*	*0.84*	*0.76*	*0.98*	-

Table 19–2f. Disattenuated Strand Correlations for Mathematics: Grade 11

	M.A	M.B	M.C	M.D	M.E
M.A	-				
M.B	*1.14*	-			
M.C	*0.96*	*1.25*	-		
M.D	*0.91*	*1.23*	*0.86*	-	
M.E	*0.93*	*1.14*	*0.89*	*0.91*	-

It should be noted that much caution is needed in interpreting the disattenuated results because the reliabilities used to calculate the disattenuated correlations are subject to both upward and downward biases. (These are discussed in some detail in Chapter Eighteen.) Consequently, some of the values tabled above may be higher or lower than they should be, depending on which bias prevails for any given pair of strand scores. When the reliabilities are lower than they should be, the disattenuated correlations will be inflated (and in many instances appear larger than the theoretical correlation maximum value of 1.0).

Exploratory Factor Analysis

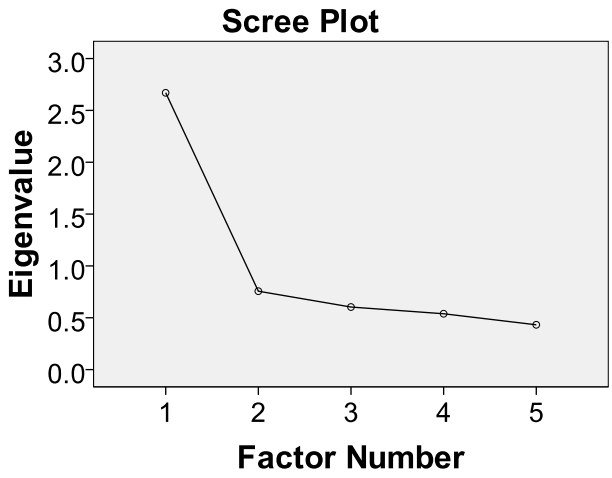
In order to further explore the internal structure of the PSSA-M, an exploratory factor analysis (EFA) of the mathematics' strand scores was conducted. The PSSA-M final data file (see Chapter Nine) was used to create the observed correlation matrices shown in Tables 19–1a through 19–1g, which in turn were used in the EFAs. In SPSS, Principle Axis Factor extraction was utilized with an oblique rotation (Promax) of the initial factor solution to improve interpretability. Oblique rotations allow for correlated factors which seemed more appropriate for the PSSA-M tests because of apriori expectations that academic achievement across subject areas should be correlated.

Table 19–3 presents the eigenvalues and the explained variance for the extracted factors for the Grade 4 PSSA-M test. The Scree Plot graphing the eigenvalues against the factor number is shown in Figure 19–1. The first factor accounted for over 50 percent of the total variance, while the second factor explained about 15 percent of the total variance. Only the first factor had an eigenvalue greater than 1.0, typically suggesting a one-factor solution using the Kaiser criterion. Also, the one-factor solution did not yield many large fitted residual values in the reproduced correlation matrix (i.e., only 1 of 10 residuals was greater than 0.05).

Table 19–3. Eigenvalues and Explained Variance for Grade 4

Factor	Eigenvalue	%
1	2.669	53.374
2	0.757	15.131
3	0.604	12.071
4	0.539	10.776
5	0.432	8.648

Figure 19–1. Scree Plot for Grade 4



The loadings resulting from one-factor solution are presented in Table 19–4a. The loadings were reasonably high.

Table 19–4a. Factor Loadings for Grade 4

	Factor
Domain	**1**
Mathematics	
M.A	.774
M.B	.635
M.C	.503
M.D	.663
M.E	.648

Similar results were found at the other grades. The eigenvalue scree plots consistently indicted a one-factor solution. (The eigenvalues and explained variances are not shown for the other grades due to space considerations.) Factor loadings are reported in Tables 19–4b through 19–4f for the remaining grades.

Table 19–4b. Factor Loadings for Grade 5

	Factor
Domain	**1**
Mathematics	
M.A	.730
M.B	.674
M.C	.551
M.D	.706
M.E	.611

Table 19–4c. Factor Loadings for Grade 6

	Factor
Domain	**1**
Mathematics	.773
M.A	.538
M.B	.593
M.C	.629
M.D	.639
M.E	.773

Table 19–4d. Factor Loadings for Grade 7

	Factor
Domain	**1**
Mathematics	
M.A	.771
M.B	.636
M.C	.693
M.D	.653
M.E	.607

Table 19–4e. Factor Loadings for Grade 8

	Factor
Domain	**1**
Mathematics	
M.A	.669
M.B	.599
M.C	.574
M.D	.764
M.E	.666

Table 19–4f. Factor Loadings for Grade 11

	Factor
Domain	**1**
Mathematics	
M.A	.612
M.B	.606
M.C	.581
M.D	.788
M.E	.619

Taken as a whole, all the internal structure evidence presented above generally indicates that related elements of each of the PSSA-M mathematics tests were correlated in the intended manner. Studies in future years will investigate whether different PSSA-M subject area tests seem to measure different constructs.

Since the strands in each content area will be designed to measure distinct components, it is reasonable to expect that the inter-subject strand correlations should be positive and strong, but ideally, not extremely high. While there is content rationale underlying the creation of each strand score, the empirical correlations will provide additional evidence about the reasonableness of using strand scores as a way to identify an individual student's strengths and weaknesses. As of now, instructional programs should not be based on strand score information alone, but used only in conjunction with other sources of evidence available (e.g., teacher observations and other exam performance).

Differential Item Functioning

Differential item functioning (DIF) occurs when examinees with the same ability level but different group memberships do not have the same probability of answering the item correctly. This pattern of results may suggest the presence of item bias. As a statistical concept, however, DIF can be differentiated from item bias, which is a content issue that can arise when an item presents negative group stereotypes, uses language that is more familiar to one subpopulation than to another, or is presented in a format that disadvantages certain learning styles. While the source of item bias is often plain to trained judges, DIF may have no clear cause. However, studying how DIF arises and how it presents itself has an effect on how to detect and correct it.

LIMITATIONS OF STATISTICAL DETECTION

No statistical procedure should be used as a substitute for rigorous, hands-on reviews by content and bias specialists. The statistical results can help organize the review so the effort is concentrated on the most problematic cases. Further, no items should be automatically rejected simply because a statistical method flagged them or accepted because they were not flagged.

Statistical detection of DIF is an inexact science. There have been a variety of methods proposed for detecting DIF, but no one statistic can be considered either necessary or sufficient. Different methods are more or less successful depending on the situation. No analysis can guarantee that a test is free of bias, but almost any thoughtful analysis will uncover the most flagrant problems.

A fundamental shortcoming of all statistical methods used in DIF evaluation is that all are intrinsic to the test being evaluated. If a test is unbiased overall but contains one or two DIF items, any method will locate the problems. If, however, all items on the test show consistent DIF to the disadvantage of a given subpopulation, a statistical analysis of the items will not be able to separate DIF effects from true differences in achievement.

MANTEL-HAENSZEL PROCEDURE FOR DIFFERENTIAL ITEM FUNCTIONING

The *Mantel-Haenszel* procedure for detecting differential item functioning is a commonly used technique in educational testing. It does not depend on the application or the fit of any specific measurement model. However, it does have significant philosophical overlap with the Rasch model since it uses a test's total score to organize the analysis.

The procedure as implemented by DRC contrasts a focal group with a reference group. While it makes no practical difference in the analysis which group is defined as the focal group, the group most apt to be disadvantaged by a biased measurement is typically defined as the focal group. In these analyses, the focal group was female for gender-based DIF and black for ethnicity-based DIF; reference groups were male and white, respectively. The Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) for each item is computed from a contingency table. It has two groups (focal and reference) and two outcomes (right or wrong). The ability groups are defined by the test's score distribution for the total examinee populations.

The basic MH statistic is a single degree of freedom chi-square that compares the observed number in each cell to the expected number. The expected counts are computed to ensure that the analysis is not confounded with differences in the achievement level of the two groups.

For OE items, a comparable statistic is computed based on the standardized mean difference (SMD) (Dorans, Schmitt & Bleistein, 1992), computed as the differences in mean scores for the focal and reference groups if both groups had the same score distribution.

To assist the review committees in interpreting the analyses, the items are assigned a severity code based on the magnitude of the MH statistic. Items classified as A+ or A- have little or no statistical indication of DIF. Items classified as B+ or B- have a moderate indication of DIF but may be judged to be acceptable for future use. Items classified as C+ or C- have strong evidence of DIF. The plus sign indicates that the item favors the focal group and a minus sign indicates that the item favors the reference group.

Counts of the number of items from each grade and content area that were assigned to each severity code are shown below in Table 19–5a (MC items) and 19–5b (OE items). DIF analyses were conducted only on operational items. Only a handful of items reached the C magnitude[18].

Table 19–5a. DIF Summary—MC Items

	Male/Female							White/Black						
	A+	**A-**	**B+**	**B-**	**C+**	**C-**	**Tot**	**A+**	**A-**	**B+**	**B-**	**C+**	**C-**	**Tot**
4	15	15	0	0	0	0	30	13	16	0	1	0	0	30
5	13	14	1	1	0	1	30	13	16	0	0	0	1	30
6	15	15	0	0	0	0	30	15	13	0	2	0	0	30
7	20	7	0	1	0	2	30	12	17	0	1	0	0	30
8	14	14	1	1	0	0	30	16	14	0	0	0	0	30
11	19	7	0	4	0	0	30	9	21	0	0	0	0	30

[18] These results are based on the final data set as described in Chapter Nine. Nearly all PSSA-M items are modified versions of general PSSA items that were previously screened for DIF and approved for use on the general assessment.

Table 19–5b. DIF Summary—OE Items

	Male/Female							White/Black						
	A+	**A-**	**B+**	**B-**	**C+**	**C-**	**Tot**	**A+**	**A-**	**B+**	**B-**	**C+**	**C-**	**Tot**
4	1	1	0	0	0	0	2	0	2	0	0	0	0	2
5	2	0	0	0	0	0	2	1	0	1	0	0	0	2
6	1	1	0	0	0	0	2	1	1	0	0	0	0	2
7	1	1	0	0	0	0	2	2	0	0	0	0	0	2
8	0	2	0	0	0	0	2	0	2	0	0	0	0	2
11	1	1	0	0	0	0	2	1	1	0	0	0	0	2

EVIDENCE BASED ON CONSEQUENCES OF TESTING

Based on the *Standards* (1999), evidence of the consequences of implementing an assessment program is an additional source of validity information. One must investigate both positive and negative (intended and unintended) consequences of score-based inferences to fully evaluate the pool of validity evidence.

Lane and Stone (2002) summarized the general intended consequences for state assessments and accountability programs:

- Student, teacher, and administrator motivation and effort.
- Curriculum and instruction practices (including content and strategies).
- Improved learning for all students.
- Content and format of classroom assessments.
- Professional development support.
- Use and nature of test preparation activities.
- Student, teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, and the use of assessment results.

Evidence for the intended improvement of student learning can be seen by looking at the increasing percentage of students who are Proficient-M or Advanced-M across years. The following tables provide the percentages of students who are Proficient-M or Advanced-M by grade, year, and subject. Values were derived from the PSSA-M final data file (see Chapter Nine). As this was the first administration of the PSSA-M mathematics test, results will need to be monitored overtime using these results as the baseline.

Table 19–6a. Percentage of Students Scoring in the Proficient-M or Advanced-M Category: Mathematics[19]

Grade	PSSA-M 2010	PSSA 2010
4	59.5	84.8
5	51.0	74.4
6	48.1	78.0
7	41.3	78.0
8	40.8	75.1
11	33.2	59.6

Lane and Stone (2002) also summarized the possible unintended outcomes:

- Narrowing of curriculum and instruction to focus only the specific standards assessed and ignoring the broader construct reflected in the specified standards.
- The use of test preparation materials that are closely linked to the assessment without making changes to instruction.
- The use of unethical test preparation materials or administration procedures.
- Differential performance gains for subgroups of students.
- Inappropriate or unfair uses of test scores, such as questionable practices in reassignment of teachers or principles.
- For some students, decreased confidence and motivation to learn and to perform well on the assessment because of past experiences with assessments.

As noted above, one important piece of consequential evidence pertains to the use of assessment results. As shown in Chapter Sixteen, there are several different types of scores and score reports used for the PSSA-M. The extent to which various groups of users (e.g., students, teachers, and parents) interpret these scores and reports appropriately would affect the validity of subsequent uses of these results. Chapter Sixteen of this technical report is intended to provide accurate and clear test score and report information with the hope that this will help users avoid unintended uses and interpretations of the PSSA-M results. Nevertheless, evidence pertaining to other consequences of the PSSA-M needs continued research.

[19] These are not the final official results as the appeals process was still ongoing at the time this report was written. Official results will be posted on PDE's website. PSSA results added for reference purposes.

EVIDENCE RELATED TO THE USE OF THE RASCH MODEL

Since the Rasch model is the basis of all calibration, scaling, and linking analyses associated with the PSSA-M, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met as well as the fit between the model and test data. As discussed at length in Chapter Twelve, the underlying assumptions of Rasch models were essentially met for all the PSSA-M data, indicating the appropriateness of using the Rasch models to analyze the PSSA-M data.

In the future, the Rasch model will be used to link different operational PSSA-M tests across years. The accuracy of the linking will also affect the accuracy of student scores and the validity of score uses. As described in Chapter Fifteen, DRC Psychometric Services staffers will follow a well-prescribed linking procedure.

VALIDITY EVIDENCE SUMMARY

Validity evidence related to test content was reviewed earlier in this chapter. On the whole, the early chapters of this technical report show that a strong link can be established between each PSSA-M item and its associated eligible content. Details regarding how the PSSA-M operational assessments were assembled to reflect the state content standards and detailed information regarding educator reviews (including content, bias, and sensitivity reviews) are presented in Chapter Three.

Validity of score inferences is bolstered when test scores are consistent. Here, the reliabilities of the total test scores (presented in Chapter Eighteen) were on the low end of the adequate range. Considering the length of the tests and the relatively homogeneous achievement level of test takers, the reported values are reasonable.

As reported above, differential item functioning (DIF) with respect to gender and ethnicity helps address construct-irrelevant variance, which represents an important threat to the validity of inferences made from achievement test scores. Only a very small percentage of items was flagged for severe DIF.

References

Achieve, Inc. (2005). Measuring Up 2005: A Report on Assessment Anchors and Tests in Reading and Mathematics for Pennsylvania. Washington, DC: Achieve, Inc.

AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (1999). *Standards for Educational and Psychological Tests*. Washington, DC: American Educational Research Association.

Allman, C. (2004). Test access: Making tests accessible for students with visual impairments – A guide for test publishers, test developers, and state assessment personnel (2nd edition). Louisville, KY: American Printing House for the Blind. Available from http://www.aph.org.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5–9.

Brennan, R. (2004). *BB-Class (Version 1.0)*. CASMA: education.uiowa.edu/casma. Computer program.

Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Cook, L. L. & Eignor, D. R. (1991). NCME Instructional module: IRT equating methods. *Educational Measurement: Issues and Practice, 17*(1), 5–9.

Cronbach, L. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education. Educational Measurement: Issues and Practice, 10, 37–45.

Cronbach, L. & Shavelson R. L. (2004). My current thoughts on Coefficient Alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391–418.

Data Recognition Corporation. (2000). *Item Viewer and Authoring Network (IVAN): Informational Guide*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2003–2007). *Fairness in Testing: Training Manual for Issues of Bias, Fairness, and Sensitivity*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2004–2007). *Pennsylvania System of School Assessment (PSSA) Style Guide.* Maple Grove, MN: DRC.

Data Recognition Corporation. (2005, December). *Technical Report for the PSSA 2005 Reading and Mathematics*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2007, May). *Technical Report for the PSSA 2006 Reading and Mathematics: Grades 4, 6, and 7*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2007, May). *Technical Report for the PSSA 2006 Writing: Grades 5, 8, and 11*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2010). *Technical Report for the 2010 Pennsylvania System of School Assessment*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2007, July). *PSSA Writing Test Score Reliability: Some Available Approaches and Possible Alternatives.* (PSSA TAC Document 071907_5). Maple Grove, MN: Bishop, N.

Data Recognition Corporation. (2007). *Preliminary Technical Report for 2008 PSSA Science*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2008, February). T*echnical Report for the PSSA 2007 Writing: Grades 5, 8, and 11*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2008, February). *Technical Report for the PSSA 2007 Reading and Mathematics: Grades 3, 4, 5, 6, 7, 8, and 11*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2008, February). *Preliminary Technical Report for 2008 PSSA Science,* Maple Grove, MN: DRC.

Data Recognition Corporation. (2009, June). *Rater Effect Study Results.* (PSSA TAC Document 06.03.09 E). Maple Grove, MN: Stearns, M.

Dorans, N., Schmitt, Al, & Bleistein, C. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, 29, 309–319.

Ericsson, K. & Simon, H. (1980). Verbal reports as data. *Psychological Review*, 87, 215–250.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data.* Cambridge: MIT.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 105–146). New York, NY: ACE/Macmillan.

Frisbie, D. A. (2005). Measurement 101: Some Fundamentals Revisited. *Educational Measurement: Issues and Practice, 24*(3) 21–28.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley and Sons.

Haertel, E. H. (2006). Reliability. In Brennan, R. L. (Ed.). *Educational Measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Hambleton, R. & Novick, M. (1973). "Toward an Integration of Theory and Method for Criterion-Referenced Tests." *Journal of Educational Measurement*, 10, 159–170.

Hanson, B. A., & Brennan, R. L. (1990). An Investigation of Classification Consistency Indexes Estimated Under Alternative Strong True Score Theory Models. *Journal of Educational Measurement, 27(4),* 345–359.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices, 10*(2) 33–41.

Huynh, H. (1976). "On the Reliability of Decisions in Domain-Referenced Testing." *Journal of Educational Measurement*, 13, 253–264.

Johnstone, C., Altman, J., & Thurlow, M. (2006). "A state guide to the development of universally designed assessments." University of Minnesota, Minneapolis, MN: National Center on Educational Outcomes.

Koger, M. E., Thacker, A. A. & Dickinson, E. R. (2004). *Relationships among the Pennsylvania System of School Assessment (PSSA) scores, SAT scores, and self-reported high school grades for the classes of 2002 and 2003*. (HumRRO Report FR-04-26), Louisville, KY: Human Resources Research Organization.

Kopriva, R. (2001). *ELL validity research designs for state academic assessments: An outline of five research designs evaluating the validity of large-scale assessments for English language learners and other test takers.* Paper presented at the CCSSO Annual Conference on Large Scale Assessment, Houston, TX.

Lane, S. (1999). *Validity evidence for assessments.* Paper presented at the 1999 Edward F. Reidy Interactive Lecture Series, Providence, RI.

Lane, S. & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23–30.

Lewis, D. M., Mitzel, H. C. & Green, D. R. (1996). *Standard Setting: A Bookmark Approach*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment: Phoenix, AZ.

Linacre, J. M. (2009). *A user's guide to WINSTEPS MININSTEP Rasch-model computer programs*. Chicago, IL: Winsteps.

Linacre, J. M., & Wright, B. D. (2003). *WINSTEPS 3.54: Multiple-choice, rating scale, and partial credit Rasch analysis* [computer software]. Chicago: MESA Press.

Livingston, S. & Lewis, C. (1995). "Estimating the Consistency and Accuracy of Classifications Based on Test Scores." *Journal of Educational Measurement* 32, 179–197.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

Marais, I. & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research, 14,* 21–38.

Messick, S. (1989). Validity. In R. L. (Ed.) *Educational Measurement* (3rd ed., p.3–104). New York: American Council on Education.

National Research Council. (2001). *Knowing what students know.* Washington, DC: National Academy of Sciences.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Paulsen & Levine, R. (1999). *The applicability of the cognitive laboratory method to the development of achievement test items.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Pennsylvania State Board of Education. (1999, January). *Chapter 4. Academic Standards and Assessment*. Harrisburg, PA: Pennsylvania State Board of Education. Retrieved November 8, 2004, from http://www.pde.state.pa. Also available from http://www.pacode.com/secure/data/022/Chapter4/s4.51.html.

Pennsylvania Department of Education. (2004). *Mathematics Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved December 13, 2004, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2004). *Reading Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved December 13, 2004, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2004, April). *Assessment Anchors and Eligible Content*. Harrisburg, PA: PDE. Retrieved December 13, 2004, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2004, November). *Mathematics Assessment Handbook*. Harrisburg, PA: PDE. Retrieved December 13, 2004, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2004, November). *Reading Assessment Handbook*. Harrisburg, PA: PDE. Retrieved December 13, 2004, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005, December). *2005–2006 Mathematics Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005, December). *2005–2006 Reading Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005). *2005–2006 Mathematics Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005). *2005–2006 Reading Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005, December). *2005–2006 Writing Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2005). *2005–2006 Writing Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved September 14, 2005, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006). *2006–2007 Mathematics Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006). *2006–2007 Reading Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006). *2006–2007 Writing Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006, December). *2006–2007 Writing Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2006, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006). *2006–2007 Science Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved March 15, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2006, November). *Science Assessment Handbook*. Harrisburg, PA: PDE. Retrieved March 15, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2007, January). *2006–2007 Mathematics Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2007, January). *2006–2007 Reading Assessment Handbook*. Harrisburg, PA: PDE. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2007, January). *2007 Accommodations Guidelines for Students with IEPs, Students with 504 Plans, English Language Learners, and All Students*. Harrisburg, PA: PDE. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2007). *Assessment Anchors and Eligible Content*. Harrisburg, PA: PDE. Retrieved May 27, 2010 from http://www.pdesas.org/standard/AnchorsDownloads

Pennsylvania Department of Education. (2007). *PSSA 2007 Handbook for Assessment Coordinators and Administrators: Grades 3–8 and 11 Reading and Mathematics*. Harrisburg, PA: PDE. Retrieved January 30, 2007, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2007, March). *PSSA Reading and Mathematics Directions for Administration Manual*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved April 2, 2007, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2007). *2008 PSSA Accommodations Guidelines for Students with IEPs and Students with 504 Plans.* Harrisburg, PA: PDE. Retrieved March 4, 2008, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2008). *PSSA 2008 Handbook for Assessment Coordinators and Administrators: Grades 3–8 and 11 Reading and Mathematics*. Harrisburg, PA: Retrieved March 4, 2008, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2009). *PSSA Accommodations Guidelines for Students with IEPs and Students with 504 Plans.* Harrisburg, PA: PDE. Retrieved February 10, 2009, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2009). *Cognitive Interviews in Pennsylvania: Report on Data Collection for the Pennsylvania System of School Assessment Alternate Assessment with Modified Achievement Standards (PSSA-M) Study.* Harrisburg, PA: PDE.

Pennsylvania Department of Education. (2010). *PSSA and PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans, Revised 1-11-2010.* Harrisburg, PA: PDE. Retrieved February 24, 2010, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2009). *2008–2009 Assessment Handbook*. Harrisburg, PA: PDE. Retrieved February 10, 2009, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2010). *2009–2010 Assessment Handbook*. Harrisburg, PA: PDE. Retrieved February 24, 2010 from http://www.pde.state.pa.

Pennsylvania Department of Education. (2009). The *2008–2009 PSSA Handbook for Assessment Coordinators: Writing, Reading and Mathematics, Science*. Harrisburg, PA: Retrieved February 10, 2009, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2010). The *2009-2010 PSSA Handbook for Assessment Coordinators: Writing, Reading and Mathematics, Science*. Harrisburg, PA: Retrieved February 24, 2010, from http://www.pde.state.pa.us.

Pennsylvania Department of Education. (2008). *2008–2009 Mathematics Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved February 10, 2009, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2008). *2008–2009 Reading Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved February 10, 2009, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2008). *2008–2009 Science Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved February 10, 2009, from http://www.pde.state.pa.us

Pennsylvania Department of Education. (2008). *2008–2009 Writing Item and Scoring Sampler*. Harrisburg, PA: PDE. Posted separately by grade level. Retrieved February 10, 2009, from http://www.pde.state.pa.us

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*(2), 111–120.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.W. (in press). Accommodations for English Language Learners. San Francisco, CA. WestEd.

Swineford, F. (1956). *Technical Manual for Users of Test Analysis*. Statistical Report 56-42. Princeton, NJ: Educational Testing Service.

Sinclair, A. L. & Thacker, A. A. (2005). *Relationships among Pennsylvania System of School Assessment (PSSA) scores, university proficiency exam Scores, and college course grades in English and Math*. (HumRRO Report FR-05-55), Louisville, KY: Human Resources Research Organization.

Solano-Flores, G. & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English language learners. *Educational Researcher*, *32*(2), 3–13.

Spearman C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Stearns, M., & Smith R. M. (2007). *Estimation of classification consistency indices for complex assessments: Model based approaches*. Paper presented at the 2007 Annual Convention of the American Educational Research Association. Chicago, IL.

Swineford, F. (1956). Technical Manual for Users of Test Analysis. Statistical Report 56–42. Princeton, NJ: Educational Testing Service.

Thacker, A. A. & Dickinson, E. R. (2004). *Item Content and Difficulty Mapping by Form and Item Type for the 2001–2003 Pennsylvania System of School Assessment (PSSA)*. Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Dickinson, E. R., & Koger, M. E. (2004). *Relationships among the Pennsylvania System of School Assessment (PSSA) and other commonly administered assessments*. (HumRRO Report FR-04-33), Louisville, KY: Human Resources Research Organization.

Thompson, S., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44), Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and application* (Measurement) Methods for the Social Science, Vol. 3). Thousand Oaks: Sage publications.

Webb, N.L. (1997). Criteria for alignment of expectations and tests in mathematics and science education (NISE Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Webb, N. L. (1999). *Research Monograph No. 18: Alignment of Science and Mathematics Standards and Assessments in Four States.* Madison, WI: National Institute for Science Education.

Webb, N.L. (2002). Alignment study in language arts, mathematics, science, and social studies of state standards and tests for four states: State Collaborative on Test and State Standards (SCASS). Technical Issues in Large-Scale Test (TILSA): University of Wisconsin, Wisconsin Center for Education Research.

WINSTEPS (2000). *WINSTEPS® Rasch Measurement.* Copyright John M. Linacre.

Wright, B. & Masters, G. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local Item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Appendix A:

Assessment Anchor Explanations

About the Mathematics Assessment Anchors*

Introduction
This is a brief introduction to the Mathematics Assessment Anchors for the PSSA-M (found on the PDE website). The Assessment Anchors for the PSSA-M are exactly the same as the Assessment Anchors for the PSSA. For more information on the Assessment Anchors and how they were developed, please read the *Assessment Anchor Introduction* provided on the website at http://www.education.state.pa.us.

How the Assessment Anchors Connect to the Standards
The PA Academic Standards for Mathematics are:

2.1 Numbers, Number Systems and Number Relationships
2.2 Computation and Estimation
2.3 Measurement and Estimation
2.4 Mathematical Reasoning and Connections
2.5 Mathematical Problem Solving and Communication
2.6 Statistics and Data Analysis
2.7 Probability and Predictions
2.8 Algebra and Functions
2.9 Geometry
2.10 Trigonometry
2.11 Concepts of Calculus

All of the Mathematics Standards categories are still included on the PSSA and PSSA-M but the Assessment Anchors tighten the focus of what is assessed. The Assessment Anchors also clarify what is expected from grade level to grade level. There is a clear vertical alignment in the Assessment Anchors that did not exist in the standards. Teachers will be able to see how concepts build on one another from year to year. In addition, the Assessment Anchors have fewer Reporting Categories to help create more valid scores (there are more items per reporting category). Rather than report student results in all 11 standards, the reports will be organized into five major categories.

How the Assessment Anchors are Organized
These categories are similar to the five NCTM (National Council of Teachers of Mathematics) Standards and the five NAEP (National Assessment of Educational Progress) Reporting Categories. Each PA Standard Category was examined and then placed in the appropriate Reporting Category. Some of the specific Standards Statements cut across different Reporting Categories (e.g., 2.11- Concepts of Calculus, which occurs in different categories rather than being a separate category). The following is a general summary of where the bulk of the PA Mathematics Standards can be found in the Reporting Categories:

*Modified from the document originally created by the Pennsylvania Department of Education

Reporting Category	Standard
A. Numbers & Operations	2.1 (Numbers) & 2.2 (Computation)
B. Measurement	2.3 (Measurement)
C. Geometry	2.9 (Geometry) & 2.10 (Trigonometry)
D. Algebraic Concepts	2.8 (Algebra)
E. Data Analysis & Probability	2.6 (Statistics & Data) & 2.7 (Probability)

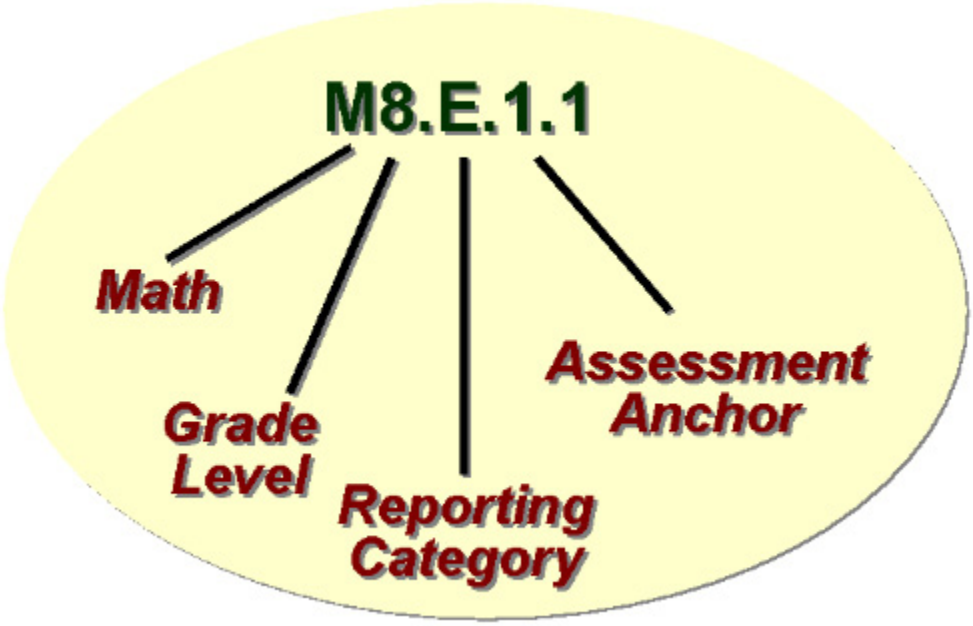
Important Patterns

The PA Mathematics Standards 2.4 (Reasoning) and 2.5 (Problem Solving) are not listed in the chart above. These two standards are not included because the above Reporting Categories focus on **content** (not **process)** and both Reasoning and Problem Solving are processes. However, knowing how to perform these processes is a very important part of the PSSA-M. Most of the multiple-choice items and all of the open-ended items will require students to know how to reason and solve problems, in addition to being knowledgeable about the content area being assessed. Even though Problem Solving is not one of the five content Reporting Categories, the PSSA-M will still show a separate score for the open- ended items on the school report, reflecting students' problem solving performance.

How to Read the Assessment Anchors

The Mathematics Assessment Anchors begin with an "M" to distinguish them from the Reading Assessment Anchors, which use an "R". The number after the "M" in the label is the grade level (e.g., M8 would be Mathematics at eighth grade). The second letter in the code system is the Reporting Category (A through E). The same reporting categories continue across all Grade levels, 4 through 8 and 11. The final number in the code is the actual Assessment Anchor (e.g., 1.1, 1.2, 1.3 etc.). Essentially, you read the Assessment Anchors like an outline, with the Assessment Anchor shaded across the top of the page and more specific details underneath.

For example, M8.E.1.1 is a Mathematics Assessment Anchor (M stands for Math) at 8th Grade (8). The E indicates that this Anchor is in the Data Analysis and Probability Reporting Category, and the 1.1 means that it is the first Assessment Anchor in the Data Analysis and Probability Reporting Category (1.1). (*See below*)



NOTE: Below each specific descriptor of the Assessment Anchor is a reference in italics. This reference relates to the Pennsylvania Academic Standards and helps you cross-walk the Anchors to the Standards.

Eligible Content and Sample Items
Two other important features appear in this document:
Eligible Content. The column on the right-hand side of the page underneath each Assessment Anchor is the Eligible Content. This is often known as the "assessment limits" and helps teachers identify how the anchor will be assessed. Not all of the Eligible Content is assessed on the PSSA and the PSSA-M, but it shows the range of knowledge drawn upon to design the tests.

Sample Items. The sample items appear on the bottom half of the page. These are examples of how the Assessment Anchor might appear on the PSSA. Some of the pages may not have any sample items because only three items were created per Assessment Anchor.

For sample items specific to the PSSA-M, teachers should consult the *PSSA Modified Mathematics Item and Scoring Sampler* located on the state website.

PENNSYLVANIA DEPARTMENT OF EDUCATION

Overview of Mathematics Assessment Anchors

**Note that on this overview document, the grade level does not appear because these anchors occur at all Grade levels 4 through 8 and 11.*

MA. Numbers and Operations

MA.1 Demonstrate an understanding of numbers, ways of representing numbers, relationships among numbers and number systems.

MA.2 Understand the meanings of operations, use operations and understand how they relate to each other.

MA.3 Compute accurately and fluently and make reasonable estimates.

MB. Measurement

MB.1 Demonstrate an understanding of measurable attributes of objects and figures, and the units, systems and processes of measurement (not assessed at Grade 11).

MB.2 Apply appropriate techniques, tools and formulas to determine measurements.

MC. Geometry

MC.1 Analyze characteristics and properties of two- and three- dimensional geometric shapes and demonstrate understanding of geometric relationships.

MC.2 Identify and/or apply concepts of transformations or symmetry (not assessed at Grades 6, 7 or 11).

MC.3 Locate points or describe relationships using the coordinate plane.

MD. Algebraic Concepts

MD.1 Demonstrate an understanding of patterns, relations and functions.

MD.2 Represent and/or analyze mathematical situations using numbers, symbols, words, tables and/or graphs.

MD.3 Analyze change in various contexts (not assessed at Grades 4 or 8).

MD.4 Describe or use models to represent quantitative relationships (not assessed at Grade 4, 5, 6 or 7).

ME. Data Analysis and Probability

ME.1 Formulate or answer questions that can be addressed with data and/or organize, display, interpret or analyze data.

ME.2 Select and/or use appropriate statistical methods to analyze data.

ME.3 Understand and/or apply basic concepts of probability or outcomes.

ME.4 Develop and/or evaluate inferences and predictions or draw conclusions based on data or data displays (not assessed at Grades 4, 5 or 6).

Appendix B:

PSSA and PSSA-M General Scoring Guidelines

PENNSYLVANIA DEPARTMENT OF EDUCATION

PSSA*

General Description of Mathematics Scoring Guidelines

4 – The response demonstrates a *thorough* understanding of the mathematical concepts and procedures required by the task.

The response provides correct answer(s) with clear and complete mathematical procedures shown and a correct explanation, as required by the task. Response may contain a minor "blemish" (e.g., missing $) or omission in work or explanation that does not detract from demonstrating a *thorough* understanding.

3 – The response demonstrates a *general* understanding of the mathematical concepts and procedures required by the task.

The response and explanation (as required by the task) are mostly complete and correct. The response may have minor errors or omissions that do not detract from demonstrating a *general* understanding.

2 – The response demonstrates a *partial* understanding of the mathematical concepts and procedures required by the task.

The response is somewhat correct with *partial* understanding of the required mathematical concepts and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

1 – The response demonstrates a *minimal* understanding of the mathematical concepts and procedures required by the task.

0 – The response has no correct answer and *insufficient* evidence to demonstrate any understanding of the mathematical concepts and procedures required by the task for that grade level.

Response may show only information copied from the question.

Special Categories within zero reported separately:
BLK (blank) ...Blank, entirely erased, or written refusal to respond
OT .....Off task
IL.....Illegible
LOE.....Response in a language other than English

This document is available on the PDE website at http://www.education.state.pa.us.

*Note: The PSSA General Scoring Guidelines for Mathematics also apply to the PSSA-M mathematics tests.

Appendix C:

2010 Modified PSSA Tally Sheets

Grade 4

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	1	
A.1.1.1	2	1	1	1	
A.1.1.2	2	2	2	0	
A.1.1.3	1	1	2	1	
A.1.1.4	3	1	2	1	
A.1.2.1	2	1	2	1	
A.1.2.2	0	1	1	1	
A.1.3.1	4	1	2	1	
A.1.3.2	2	1	2	1	
A.2	1	1	1	0	
A.2.1.1	2	0	0	0	
A.2.1.2	3	2	2	1	
A.3	0	0	1	0	
A.3.1.1	2	1	2	1	
A.3.1.2	1	1	2	1	
A.3.1.3	2	0	0	1	
A.3.2.1	1	0	0	0	non-calculator only
A.3.2.2	1	1	2	1	
B.1	0	1	1	0	
B.1.1.1	0	0	1	1	
B.1.1.2	2	0	1	1	
B.1.1.3	1	0	1	1	
B.1.1.4	2	0	0	1	
B.2	1	0	0	0	
B.2.1.1	2	0	1	1	
B.2.2.1	0	1	2	0	
C.1	0	0	0	1	
C.1.1.1	2	1	1	0	
C.1.1.2	1	0	0	1	
C.1.2.1	2	1	1	0	
C.1.2.2	1	1	2	0	
C.2	0	0	0	0	
C.2.1.1	3	1	2	0	
C.3	0	0	1	0	
C.3.1.1	1	1	1	0	
D.1	1	0	1	0	
D.1.1.1	1	1	1	0	

G4 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	16	22%	9	24%
A.2	9	13%	6	16%
A.3	7	10%	3	8%
B.1	5	7%	4	11%
B.2	6	8%	1	3%
C.1	6	8%	3	8%
C.2	3	4%	1	3%
C.3	1	1%	1	3%
D.1	8	11%	3	8%
D.2	2	3%	2	5%
E.1	7	10%	4	11%
E.3	2	3%	1	3%

G4 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	32	44%	18	47%
B	11	15%	5	13%
C	10	14%	5	13%
D	10	14%	5	13%
E	9	13%	5	13%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.1.1.2	1	1	1	0	
D.1.1.3	0	1	1	0	
D.1.2.1	1	0	0	0	
D.1.2.2	1	0	1	0	
D.2	0	0	0	1	
D.2.1.1	1	0	0	1	
D.2.2.1	0	1	1	0	
D.2.2.2	1	1	1	0	
E.1	0	0	1	0	
E.1.1.1	4	2	2	1	
E.1.2.1	1	1	1	1	
E.1.2.2	2	1	2	1	
E.3	0	0	0	0	
E.3.1.1	2	1	2	2	
Totals	**63**	**32**	**54**	**26**	

Grade 5

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	1	0	
A.1.1.1	2	1	2	1	
A.1.2.1	2	1	3	0	
A.1.2.2	1	1	1	1	
A.1.3.1	0	1	1	1	
A.1.3.2	1	0	0	1	
A.1.3.3	3	1	2	0	
A.1.4.1	1	1	2	0	
A.1.4.2	1	1	1	1	
A.1.5.1	2	2	2	1	
A.1.6.1	2	1	1	1	
A.1.6.2	1	1	1	1	
A.2	1	1	1	1	
A.2.1.1	3	0	0	1	
A.2.1.2	1	0	1	1	
A.2.1.3	2	0	0	1	
A.3	0	0	0	0	
A.3.1.1	3	1	3	1	
A.3.1.2	2	1	1	0	
A.3.2.1	1	0	0	0	non-calculator only
B.1	1	0	0	0	
B.1.1.1	1	1	1	0	
B.1.2.1	1	1	1	0	
B.1.2.2	0	0	0	2	
B.1.3.1	0	0	1	0	
B.1.3.2	1	0	0	1	
B.2	0	0	1	0	
B.2.1.1	1	1	1	0	
B.2.2.1	1	0	0	1	
B.2.2.2	1	1	1	1	
B.2.2.3	1	1	1	0	
C.1	1	0	1	0	
C.1.1.1	1	1	2	0	
C.1.1.2	1	0	0	1	
C.1.2.1	1	1	1	0	
C.2	0	0	0	1	

G5 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	16	22%	11	29%
A.2	10	14%	4	11%
A.3	6	8%	2	5%
B.1	7	10%	2	5%
B.2	4	6%	3	8%
C.1	7	10%	2	5%
C.2	3	4%	3	8%
D.1	5	7%	4	11%
D.2	5	7%	2	5%
E.1	0	0%	1	3%
E.2	6	8%	2	5%
E.3	3	4%	2	5%

G5 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	32	44%	17	45%
B	11	15%	5	13%
C	10	14%	5	13%
D	10	14%	6	16%
E	9	13%	5	13%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
C.2.1.1	1	2	2	0	
C.2.1.2	2	1	2	0	
D.1	0	1	1	0	
D.1.1.1	1	0	2	0	
D.1.1.2	2	0	1	1	
D.1.2.1	2	0	1	1	
D.2	0	0	0	0	
D.2.1.1	1	1	2	1	
D.2.1.2	4	1	1	2	
E.1	0	0	1	0	
E.1.1.1	0	1	1	0	
E.2	0	0	0	1	
E.2.1.1	4	1	1	0	
E.2.1.2	2	1	2	0	
E.3	0	0	0	0	
E.3.1.1	1	1	2	1	
E.3.1.2	2	1	1	0	
Totals	**63**	**32**	**53**	**27**	

Grade 6

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	1	0	
A.1.1.1	1	1	1	1	
A.1.1.2	2	1	1	1	
A.1.1.3	1	1	1	1	
A.1.1.4	2	1	1	1	
A.1.2.1	2	0	1	1	
A.1.3.1	1	1	2	0	
A.1.3.2	2	1	2	1	
A.1.3.3	0	0	1	0	
A.1.4.1	2	0	1	1	
A.2	1	0	0	1	
A.2.1.1	2	2	2	0	
A.3	0	1	1	0	
A.3.1.1	2	0	0	0	non-calculator only
A.3.2.1	1	0	1	1	
B.1	0	0	1	0	
B.1.1.1	3	1	2	1	
B.2	0	0	0	1	
B.2.1.1	1	1	1	0	
B.2.1.2	0	0	1	0	
B.2.1.3	2	1	1	0	
B.2.2.1	2	1	1	0	
B.2.3.1	1	1	1	0	
C.1	1	0	0	0	
C.1.1.1	1	1	1	1	
C.1.1.2	1	2	2	0	
C.1.1.3	1	1	1	1	
C.1.1.4	1	1	1	0	
C.1.2.1	2	1	1	1	
C.1.2.2	1	0	1	1	
C.3	0	0	1	0	
C.3.1.1	3	1	2	1	
D.1	0	0	0	0	
D.1.1.1	1	1	1	1	
D.1.2.1	4	2	2	1	
D.2	0	0	1	1	

G6 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	13	18%	6	16%
A.2	6	8%	2	5%
A.3	3	4%	4	11%
B.1	3	4%	1	3%
B.2	6	8%	4	11%
C.1	11	15%	6	16%
C.3	3	4%	1	3%
D.1	5	7%	3	8%
D.2	8	11%	4	11%
E.1	7	10%	1	3%
E.2	2	3%	5	13%
E.3	5	7%	1	3%

G6 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	22	31%	12	32%
B	9	13%	5	13%
C	14	19%	7	18%
D	13	18%	7	18%
E	14	19%	7	18%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.2.1.1	3	1	2	1	
D.2.1.2	2	2	3	0	
D.2.2.1	3	1	1	1	
E.1	1	0	0	0	
E.1.1.1	1	1	2	1	
E.1.1.2	1	0	1	1	
E.1.1.3	1	0	1	2	
E.2	0	1	1	0	
E.2.1.1	2	1	2	1	
E.3	0	0	0	0	
E.3.1.1	3	0	1	1	
E.3.1.2	2	1	2	0	
Totals	**63**	**32**	**53**	**27**	

Grade 7

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	0	
A.1.1.1	2	1	1	1	
A.1.2.1	2	2	2	1	
A.1.2.2	0	0	1	0	
A.2	0	1	1	0	
A.2.1.1	1	0	1	1	
A.2.2.1	1	0	1	0	
A.2.2.2	1	0	1	0	
A.2.2.3	1	0	0	2	
A.2.2.4	1	0	0	1	
A.2.2.5	2	0	1	0	
A.2.2.6	1	1	1	0	
A.3	0	0	0	0	
A.3.1.1	2	0	0	0	non-calculator only
A.3.2.1	1	0	1	1	
A.3.2.2	0	1	1	1	
B.1	0	0	0	1	
B.1.1.1	2	1	1	0	
B.2	0	0	1	0	
B.2.1.1	1	0	1	1	
B.2.1.2	1	1	2	0	
B.2.1.3	2	2	2	0	
B.2.2.1	3	1	1	1	
B.2.2.2	1	0	0	0	
C.1	1	0	0	1	
C.1.1.1	1	1	1	1	
C.1.1.2	1	1	2	0	
C.1.1.3	1	1	1	0	
C.1.2.1	1	1	1	1	
C.1.2.2	1	1	2	0	
C.3	0	0	1	0	
C.3.1.1	3	1	1	0	
C.3.1.2	2	1	1	1	
D.1	0	0	0	0	
D.1.1.1	4	3	3	1	
D.2	1	0	1	0	

G7 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	4	6%	3	8%
A.2	8	11%	5	13%
A.3	3	4%	1	3%
B.1	2	3%	1	3%
B.2	8	11%	4	11%
C.1	9	13%	5	13%
C.3	5	7%	2	5%
D.1	4	6%	3	8%
D.2	9	13%	3	8%
D.3	6	8%	5	13%
E.1	2	3%	1	3%
E.2	3	4%	3	8%
E.3	7	10%	1	3%
E.4	2	3%	1	3%

G7 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	15	21%	9	24%
B	10	14%	5	13%
C	14	19%	7	18%
D	19	26%	11	29%
E	14	19%	6	16%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.2.1.1	2	1	2	1	
D.2.1.2	2	1	2	1	
D.2.2.1	1	1	3	1	
D.3	0	1	1	0	
D.3.1.1	3	1	1	3	
D.3.1.2	3	0	1	1	
E.1	0	0	0	0	
E.1.1.1	2	1	1	1	
E.2	0	0	0	1	
E.2.1.1	2	2	2	0	
E.2.1.2	1	1	1	0	
E.3	1	0	0	0	
E.3.1.1	1	1	2	0	
E.3.1.2	1	0	1	1	
E.3.1.3	1	0	1	0	
E.4	0	0	1	0	
E.4.1.1	2	1	1	1	
Totals	**63**	**32**	**54**	**27**	

Grade 8

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	0	
A.1.1.1	2	1	1	2	
A.1.1.2	0	1	1	0	
A.2	1	1	1	0	
A.2.1.1	0	1	1	1	
A.2.2.1	2	0	2	1	
A.2.2.2	1	0	2	0	
A.3	0	0	0	0	
A.3.1.1	1	0	1	1	
A.3.1.2	1	0	1	0	
A.3.2.1	2	0	0	0	non-calculator only
A.3.3.1	2	1	1	1	
B.1	0	0	1	0	
B.1.1.1	0	1	1	0	
B.1.1.2	2	1	1	0	
B.1.1.3	1	0	0	0	
B.1.1.4	1	0	0	1	
B.2	0	0	1	1	
B.2.1.1	1	1	1	0	
B.2.1.2	1	1	1	0	
B.2.1.3	0	0	0	0	
B.2.2.1	2	0	1	0	
B.2.2.2	1	1	1	0	
B.2.2.3	1	0	0	1	
C.1	0	0	0	1	
C.1.1.1	1	2	2	0	
C.1.1.2	2	1	2	1	
C.1.1.3	1	1	2	1	
C.1.2.1	3	2	2	0	
C.3	1	0	1	0	
C.3.1.1	3	1	1	1	
D.1	1	0	0	0	
D.1.1.1	1	2	2	1	
D.1.1.2	1	0	1	1	
D.1.1.3	1	1	1	1	
D.2	0	1	1	0	

G8 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	2	3%	2	5%
A.2	7	10%	5	13%
A.3	6	8%	1	3%
B.1	4	6%	2	5%
B.2	6	8%	3	8%
C.1	7	10%	6	16%
C.3	7	10%	1	3%
D.1	7	10%	3	8%
D.2	6	8%	6	16%
D.4	6	8%	2	5%
E.1	7	10%	4	11%
E.3	3	4%	1	3%
E.4	4	6%	2	5%

G11 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	15	21%	8	21%
B	10	14%	5	13%
C	14	19%	7	18%
D	19	26%	11	29%
E	14	19%	7	18%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.2.1.1	2	0	1	1	
D.2.1.2	0	0	1	1	
D.2.1.3	1	2	2	0	
D.2.2.1	1	0	1	2	
D.2.2.2	2	0	1	1	
D.4	0	0	0	0	
D.4.1.1	2	1	1	1	
D.4.1.2	2	1	1	1	
D.4.1.3	2	0	2	0	
E.1	0	0	1	0	
E.1.1.1	3	1	1	0	
E.1.1.2	2	1	1	1	
E.1.1.3	2	2	2	0	
E.3	0	0	0	1	
E.3.1.1	2	1	1	0	
E.3.2.1	1	0	1	1	
E.4	0	0	0	0	
E.4.1.1	1	1	1	0	
E.4.1.2	3	1	2	1	
Totals	**63**	**32**	**54**	**27**	

Grade 11

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	1	0	0	0	
A.1.1.1	0	1	1	0	
A.1.1.2	1	0	0	0	
A.1.1.3	0	0	0	1	
A.1.2.1	0	0	1	0	
A.1.3.1	1	1	1	0	
A.1.3.2	0	0	0	0	
A.2	0	0	1	0	
A.2.1.1	1	1	1	0	
A.2.1.2	0	1	1	0	
A.2.1.3	0	0	0	1	
A.2.2.1	1	0	0	1	
A.2.2.2	0	0	0	0	
A.3	0	0	0	0	
A.3.1.1	1	1	1	1	
A.3.2.1	2	0	0	0	non-calculator only
B.2	0	1	1	0	
B.2.1.1	1	1	1	1	
B.2.2.1	2	0	1	0	
B.2.2.2	1	0	0	1	
B.2.2.3	2	0	1	0	
B.2.2.4	2	0	1	1	
B.2.3.1	1	0	2	1	
C.1	0	0	1	0	
C.1.1.1	1	1	1	0	
C.1.1.2	2	1	1	0	
C.1.2.1	1	0	1	0	
C.1.2.2	0	0	0	1	
C.1.2.3	2	1	1	0	
C.1.3.1	1	1	1	0	
C.1.4.1	2	1	1	0	
C.3	0	0	0	1	
C.3.1.1	1	1	1	0	
C.3.1.2	1	0	1	1	
D.1	0	1	1	0	
D.1.1.1	2	1	1	1	

G11 Anchor Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A.1	6	8%	2	5%
A.2	2	3%	2	5%
A.3	3	4%	1	3%
B.2	9	13%	5	13%
C.1	9	13%	5	13%
C.3	2	3%	1	3%
D.1	5	7%	5	13%
D.2	14	19%	5	13%
D.3	7	10%	4	11%
D.4	3	4%	2	5%
E.1	1	1%	1	3%
E.2	3	4%	2	5%
E.3	2	3%	0	0%
E.4	6	8%	3	8%

G11 Reporting Category Summary (by points)

	2010 PSSA Core		2010 PSSA-M Core	
A	11	15%	5	13%
B	9	13%	5	13%
C	11	15%	6	16%
D	29	40%	16	42%
E	12	17%	6	16%

Anchor or Eligible Content	2010 PSSA Core	2010 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.1.1.2	1	0	1	1	
D.1.1.3	2	0	0	1	
D.2	1	0	0	1	
D.2.1.1	1	0	1	0	
D.2.1.2	1	0	1	1	
D.2.1.3	1	1	1	0	
D.2.1.4	1	1	2	1	
D.2.1.5	1	1	1	0	
D.2.2.1	3	1	2	0	
D.2.2.2	2	1	1	1	
D.2.2.3	0	0	1	1	
D.3	0	0	1	0	
D.3.1.1	1	1	1	1	
D.3.1.2	1	1	1	0	
D.3.2.1	2	1	1	1	
D.3.2.2	1	1	1	1	
D.3.2.3	2	0	2	0	
D.4	0	0	0	0	
D.4.1.1	3	2	2	2	
E.1	0	0	1	0	
E.1.1.1	1	0	0	0	
E.1.1.2	0	1	1	0	
E.2	0	0	0	0	
E.2.1.1	1	1	1	0	
E.2.1.2	0	1	1	1	
E.2.1.3	2	0	0	0	
E.3	0	0	0	1	
E.3.1.1	1	0	0	0	
E.3.1.2	0	0	1	1	
E.3.2.1	1	0	1	0	
E.4	1	0	0	0	
E.4.1.1	0	1	1	0	
E.4.1.2	1	0	0	0	
E.4.2.1	0	1	1	0	
E.4.2.2	1	1	1	0	
Totals	**62**	**32**	**54**	**27**	

Appendix D:

Item and Test Development Process

Guidelines for Item Revision and Enhancement

The following guidelines were used in the development of the modified items. The process used to implement these guidelines is described in the chart that follows this section.

Overview

The PSSA-M will be developed to facilitate students' ability to demonstrate their grade-level content knowledge and skills, as specified in the Pennsylvania Academic Assessment Anchor Content Standards as defined by the Eligible Content. The assessment tasks (items and graphics/stimuli) will be designed with the goal (revised and/or enhanced) to minimize or remove the effects of processing (e.g., cognitive, linguistic) or physical challenges related to students' disabilities without significant alteration of the assessed construct. Therefore, the PSSA-M design considers the particular needs of students eligible for this assessment in order to increase their access to the assessed content—appropriate access to test content is necessary to ensure the validity of the assessment results. Lack of access could result in the measurement of sources of variance that are not related to the intended test content (*construct irrelevance*) or could allow construct-irrelevant factors to interfere with the student's ability to fully demonstrate what he or she knows and can do, and subsequently the test results could underestimate the student's actual level of achievement. Therefore, for the initial field test (spring 2009), PSSA-M assessment items, tasks, etc. will be revised and/or enhanced to maintain the integrity of the grade-level content; however, revisions and/or enhancements will be purposefully and necessarily made in the operationalization of the grade-level content in order to address the specific access needs of the students who will be eligible for the PSSA-M.

Three main areas of consideration will affect the initial revision and/or enhancement process involved in the PSSA-M items: student characteristics, assessed content, and item format. Although each of these areas is discussed separately below, the areas interact and have real implications for item revisions and/or enhancements.

Student Characteristics

Students who will be eligible for the PSSA-M generally have difficulty processing information (e.g., working memory limitations, attention deficits). Therefore, reflected in the item revisions and/or enhancements will be methods for (1) appropriately reducing the cognitive load (e.g., amount and complexity of information), (2) appropriately reducing language load (i.e., construct-irrelevant language) of the assessed content, and/or (3) supporting students' processing of information (e.g., by segmenting or chunking information or by providing graphics that support understanding) in order to address their access needs and increase the validity of assessment results for these students.

Assessed Content

Given the capabilities and limitations of students eligible for the PSSA-M, some grade-level content may be less accessible to these students. For example, the ability to infer and to make connections among multiple pieces of information is a common challenge for learning disabled (LD) students in this population. Therefore, reflected in each item will be specific parameters for content that ensure it (1) is appropriate for the student population, (2) is consistent with the intention of the grade-level Assessment Anchor Content Standard as defined by the Eligible

Content, and (3) adequately represents the breadth and depth of the Assessment Anchor Content Standard as defined by the Eligible Content (i.e., does not under-represent the targeted construct). This may well mean that some of the Eligible Content may be simplified and/or eliminated.

Note: The initial phases of PSSA-M item revisions and/or enhancement of the items will rely primarily on expert judgment (e.g., PDE content-area experts and special educators; Pennsylvania content-area experts and special education experts; and additional content-area experts and special education experts from WestEd and DRC). Expert judgment will be supplemented with PDE's analyses of the 2009 student performance data (e.g., p-values, point biserials, and omission rates). In addition, Cognitive Interviews will also be conducted prior to the spring 2009 field test. Additional data will also be collected after the spring 2009 field test to further validate the design and revision and/or enhancement of the PSSA-M items.

Item Format

Item formats involve consideration of the degree to which the item format could (1) reliably measure the student's knowledge/skill, (2) yield an accurate measure of the student's knowledge/skill, and (3) have embedded the type of support or enhancement (e.g., graphic, context clues, range of permissible ways the student can process—reception and/or production—the assessed content) the student needs to access and demonstrate understanding of the assessed content. Item format considerations are as follows:

Font (Typeface)

- Introducing bolding, underlining, and other text changes (font size, italics, etc.) if item validity and construct alignment are not affected.
- Adding more space between letters and words if item validity is not affected.

Item Layout

- Adding more white space between items or having fewer items per page, when appropriate.
- Increasing the width of an item or line length (two column to one, single column layout), when appropriate.
- Restructuring the stem of an item into a "stacked" format. (Indenting stacked facts may be also be used.)
- Inserting bullets to organize complex information or inserting bullets to break complex text within an item stem into smaller parts.

Scaffolding

- For reading, segmenting Passages/Prompts/Scenarios (For example, students are provided the same passage/prompt/scenario as the general education PSSA at a given grade level, but the passage is "segmented" or divided into meaningful parts. Those items that apply directly to each segment would appear right after or adjacent to the referenced section of the text. In other words, questions would follow an order that parallels how information generally appears in the passage, prompt, and/or scenario. For reading, inferential questions, such as author's purpose or theme, would appear at the end after the entire passage had been read.

- Other types of scaffolding include, but are not limited to, the following:
    - Adding helpful hints or thought boxes (visual cues) to provide further definition of words and terminology and/or to support the text or emphasize main ideas.
    - Providing support or scaffolding for the number of steps and/or operations in a multi-step item such as adding sub-questions or steps to break up or help students think through multi-step problems/items.
    - Adding additional directions to explain a process or activity.
    - Adding pre-reading information to clarify the purpose of a passage, prompt, or scenario, such as the topic of the science scenario.
    - Embedding a formula (as appropriate for intention of the assessed standard).

General Guidelines for Revising and/or Enhancing PSSA-M Items

Guidelines for revising and/or enhancing PSSA-M items will include, but will not be limited to, those listed below. While many of these guidelines are common "best practices" and are included in guidelines for writing, reviewing, and revising items for the PSSA, further revisions and/or enhancements may apply.

Context

Context helps make language that is reflective of abstract/highly-generalized situations more concrete and relevant in order to ground the content being tested. Context that facilitates access includes the following:

- Concrete language
- Illustrative language
- Illustration/graphic

Graphics: Best Practices for the PSSA and the PSSA-M (Note: With graphics, the visual discrimination and visual processing challenges of students are considered).

- Graphic and labeling/naming conventions should be consistent.
- Graphics should support students' understanding of assessed content.
- Graphics should clarify (1) key aspects of the content/construct assessed and/or (2) what the student is expected to do (graphics used should be purposeful).
- Graphics should support context without requiring additional language (and may reinforce what is in the text of the item).
- Graphics should help students shift from one context to another within an assessment (e.g., from one type of item to another).
- Graphics should allow students to verify understanding of key elements of the text of the item.
- Graphics should allow representation of key elements of the problem (necessary information; construct-relevant) so that this information does not need to be presented in words.

  Consideration: How central is the information in the graphic to the construct? For example, if the graphic helps clarify construct-irrelevant information, then it may not be necessary—perhaps it would be better to alter the construct-irrelevant information. But, if the graphic helps to clarify the context or content that is construct-relevant or

an operation related to the construct, then it may be necessary; otherwise, the graphic may be misleading or distracting. Note: Certain graphics are required/assessed in mathematics and science.

Consideration: Can the graphic accurately represent the complexity of the problem in its totality? If not, then the graphic may be misleading.

- If the problem has a number of operations/steps, then it is important to simplify structures of the item (e.g., bulleted list with context or a graphic, diagram that accurately reflects the problem in its totality).
- Graphics should allow for reduction of language and/or complexity of language.
- A graphic needs to be consistent with the key elements of the item.
- Intervals (e.g., on number lines) should be consistent/equal.

Graphics: Additional Considerations for the PSSA-M

- Adding graphic organizers as enhancements: Graphic organizers (e.g., Venn diagram for compare and contrast, timelines, story maps).
- Altering a graphic or adding or expanding a graphic to duplicate text-described context (e.g., the stem in the unaltered item may refer to the weight of a car; for the altered version, a graphic showing a car with the weight written on or near it may be included. The graphic should reinforce or clarify the text, not replace it. The text should be removed and replaced with a graphic only in exceptionally rare and unique instances.).
- Adding a graphic to illustrate a term.
- Adding a support that provides a visual representation for helping students determine a solution to a problem (adding a blank grid or a blank number chart).

Item Sentence Structure: *Best Practices for the PSSA and the PSSA-M* (Note: The closed stem format is preferable to the open-stem format as the closed stem helps to reduce the retention load of content for the student as the student formulates the answer to a given question.)

- Referents should be clear; noun-pronoun relationships should be clear; antecedent references should be clear.
- Grammatical structures should be clear. Typically,
    - past or future-tense verb forms are changed to present tense,
    - passive verb forms are changed to active verb forms,
    - complex structures are changed to subject-verb-object structures,
    - long nominals/names/phrases are shortened (e.g., "last year's class vice president" becomes "a student leader"),
    - compound sentences are replaced with two separate sentences, especially in comparative structures,
    - long prepositional phrases are reduced or removed,
    - conditional clauses are replaced with separate sentences or the ordering of clauses within a sentence is changed for clarity, and
    - relative clauses are removed or rephrased for clarity.
- Questions framed in negative terms are rephrased.
- Changing tense may help remove passive-voice construction.

- Identifying the agent (e.g., proper noun) helps remove passive voice constructions.
- The verb should follow the subject (subject and verb should be adjacent to each other) — use common construction.
- One sentence per idea for each complex item helps reduce inappropriate complexity of sentence structure (e.g., could use bulleted lists).
- Introductory phrases are removed (e.g., last week)—unless necessary for the item.
- Key information is presented up front (first/early in item) and typically in simple sentence structure.
- Proper nouns should be ones that are familiar to students.
- Complexity of sentence structure should be at or below grade level (depends on intention of assessed standard).
- Traditional constructions should be used—e.g., _'s for possessive; _s or _es for plural.

Vocabulary/Wording: *Best Practices for the PSSA and the PSSA-M*

Use words/phrases consistently within the context of the item—(also consider consistency within a strand—e.g., reading, measurement).

- Support with context-familiar content-based abbreviations; make explicit connections between terms/abbreviations.
- Avoid words that are both nouns and verbs (e.g., race, value, cost); however, if a choice needs to be made, then the tendency is to use the word as a noun.
- Avoid hyphenated and compound words.
  Consideration: Balance the amount and complexity of language with the amount of information necessary for the student to understand/access the item (economy of language with meaning—purposeful use of language).
- Relative pronouns (e.g., which) should have a referent (e.g., which expression, which adjective) Note: This is preferable, but may not always be possible for a given content area or at a given grade level within a content area.
- Use construct-irrelevant vocabulary/phrases are at or below grade level.

Vocabulary/Wording: *Additional Considerations for the PSSA-M*

Repeat key words/phrases needed by the student to understand and respond to the item—providing synonyms for a key word may not always be helpful, given length and/or context of item; sometimes repeating the same key word is more appropriate (keep in mind the difference between instructional and assessment settings).

The following is a chronological description of the steps involved in the development process.

Item and Test Development Process for the PSSA-M

Step	Description
1. Create and Review Guiding Documentation	Item and test development specialists meet internally to review all guiding documentation related to the PSSA and PSSA-M. Documentation reviewed includes the test design blueprints, the Pennsylvania Assessment Anchors and Eligible Content, the test item specifications, and all test content descriptions. In addition, the test style specifications (style guide) are updated with new styles and formats specifically designed for the PSSA-M assessment.
2. Meet with PDE to Confirm Understanding of Program	The goal of the meeting each year is to ensure that item and test development teams have a clear understanding of PDE's vision for test development. A successful development cycle requires a clear understanding of Pennsylvania's content-area test specifications and of any unique interpretations of the Pennsylvania Assessment Anchors (if any).
3. Create Preliminary Test Item Development Plan	Item and test development specialists generate a preliminary development plan which includes an overview of the program, the internal and external (PDE) review and approval processes, a projected schedule for the modification of test items—including the number of test items to be modified for review by PDE and subsequent review by the committees of Pennsylvania educators.
4. Meet with PDE to Finalize Test Item Development Plan	Over the course of the meeting, item and test development specialists verify all steps in the development process including timelines and schedules for item modifications and test development.
5. Analyze Item Bank	Existing test items in the current PSSA Item Bank are reviewed as potential candidates for modification and enhancement. During this phase, test development specialists also make a tally of the modified item candidates in the PSSA-M item pool by assessment anchor.
6. Refine Modified Test Item Development Plan to Include Reviewers and Subcontractors	Item and test development specialists identify the item reviewers who will modify the test items (test development specialists or other professional item writers, subcontractors, etc.), the estimated number of item reviewers needed, the qualifications of the item reviewers, and the approximate number of modified test items to be submitted by each source.
7. Train Reviewers	Item and test development specialists train item reviewers, as needed. Item reviewers who have written for the PSSA in the past receive updated information concerning modification and style guidelines for the PSSA-M as needed.

Item and Test Development Process for the PSSA-M

Step	Description
8. Modify and Review Items	Test items are modified by item reviewers after training is complete, and feedback is provided by the item and test development specialists to item reviewers on a regular basis. As test items are modified, they are reviewed and edited in a series of internal reviews. Item and test development specialists review and edit items to include, but not limited to, the following: match to assessment anchor/eligible content, relevance to purpose, accuracy of content, item difficulty, interest level, grade appropriateness, depth of knowledge and cognitive complexity, adherence to the principles of universal design, and freedom from issues of bias/fairness/sensitivity. The items are also reviewed to ensure that the PSSA-M guidelines for enhancement and modifications have been met.
9. Enter Test Items into Database	Upon acceptance from item writers, test items are entered into the item management system, IDEAS (*Item Development and Educational Assessment System*). Item data stored in the system database includes, but is not limited to, the following: readability, cognitive level, estimated level of difficulty, alignment to assessment anchors, and correlation to stimulus.
10. Prepare Item Set for Sample Item Review by PDE	Item and test development specialists prepare a subset of the items for review by PDE.
11. PDE Conducts Sample Item Review	After a subset of the items is submitted to PDE for review, PDE reviews the items and provides feedback to item and test development teams via a conference call. Items are revised per PDE feedback.
12. Continue to Modify and Review Items	The remaining items are modified, and feedback is provided by the item and test development specialists to item reviewers on a regular basis. Items are entered into the item management system, IDEAS (*Item Development and Educational Assessment System*) (See step 8 and step 9).
13. Review Items Prior to Test Item Review and Validation Sessions	Prior to New Item Content Review, all items are submitted to PDE for review. Item and test development specialists incorporate all PDE feedback, and PDE-requested edits to items are made.
14. Prepare for Test Item Review Sessions (the New Item Content Review and the Bias, Fairness, and Sensitivity Review)	Item and test development specialists prepare all items and stimuli for review by the New Item Content Review Committee (consisting of Pennsylvania educators) and by the separate Bias, Fairness, and Sensitivity Committee (consisting of a panel of experts). Item and test development specialists also prepare training materials needed for training committee members to review items for content or for bias, fairness, and sensitivity issues. All training materials and other ancillary materials (e.g. agendas, presentations, etc.) are also developed and then submitted to PDE for review and approval. Invitations are also sent to Pennsylvania educators and national experts from PDE-approved committee lists.

Item and Test Development Process for the PSSA-M

Step	Description
15. Conduct Test Item Review Sessions (the New Item Content Review and the Bias, Fairness, and Sensitivity Review)	Committees of Pennsylvania educators and national experts review items in two meetings: one addressing item quality, the other addressing bias, fairness, and sensitivity. PDE, with support from item and test development specialists, present training on how to review new test items for content considerations or bias/fairness/sensitivity issues. At the New Item Content Review, suggested edits to test items are made and/or replacement test items are written during the actual item review so that both the committee and the PDE are able to observe changes to the test items and approve the test items during the committee review process. At the Bias, Fairness, and Sensitivity Review, experts in bias, fairness, and sensitivity review all test items and come to a consensus about any issues that are noted. At both meetings, the results are carefully documented.
16. Conduct Item Review Resolution and Cleanup	Following the conclusion of the New Item Content Review Committee meetings, PDE re-examines the consensus changes suggested by the committee members during the New Item Content Review Committee meetings. DRC item and test development specialists then record all of PDE's follow-up decisions and changes. During this cleanup process, PDE either accepts the changes as requested by the committee, or PDE rejects the decision of the committee. If a committee decision is rejected, PDE provides an alternate decision for DRC to implement. During this cleanup process, PDE also interprets the report from the Bias, Fairness, and Sensitivity Committee meetings and subsequently applies changes to test items. DRC item and test development specialists then apply the changes to the test items per PDE's decisions.
17. Construct Standalone Field Test Forms	DRC item and test development specialists select test items for inclusion on the standalone field test forms. Selections are based on recommendations from item review committees and their estimations of cognitive and difficulty levels, as well as input and recommendations from PDE. The items are arranged in three unique standalone field tests for each grade.
18. Submit Field Test Forms for Final Sign-Off	PDE-approved changes are applied to the items, stimuli, etc. (Changes reflect PDE's arbitration of the committee decisions.) Once all revisions to the items and/or to the art used by test items are completed, the field test forms are submitted to PDE for final review and sign-off.

Item and Test Development Process for the PSSA-M

Step	Description
19. Review Results of the Field Test	Following the administration of a field test form and the subsequent rangefinding and field test scoring processes for field test items, performance data for all field test items are analyzed by DRC psychometricians and test development specialists. Test item performance data that meet certain triggering criteria are flagged for additional reviews by test development specialists. Flagged field test items with extreme performance data are considered psychometrically unusable and are removed from future operational consideration. Normally, only field test items with marginal performance data are prepared for the Field Test Item Data Review meeting. However, since the PSSA-M program is in its initial stages, all of the items from the field test are eligible for review.
20. Prepare for Field Test Item Data Review	Test development specialists prepare all items and stimuli for review by the Field Test Item Data Review Committee (which consists of Pennsylvania educators). Psychometricians also prepare training materials needed for training committee members to review items for their performance. All training materials and other ancillary materials (e.g. agendas, presentations, etc.) are submitted to PDE for review and approval. Invitations are also sent to Pennsylvania educators from PDE-approved committee lists.
21. Conduct Field Test Item Data Review	Committees of Pennsylvania educators review the performance data of field test items. Psychometrians present training on how to review field test items based on their performance data. At the Item Data Review, committee members examine the performance of the items and determine whether the field test item is technically sound and appropriate for use on an operational PSSA-M test. Since test items cannot be modified at the Field Test Item Data Review, the committee can either accept an item as is or the committee can reject the item.
22. Conduct Field Test Item Data Review Reconciliation	Following the conclusion of the Field Test Item Data Review Committee meetings, PDE re-examines the consensus decisions (accept or reject) suggested by the committee members during the Field Test Item Data Review Committee meetings. Test development specialists record all of PDE's follow-up decisions and changes. During this cleanup process, PDE either accepts the decisions of the data review committee, or PDE rejects the decisions of the data review committee. If a committee decision is not accepted, PDE provides an alternate decision for test development specialists to implement. All PDE-approved changes to the test items status (accepted or rejected) are incorporated into the *Item Development and Educational Assessment System*, IDEAS.
23. Select Core Items for Operational Test Forms	After the results of the prior field test have been finalized following data review, test development specialists collaborate with psychometricians to follow the Test Design Blueprints and build requirements to make the initial selection of items for core positions in all test forms.

Item and Test Development Process for the PSSA-M

Step	Description
24. Review Core Selections	After test content and psychometric requirements have been achieved for core positions, the core items are provided to PDE for review and approval. Any changes to the content of the core requested by PDE are balanced with psychometric requirements until all core positions are approved by PDE, test development specialists, and psychometricians.
A new cycle of development has begun during this time, and steps 1–16 are repeated for a new set of items to be modified from the PSSA item pool. The newly modified items are guided through the same process as before in preparation for field testing.	
25. Construct Test Forms	Items and test components are assembled into forms using the form construction and typesetting function of DRC's *Item Development and Educational Assessment System*, IDEAS. Forms are reviewed internally for style and formatting requirements. New field test items (from the new cycle of development) are also selected to appear on the test forms for future operational use. As a result, three individual forms are created for each grade; each form contains the same core items, but a unique set of field test items.
26. Review Typeset Forms	After forms are constructed in IDEAS, draft hard copies of the forms are produced and presented to PDE for review and approval. Any changes to the content of the core requested by PDE are balanced with psychometric requirements until all core items are approved by PDE, test development specialists, and psychometricians. PDE also re-reviews all field test items appearing in the test forms. DRC applies changes to the field test items as required.
27. Print Test Forms	Following PDE's approval of the test forms, DRC completes a series of final proofing of all test forms. Final forms (along with ancillary materials) are then approved for printing.
28. Assemble Documentation of Test Materials	Metadata for each test item and form is documented and proofed, including: grade, form, session/section, item sequence, reporting category, assessment anchor, descriptor (sub-anchor), eligible content, number of points, item type, number of answer options, item usage, stimulus ID, etc.
The new field test items will then undergo the process described in steps 19–24 in preparation for use on the second operational test.	

Appendix E:

PSSA-M Item Review Cards

1. A bag has 10 marbles in it. The marbles are described below.

- 3 green marbles
- 3 blue marbles
- 4 white marbles

Lori selects 1 marble from the bag without looking.

What is the probability Lori selects a **white** marble?

- ○ $\frac{1}{10}$
- ○ $\frac{1}{4}$
- ○ $\frac{4}{10}$
- ○ $\frac{4}{6}$

PSSA-M Item Card

Item ID
[illegible]
Content Area
Mathematics
Passage ID
Passage Title
Grade
5
AACS Standards
E.3.1.2
Item Type
Multiple Choice
Points
1
Depth of Knowledge
2
Est Difficulty
Low
Key
C
Calculator
C
Focus
Probability

1. Anna has a toy animal collection on 3 shelves.

- 15 animals are on the 1st shelf.
- 21 animals are on the 2nd shelf.
- n animals are on the 3rd shelf.
- There is a total of 54 animals in Anna's collection.

A. Write an equation that can be used to find the number of animals (n) on the 3rd shelf.

Equation: _____

B. Solve the equation from part A to find the number of animals on the 3rd shelf. Show or explain all your work.

Show or explain your work here:

Number of animals on 3rd shelf: _____

PSSA-M Item Card

Item ID	
Content Area	Mathematics
Passage ID	
Passage Title	
Grade	5
AACS Standards	D.2
Item Type	Open Ended
Points	4
Depth of Knowledge	3
Est Difficulty	High
Calculator	C
Focus	Solve for missing nu

Kelly has a toy animal collection. Kelly's toy animal collection is larger than Anna's by **more than** 26 animals.

E. Write an inequality that can be used to show the number of animals (*a*) in Kelly's collection.

Inequality: _____

1. Simplify:

$$\frac{15 - 2^2 + (3 - 2)}{2}$$

(Hint: Remember to use order of operations.)

- ○ 4
- ○ 5
- ○ 10
- ○ 12

PSSA-M Data Card

Item ID
[illegible]
Content Area
Mathematics
Passage ID
Passage Title
Grade
8
Standards
AACS: A.2.1.1
Item Type
Multiple Choice
Points
1
Depth of Knowledge
1
Est Difficulty
Low
Calculator
Yes
Key
A
Focus
Order of operations

PSSA-M Data Card continued

Administration

Name	Use Function	Rptg Flag	Seq	Period	Year	Day	Session	Calc	Model/Ext	Grade
23_M	FT			Spring	2010		2	Yes		8

Traditional Statistics

N	P-Val	Mean	Item Total Corr
999	0.50		0.47

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Mean Sq	Fit
-5.0	-5.7	0.85	0.87				

IRT Statistics

Label	Final	Final S.E.	Preliminary	Preliminary S.E.
Location	0.24	0.07		

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
A*	0.50	0.47		
B	0.13	-0.34		
C	0.22	-0.45		
D	0.14	-0.44		
MULTS	0.00			
OMITS	0.01			

DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal

Appendix F:

Item Rating Sheet and Item Review Criteria Guidelines

Reviewer Signature: _____

Item Rating Sheet

Content Area: _____ **Grade:** _____

	Content Alignment	Rigor Level Alignment				Technical Design			Universal Design		STATUS
	Standards	Grade	Difficulty	Depth of Knowledge	Source of Challenge	Correct Answer	Distractors	Graphics	Language Demand	Bias	Acceptance Status
Unique ID number	—Higher —Lower —None	— Above —At — Below	—Hard —Medium —Easy	—Recall —Application —Strategic Thinking	—Yes —No	—Yes —No	—Yes —No	—Yes —No	—Yes —No	—Yes —No	— Approved as is — Accepted with suggested revisions — Dissenting View

Item Review Criteria Guidelines

The purpose of this form is to provide guidelines to the item review process in terms of item characteristics that are essential in building a fair and balanced assessment. Use these guidelines in conjunction with the Item Rating Sheet when recording your feedback on individual items.

Content Alignment		Options
Standards, Anchors, Eligible Content	Does the content of the item align with the Standard/Anchor/Eligible Content? Each item was written to assess a particular Standard/Anchor/ Eligible Content statement which is indicated on the individual Item Card. Consider the degree to which the item is, in fact, aligned with the indicated eligible content. In making this judgment, it is important to consider whether the **content** is aligned (e.g., do the eligible content and the item both deal with fractions) and whether the required **performance** is aligned (e.g., if the eligible content calls for a comparison to be made, is this reflected in the item).	**HIGHER**—Aligns to the higher level of the EC **LOWER**—Aligns to the lower level of the EC **NONE**—No alignment with EC

Rigor Level Alignment		Options
Grade	Is the item grade-level appropriate? Is the content consistent with the experiences of a student at the grade level assessed? Is the challenge level appropriate for the grade?	**ABOVE** Grade Level **AT** Grade Level **BELOW** Grade Level
Difficulty	Do you agree with the item's difficulty rating? Item Difficulty is indicated as Easy, Medium, and Hard. Is your rating in agreement with the difficulty rating on the Item Form?	**HARD** **MEDIUM** **EASY**
Depth of Knowledge	Depth of Knowledge is based on the alignment work of Norman Webb. Rate each item based on the cognitive demand, using the following levels: 1. Recall – ***Recall*** of a fact, information, or procedure. 2. Basic Application of Skill or Concept – ***Use*** of information, conceptual knowledge, procedures, two or more steps, etc. 3. Strategic Thinking – Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer. 4. Extended Thinking – Requires an investigation, time to think and process multiple conditions of the problem or task, and more than 10 minutes to do non-routine manipulations. (This level is generally not assessed in on-demand assessments.)	**4** = Extended Thinking **3** = Strategic Thinking **2** = Basic Application **1** = Recall

Source of Challenge	Is the source of challenge appropriately targeted to the content? The hardest part of the item (i.e., source of challenge) should be the content that is targeted. For example, in mathematics, the mathematics should be the major source of challenge rather than the wording or graphic. Students should not give an incorrect answer to a mathematics item because the reading level is too high or a graphic is flawed. Conversely, students should not give correct answers for reasons such as prior knowledge that make the answer to the question obvious (e.g., if the question asks which country has the largest population and students are to read a graph that includes China, there is no need to read the graph to answer the question).	**Y** = Yes **N** = No

Technical Design		**Options**
Correct Answer	Is there one clear, correct answer? There should be no other answer that "could" be correct. CAUTION: This does not mean that "good" distractors are unfair.	**Y** = Yes **N** = No
Distractors	Are distractors fair and appropriate? Distractors that are appropriate offer students reasonable choices that can be arrived at by making common errors. There should be no distractors that make no sense at all. It should be possible to examine each option and to reason how a student with some deficiency in knowledge or skill could choose it. The distractors should be formatted according to acceptable standards of test construction (e.g., a phrase that is common to each distractor should be placed in the stem).	**Y** = Yes **N** = No
Graphics	Are the graphics clear and accurate?	**Y** = Yes **N** = No

Universal Design		**Options**
Language Demand	Is language clear, well-formatted, and precise? Does the item use correct terminology for the content area? In order for all students to enter into the questions of the assessment, they must be able to understand them. If the items are formatted poorly, use unnecessarily complex words or phrases, or use figures or layouts that are difficult to understand, some students will give incorrect answers due to these factors rather than the content that is being assessed.	**Y** = Yes **N** = No
Bias	Is the item free of bias? All students will not be able to enter into the assessment if bias considerations are not resolved. Does the item contain clear bias problems? *A thorough, independent bias review* (separate from this meeting) *will be completed for all items.*	**Y** = Yes **N** = No

Status		**Options**
Acceptance Status	This is an overall judgment about the item. Based on the consensus of the committee, indicate whether the item was approved without revision to the content of the item or whether the item was accepted by the committee after revision of the content of the item. If there is a dissenting view (opposed to the committee consensus), record a brief explanation of the dissenting view on the back of the Item Rating Sheet.	**—Approved** as is **—Accepted** with suggested revisions **—Dissenting View**

NOTES:

- If you leave a box blank on the Item Rating Sheet, it will be recorded to indicate that you did not have any specific feedback for that item or issue.
- If you object to the consensus of the committee, please note this on the item rating sheet and then record a brief explanation of the dissenting view on the back of the Item Rating Sheet.
- **Do NOT remove any items from the item binder at any time.**
- You must sign your item rating sheet.

Appendix G:

2010 Test Book Section Layout Plans

2010 Modified Mathematics Test Book Section Layout for Grades 4 – 8 & 11

Mathematics Core

Core/common MC items	30
2 core 4 pt OE items	8
Total	38 points

The estimated testing time for mathematics is approximately 135-155 minutes (including embedded field test items). [Timing assumes 10 min per OE and 3 min per MC.]

Section	Content	Number of MC	MC Item Breakdown	Number of OE	OE Item Breakdown	Section Time (in minutes)
1	Mathematics	15	15–common (core) items	2	1–common (core) item 1–field test	60–70
2	Mathematics	23	15–common (core) items 8–field test items	1	1–common (core) item	75–85

Notes:

1) There are 3 forms.

2) The ruler items may fall in Sections 1 or 2.

3) All items in the PSSA-M mathematics test allow for calculator use.

Appendix H:

Mean Raw Scores by Form

	Form	N	L	Pts	Min	Max	Mean	Med	*SD*
4	0	2169	32	38	4	38	22.9	23.0	6.55
	1	735	32	38	4	38	22.8	23.0	6.53
	2	716	32	38	4	37	22.7	23.0	6.77
	3	718	32	38	4	38	23.1	23.0	6.36
5	0	2552	32	38	5	37	21.9	22.0	6.43
	1	881	32	38	6	37	22.2	22.0	6.47
	2	836	32	38	5	37	21.8	22.0	6.49
	3	835	32	38	6	37	21.6	21.0	6.32
6	0	2700	32	38	4	37	18.3	18.0	6.01
	1	919	32	38	4	37	18.2	18.0	6.17
	2	899	32	38	4	36	18.6	19.0	5.90
	3	882	32	38	4	34	18.1	18.0	5.94
7	0	2817	32	38	3	38	19.1	19.0	7.00
	1	947	32	38	3	37	19.0	19.0	6.92
	2	944	32	38	3	38	19.0	19.0	7.02
	3	926	32	38	3	38	19.2	19.0	7.06
8	0	3019	32	38	2	37	18.9	19.0	6.41
	1	1011	32	38	2	36	19.2	19.0	6.31
	2	1007	32	38	3	37	18.8	19.0	6.31
	3	1001	32	38	3	36	18.8	19.0	6.59
11	0	3536	32	38	0	37	17.0	16.0	6.27
	1	1179	32	38	3	35	17.1	17.0	6.39
	2	1161	32	38	0	37	16.9	17.0	6.13
	3	1196	32	38	0	36	16.9	16.0	6.30

Appendix I:

Item Statistics

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	4	561441	1	0	1	A.1.1.1	1	2169	0.85	0.00	0.21	-0.9424	0.0620	1.3	1.1	2.5	1.2
Math	4	560699	2	0	2	A.1.1.2	1	2169	0.68	0.00	0.34	0.1173	0.0494	-0.2	1.0	-0.5	1.0
Math	4	561480	3	0	3	A.1.1.2	1	2169	0.56	0.00	0.23	0.7251	0.0466	6.8	1.1	6.2	1.2
Math	4	561509	4	0	4	A.1.1.3	1	2169	0.46	0.00	0.32	1.2120	0.0464	1.7	1.0	2.5	1.1
Math	4	561478	5	0	5	A.1.1.4	1	2169	0.49	0.00	0.34	1.0536	0.0462	0.3	1.0	0.9	1.0
Math	4	561493	6	0	6	A.1.2.1	1	2169	0.38	0.00	0.29	1.5864	0.0474	1.5	1.0	4.3	1.1
Math	4	561515	7	0	7	A.1.2.2	1	2169	0.60	0.00	0.33	0.5382	0.0472	0.8	1.0	0.1	1.0
Math	4	561495	8	0	8	A.1.3.1	1	2169	0.48	0.00	0.32	1.1137	0.0463	1.9	1.0	2.3	1.1
Math	4	561517	9	0	9	A.1.3.2	1	2169	0.44	0.00	0.43	1.3200	0.0465	-4.9	0.9	-3.3	0.9
Math	4	561475	10	0	10	A.2.1.2	2	2169	0.80	0.01	0.36	-0.6429	0.0573	-1.8	0.9	-1.6	0.9
Math	4	560696	11	0	11	A.2.1.2	2	2169	0.72	0.00	0.39	-0.0594	0.0508	-2.5	0.9	-2.8	0.9
Math	4	561508	12	0	12	A.3.1.1	1	2169	0.55	0.00	0.35	0.7971	0.0465	0.3	1.0	0.0	1.0
Math	4	561471	13	0	13	A.3.1.2	1	2169	0.40	0.00	0.28	1.4911	0.0470	3.1	1.1	3.7	1.1
Math	4	561446	14	0	14	A.3.2.2	1	2169	0.59	0.00	0.29	0.5910	0.0470	3.3	1.1	2.7	1.1
Math	4	560700	15	0	15	B.2.2.1	1	2169	0.56	0.00	0.33	0.7319	0.0466	1.3	1.0	1.7	1.0
Math	4	560691	17	0	18	C.1.1.1	1	2169	0.60	0.00	0.30	0.5289	0.0472	2.2	1.0	2.5	1.1
Math	4	561702	18	0	19	C.1.2.1	1	2169	0.86	0.00	0.22	-1.0589	0.0642	1.0	1.0	2.0	1.2
Math	4	561760	19	0	20	C.1.2.2		2169	0.79	0.00	0.26	-0.5233	0.0557	1.0	1.0	1.2	1.1
Math	4	561875	20	0	21	C.2.1.1	1	2169	0.82	0.00	0.27	-0.7483	0.0589	0.5	1.0	1.2	1.1
Math	4	560701	21	0	22	C.3.1.1	1	2169	0.66	0.00	0.20	0.2294	0.0487	6.0	1.1	6.6	1.2
Math	4	560703	22	0	23	D.1.1.1	2	2169	0.93	0.00	0.25	-1.8584	0.0841	-0.6	1.0	-1.8	0.8
Math	4	561756	23	0	24	D.1.1.2	2	2169	0.67	0.00	0.38	0.1626	0.0491	-1.5	1.0	-2.2	0.9
Math	4	561693	24	0	25	D.1.1.3	2	2169	0.40	0.00	0.32	1.4888	0.0470	0.6	1.0	2.4	1.1
Math	4	561805	25	0	26	D.2.2.1	1	2169	0.83	0.00	0.39	-0.8069	0.0598	-3.0	0.9	-3.8	0.8
Math	4	561629	26	0	27	D.2.2.2	1	2169	0.92	0.00	0.31	-1.7437	0.0806	-1.2	0.9	-2.4	0.8
Math	4	561644	27	0	28	E.1.1.1	2	2169	0.60	0.00	0.44	0.5150	0.0472	-5.8	0.9	-5.2	0.9
Math	4	561848	28	0	29	E.1.1.1	1	2169	0.88	0.00	0.35	-1.3053	0.0693	-2.1	0.9	-2.8	0.8
Math	4	561878	29	0	30	E.1.2.1	1	2169	0.88	0.00	0.39	-1.3155	0.0695	-2.6	0.9	-3.8	0.7
Math	4	561893	30	0	31	E.1.2.2	1	2169	0.92	0.00	0.38	-1.7097	0.0796	-2.3	0.9	-4.6	0.6
Math	4	560690	31	0	32	E.3.1.1	2	2169	0.90	0.00	0.17	-1.4874	0.0736	0.7	1.0	1.5	1.1
Math	4	597214	34	1	33	A.1.1.1	1	735	0.54	0.00	0.36	0.8013	0.0796	-0.2	1.0	-0.6	1.0
Math	4	597383	35	1	34	A.2.1.2	2	735	0.52	0.00	0.43	0.8912	0.0794	-2.7	0.9	-2.1	0.9
Math	4	561763	36	1	35	B.1.1.2	1	735	0.41	0.00	0.15	1.4130	0.0803	5.7	1.2	5.5	1.3
Math	4	561795	37	1	36	B.1.1.4	2	735	0.42	0.00	0.31	1.3541	0.0801	0.9	1.0	0.9	1.0
Math	4	560693	38	1	37	B.2.1.1	1	735	0.96	0.00	0.24	-2.5828	0.1938	-0.4	0.9	-1.3	0.7
Math	4	597509	39	1	38	C.1.1.2	1	735	0.76	0.00	0.18	-0.3309	0.0912	2.4	1.1	3.0	1.3
Math	4	597407	40	1	39	E.1.1.1	2	735	0.81	0.00	0.39	-0.6574	0.0982	-1.5	0.9	-2.5	0.8
Math	4	561698	41	1	40	E.3.1.1	2	735	0.85	0.00	0.27	-0.9767	0.1070	0.0	1.0	-0.3	1.0
Math	4	561469	43	2	33	A.1.1.3	1	716	0.58	0.00	0.44	0.5990	0.0821	-2.4	0.9	-2.8	0.9

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	4	597529	44	2	34	A.1.2.2	1	716	0.79	0.00	0.43	-0.5644	0.0977	-2.1	0.9	-2.1	0.8
Math	4	561440	45	2	35	A.1.3.1	1	716	0.42	0.00	0.32	1.3776	0.0817	0.7	1.0	1.5	1.1
Math	4	597450	46	2	36	A.3.1.1	1	716	0.37	0.00	0.37	1.6048	0.0830	-0.9	1.0	0.3	1.0
Math	4	597533	47	2	37	A.3.1.3	1	716	0.23	0.00	0.29	2.4138	0.0939	-0.9	1.0	2.3	1.2
Math	4	561652	48	2	38	B.1.1.1	1	716	0.68	0.00	0.35	0.1122	0.0863	0.2	1.0	0.1	1.0
Math	4	561630	49	2	39	D.2.1.1	1	716	0.33	0.00	0.28	1.8491	0.0852	0.7	1.0	2.6	1.2
Math	4	561666	50	2	40	E.1.2.1	2	716	0.66	0.00	0.46	0.2167	0.0851	-2.8	0.9	-2.8	0.9
Math	4	597494	52	3	33	A.1.1.4	1	718	0.77	0.00	0.34	-0.3622	0.0940	-0.6	1.0	-1.4	0.9
Math	4	597612	53	3	34	A.1.2.1	1	718	0.58	0.00	0.39	0.6551	0.0812	-1.5	1.0	-0.9	1.0
Math	4	597558	54	3	35	A.1.3.2	2	718	0.34	0.00	0.36	1.8171	0.0839	-1.2	1.0	0.1	1.0
Math	4	597506	55	3	36	A.3.1.2	1	718	0.47	0.00	0.34	1.1743	0.0802	-0.1	1.0	0.2	1.0
Math	4	597570	56	3	37	A.3.2.2	1	718	0.88	0.00	0.38	-1.2723	0.1208	-1.5	0.9	-2.7	0.7
Math	4	561747	57	3	38	B.1.1.3	1	718	0.34	0.00	0.13	1.8459	0.0842	4.6	1.2	5.5	1.3
Math	4	561803	58	3	39	E.1.2.2	2	718	0.66	0.00	0.45	0.2959	0.0840	-3.4	0.9	-2.8	0.9
Math	4	561770	59	3	40	E.3.1.1	2	718	0.77	0.00	0.35	-0.3622	0.0940	-1.4	0.9	0.6	1.1
Math	5	561806	60	0	1	A.1.1.1		2551	0.72	0.00	0.38	-0.2596	0.0466	-2.9	0.9	-3.6	0.9
Math	5	561748	61	0	2	A.1.2.1		2551	0.66	0.00	0.22	0.1008	0.0442	4.7	1.1	5.4	1.2
Math	5	561692	62	0	3	A.1.2.2	1	2551	0.63	0.00	0.37	0.2170	0.0436	-2.6	1.0	-2.0	1.0
Math	5	561677	63	0	4	A.1.3.1		2551	0.69	0.00	0.37	-0.0906	0.0453	-3.2	0.9	-3.2	0.9
Math	5	561703	64	0	5	A.1.3.3		2551	0.28	0.00	0.24	1.9581	0.0468	2.3	1.1	5.6	1.2
Math	5	561757	65	0	6	A.1.4.1		2551	0.79	0.00	0.33	-0.6598	0.0507	-2.0	0.9	-2.5	0.9
Math	5	561879	66	0	7	A.1.4.2	1	2551	0.54	0.00	0.29	0.6908	0.0423	1.9	1.0	1.8	1.0
Math	5	561797	67	0	8	A.1.5.1		2551	0.75	0.00	0.27	-0.4297	0.0482	0.9	1.0	0.4	1.0
Math	5	561733	68	0	9	A.1.5.1		2551	0.98	0.00	0.17	-3.4765	0.1499	0.1	1.0	-2.1	0.6
Math	5	560706	69	0	10	A.1.6.1	1	2551	0.28	0.00	0.09	1.9718	0.0469	7.5	1.2	9.9	1.5
Math	5	561849	70	0	11	A.1.6.2		2551	0.63	0.00	0.35	0.2367	0.0435	-1.6	1.0	-2.2	1.0
Math	5	561883	71	0	12	A.3.1.1		2551	0.50	0.00	0.30	0.8437	0.0422	1.6	1.0	1.5	1.0
Math	5	560692	72	0	13	A.3.1.2		2551	0.30	0.00	0.26	1.8285	0.0458	1.4	1.0	3.8	1.1
Math	5	561681	73	0	14	B.1.1.1		2551	0.74	0.00	0.21	-0.3051	0.0470	2.9	1.1	2.5	1.1
Math	5	561920	74	0	15	B.1.2.1		2551	0.55	0.00	0.49	0.5934	0.0424	-9.9	0.9	-8.6	0.8
Math	5	561810	76	0	18	B.2.1.1		2551	0.66	0.00	0.24	0.1008	0.0442	3.4	1.1	3.1	1.1
Math	5	561723	77	0	19	B.2.2.2		2551	0.58	0.00	0.35	0.4688	0.0427	-1.8	1.0	-2.2	1.0
Math	5	560715	78	0	20	B.2.2.3		2551	0.82	0.00	0.38	-0.8497	0.0532	-3.4	0.9	-4.3	0.8
Math	5	561865	79	0	21	C.1.1.1		2551	0.87	0.00	0.24	-1.3173	0.0611	0.0	1.0	0.0	1.0
Math	5	561913	80	0	22	C.1.2.1	1	2551	0.61	0.00	0.25	0.3519	0.0431	3.8	1.1	4.1	1.1
Math	5	561832	81	0	23	C.2.1.1		2551	0.79	0.00	0.28	-0.6867	0.0511	-0.3	1.0	0.0	1.0
Math	5	561858	82	0	24	C.2.1.1		2551	0.67	0.00	0.34	0.0374	0.0445	-0.9	1.0	-1.2	1.0
Math	5	560705	83	0	25	C.2.1.2		2551	0.52	0.00	0.32	0.7338	0.0423	0.4	1.0	0.8	1.0
Math	5	561841	84	0	26	D.2.1.1		2551	0.83	0.00	0.38	-0.9803	0.0552	-2.7	0.9	-4.2	0.8

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	5	560704	85	0	27	D.2.1.2	1	2551	0.66	0.00	0.34	0.0641	0.0444	-1.0	1.0	-1.7	1.0
Math	5	561824	86	0	28	E.1.1.1		2551	0.91	0.00	0.26	-1.7559	0.0712	-1.2	0.9	-2.0	0.8
Math	5	561914	87	0	29	E.2.1.1	1	2551	0.38	0.00	0.34	1.4489	0.0435	-1.3	1.0	-1.0	1.0
Math	5	561713	88	0	30	E.2.1.2	1	2551	0.75	0.00	0.32	-0.4081	0.0480	-0.9	1.0	-1.0	1.0
Math	5	561707	89	0	31	E.3.1.1		2551	0.82	0.00	0.28	-0.8674	0.0535	-1.0	1.0	-0.3	1.0
Math	5	560697	90	0	32	E.3.1.2		2551	0.59	0.00	0.32	0.4402	0.0428	0.3	1.0	-0.2	1.0
Math	5	597864	93	1	33	A.1.1.1	1	880	0.71	0.00	0.28	-0.1335	0.0783	0.9	1.0	1.4	1.1
Math	5	597825	94	1	34	A.1.4.2	1	880	0.61	0.00	0.36	0.3834	0.0735	-0.9	1.0	-1.2	1.0
Math	5	597772	95	1	35	B.1.2.2	2	880	0.60	0.01	0.40	0.4108	0.0733	-2.5	0.9	-2.1	0.9
Math	5	597768	96	1	36	B.1.3.2	2	880	0.30	0.00	0.23	1.8883	0.0782	2.0	1.1	2.1	1.1
Math	5	561870	97	1	37	B.2.2.1		880	0.84	0.00	0.29	-0.9751	0.0949	-0.6	1.0	-0.7	0.9
Math	5	560708	98	1	38	C.1.1.2	1	880	0.40	0.00	0.16	1.3719	0.0734	5.4	1.2	5.0	1.2
Math	5	561790	99	1	39	D.1.1.2		880	0.36	0.00	0.20	1.5675	0.0748	3.2	1.1	3.9	1.2
Math	5	597257	100	1	40	D.2.1.2	2	880	0.54	0.00	0.42	0.6849	0.0722	-3.3	0.9	-3.0	0.9
Math	5	597837	102	2	33	A.1.2.2	1	836	0.70	0.00	0.32	-0.1584	0.0802	-0.6	1.0	0.0	1.0
Math	5	597866	103	2	34	A.1.3.2	1	836	0.11	0.00	0.13	3.2227	0.1145	0.5	1.0	2.7	1.4
Math	5	597863	104	2	35	A.1.6.2	1	836	0.59	0.01	0.48	0.4108	0.0750	-5.0	0.9	-4.6	0.8
Math	5	561736	105	2	36	A.2.1.1		836	0.43	0.00	0.33	1.1773	0.0745	-0.2	1.0	0.0	1.0
Math	5	597816	106	2	37	A.2.1.3	2	836	0.31	0.00	0.09	1.7921	0.0795	4.5	1.2	4.8	1.3
Math	5	561909	107	2	38	B.2.2.2		836	0.63	0.00	0.36	0.2368	0.0761	-1.3	1.0	-1.6	0.9
Math	5	560711	108	2	39	D.1.2.1		836	0.69	0.00	0.39	-0.0682	0.0791	-2.4	0.9	-1.8	0.9
Math	5	561884	109	2	40	E.3.1.1		836	0.79	0.00	0.28	-0.6681	0.0888	-0.3	1.0	0.4	1.0
Math	5	597829	111	3	33	A.1.3.1	1	835	0.74	0.00	0.30	-0.3980	0.0829	-0.4	1.0	0.4	1.0
Math	5	597855	112	3	34	A.1.5.1	2	835	0.32	0.00	0.31	1.6743	0.0785	-0.6	1.0	1.0	1.1
Math	5	597824	113	3	35	A.1.6.1	1	835	0.21	0.00	0.01	2.3200	0.0890	3.8	1.2	6.6	1.6
Math	5	597785	114	3	36	A.2.1.2	2	835	0.63	0.00	0.21	0.2189	0.0756	2.8	1.1	2.5	1.1
Math	5	597753	115	3	37	A.3.1.1	1	835	0.59	0.00	0.31	0.3906	0.0745	0.1	1.0	0.4	1.0
Math	5	561926	116	3	38	B.1.2.2	2	835	0.37	0.00	0.32	1.4138	0.0760	-0.8	1.0	0.9	1.0
Math	5	597517	117	3	39	D.2.1.1	1	835	0.53	0.00	0.42	0.6910	0.0736	-3.4	0.9	-3.4	0.9
Math	5	561902	118	3	40	D.2.1.2	1	835	0.28	0.00	0.24	1.8815	0.0812	0.8	1.0	2.0	1.1
Math	6	560709	119	0	1	A.1.1.1		2698	0.41	0.00	0.47	0.6270	0.0413	-9.9	0.9	-8.8	0.8
Math	6	561850	120	0	2	A.1.1.2		2698	0.38	0.00	0.24	0.7741	0.0418	3.0	1.1	4.2	1.1
Math	6	562078	121	0	3	A.1.1.3		2698	0.66	0.00	0.36	-0.5513	0.0428	-3.4	0.9	-2.8	0.9
Math	6	561904	122	0	4	A.1.1.4		2698	0.51	0.00	0.46	0.1957	0.0407	-9.9	0.9	-8.5	0.9
Math	6	561931	123	0	5	A.1.3.1		2698	0.52	0.00	0.21	0.1428	0.0407	5.4	1.1	5.1	1.1
Math	6	561905	124	0	6	A.1.3.2	1	2698	0.45	0.00	0.25	0.4502	0.0409	2.7	1.0	2.9	1.1
Math	6	561910	125	0	7	A.2.1.1		2698	0.81	0.00	0.25	-1.3507	0.0502	-0.5	1.0	-0.1	1.0
Math	6	561838	126	0	8	A.2.1.1		2698	0.62	0.00	0.34	-0.3168	0.0417	-2.3	1.0	-1.7	1.0
Math	6	561906	127	0	9	B.1.1.1		2698	0.36	0.00	0.22	0.8922	0.0423	3.3	1.1	4.0	1.1

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	6	561857	128	0	10	B.2.1.1	1	2698	0.50	0.00	0.23	0.2417	0.0407	4.4	1.1	4.3	1.1
Math	6	561915	129	0	11	B.2.1.3		2698	0.56	0.00	0.23	-0.0529	0.0409	4.4	1.1	3.6	1.1
Math	6	561938	130	0	12	B.2.2.1		2698	0.94	0.00	0.26	-2.7267	0.0804	-0.7	1.0	-3.0	0.7
Math	6	561951	131	0	13	B.2.3.1	1	2698	0.72	0.00	0.33	-0.8688	0.0451	-2.3	1.0	-1.7	1.0
Math	6	560718	132	0	14	C.1.1.1	1	2698	0.68	0.00	0.27	-0.5857	0.0430	0.0	1.0	-0.4	1.0
Math	6	560720	133	0	15	C.1.1.2		2698	0.63	0.00	0.29	-0.3656	0.0419	0.1	1.0	0.0	1.0
Math	6	561936	135	0	18	C.1.1.2		2698	0.36	0.00	0.16	0.8995	0.0423	6.4	1.1	6.0	1.2
Math	6	561927	136	0	19	C.1.1.3		2698	0.49	0.00	0.27	0.2945	0.0407	1.3	1.0	1.0	1.0
Math	6	561842	137	0	20	C.1.1.4		2698	0.30	0.00	0.21	1.1907	0.0441	2.6	1.1	3.8	1.1
Math	6	561953	138	0	21	C.1.2.1		2698	0.71	0.00	0.30	-0.7734	0.0443	-1.3	1.0	0.0	1.0
Math	6	561922	139	0	22	C.3.1.1		2698	0.71	0.00	0.28	-0.7856	0.0444	-0.9	1.0	-0.9	1.0
Math	6	561853	140	0	23	D.1.1.1		2698	0.46	0.00	0.33	0.4193	0.0408	-2.0	1.0	-1.3	1.0
Math	6	561787	141	0	24	D.1.2.1	1	2698	0.39	0.01	0.11	0.7526	0.0417	9.9	1.2	9.4	1.2
Math	6	561818	142	0	25	D.1.2.1		2698	0.46	0.00	0.39	0.4347	0.0408	-5.9	0.9	-5.1	0.9
Math	6	561655	143	0	26	D.2.1.1	1	2698	0.49	0.00	0.28	0.2928	0.0407	1.2	1.0	0.9	1.0
Math	6	561696	144	0	27	D.2.1.2		2698	0.78	0.00	0.36	-1.2258	0.0486	-3.6	0.9	-4.0	0.8
Math	6	561843	145	0	28	D.2.1.2		2698	0.32	0.01	0.28	1.1078	0.0435	0.0	1.0	0.8	1.0
Math	6	561794	146	0	29	D.2.2.1		2698	0.30	0.01	0.17	1.1947	0.0441	3.7	1.1	5.9	1.2
Math	6	561800	147	0	30	E.1.1.1		2698	0.80	0.00	0.29	-1.2961	0.0495	-1.0	1.0	-1.4	0.9
Math	6	561642	148	0	31	E.2.1.1	2	2698	0.36	0.00	0.29	0.8959	0.0423	0.5	1.0	1.0	1.0
Math	6	561769	149	0	32	E.3.1.2		2698	0.53	0.00	0.31	0.0933	0.0407	-1.0	1.0	-0.6	1.0
Math	6	561932	152	1	33	B.1.1.1	2	918	0.24	0.01	0.15	1.5401	0.0812	1.9	1.1	2.9	1.2
Math	6	561846	153	1	34	C.1.1.1		918	0.32	0.00	0.32	1.0642	0.0744	-0.9	1.0	-0.8	1.0
Math	6	597165	154	1	35	C.1.2.2	1	918	0.72	0.00	0.24	-0.8733	0.0770	0.5	1.0	1.0	1.1
Math	6	560721	155	1	36	D.1.1.1		918	0.57	0.00	0.32	-0.1258	0.0704	-0.7	1.0	-1.0	1.0
Math	6	597516	156	1	37	D.2.1.1	1	918	0.39	0.00	0.17	0.6990	0.0712	4.0	1.1	4.2	1.2
Math	6	597259	157	1	38	E.1.1.1	2	918	0.27	0.00	0.30	1.3346	0.0778	-0.4	1.0	-0.4	1.0
Math	6	597254	158	1	39	E.1.1.3	2	918	0.44	0.00	0.18	0.4956	0.0703	3.5	1.1	3.4	1.1
Math	6	561712	159	1	40	E.3.1.1		918	0.17	0.00	0.14	1.9711	0.0904	1.0	1.1	2.7	1.2
Math	6	597616	161	2	33	A.1.1.1	1	899	0.39	0.00	0.18	0.7755	0.0720	3.2	1.1	3.4	1.1
Math	6	561935	162	2	34	A.1.1.4	1	899	0.51	0.01	0.44	0.2484	0.0704	-5.2	0.9	-4.7	0.9
Math	6	597807	163	2	35	A.1.3.2	1	899	0.37	0.00	0.20	0.8605	0.0726	2.7	1.1	2.5	1.1
Math	6	561900	164	2	36	A.1.4.1		899	0.62	0.00	0.30	-0.3021	0.0725	-0.3	1.0	-0.3	1.0
Math	6	561952	165	2	37	C.3.1.1		899	0.72	0.00	0.34	-0.8149	0.0780	-1.3	1.0	-2.0	0.9
Math	6	597530	166	2	38	D.1.2.1	2	899	0.44	0.00	0.29	0.5218	0.0708	0.4	1.0	0.4	1.0
Math	6	560722	167	2	39	D.2.2.1	1	899	0.52	0.00	0.12	0.1626	0.0705	6.0	1.2	5.8	1.2
Math	6	597314	168	2	40	E.1.1.2	2	899	0.57	0.00	0.39	-0.0665	0.0711	-3.2	0.9	-2.8	0.9
Math	6	561908	170	3	33	A.1.1.2		881	0.42	0.00	0.40	0.5512	0.0720	-3.4	0.9	-3.1	0.9
Math	6	597605	171	3	34	A.1.1.3	1	881	0.44	0.00	0.36	0.4621	0.0716	-1.9	1.0	-2.2	0.9

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	6	561929	172	3	35	A.1.2.1		881	0.30	0.00	0.25	1.1550	0.0772	0.3	1.0	0.8	1.0
Math	6	597799	173	3	36	A.3.2.1	2	881	0.31	0.00	0.21	1.1249	0.0768	0.6	1.0	2.8	1.2
Math	6	597210	174	3	37	C.1.1.3	1	881	0.66	0.00	0.25	-0.5450	0.0746	0.8	1.0	0.8	1.0
Math	6	597651	175	3	38	C.1.2.1	1	881	0.52	0.00	0.16	0.1054	0.0712	4.5	1.1	4.1	1.1
Math	6	560727	176	3	39	E.1.1.3		881	0.70	0.00	0.31	-0.7485	0.0768	-0.6	1.0	-1.2	0.9
Math	6	597378	177	3	40	E.2.1.1	2	881	0.24	0.00	0.16	1.4760	0.0820	2.0	1.1	3.2	1.2
Math	7	561336	178	0	1	A.1.1.1		2814	0.85	0.00	0.22	-1.6574	0.0549	0.2	1.0	1.1	1.1
Math	7	561341	179	0	2	A.1.2.1		2814	0.53	0.00	0.43	0.1868	0.0399	-7.1	0.9	-6.6	0.9
Math	7	560714	180	0	3	A.1.2.1		2814	0.57	0.00	0.44	-0.0175	0.0403	-8.4	0.9	-7.7	0.9
Math	7	561360	181	0	4	A.2.2.6	2	2814	0.56	0.00	0.32	0.0419	0.0401	-0.3	1.0	0.4	1.0
Math	7	560755	182	0	5	A.3.2.2	1	2814	0.51	0.00	0.24	0.2706	0.0399	4.0	1.1	3.4	1.1
Math	7	560756	183	0	6	B.1.1.1	2	2814	0.54	0.00	0.34	0.1549	0.0400	-2.3	1.0	-2.5	1.0
Math	7	561374	184	0	7	B.2.1.2	1	2814	0.57	0.00	0.28	-0.0311	0.0403	2.1	1.0	1.4	1.0
Math	7	561385	185	0	8	B.2.1.3	1	2814	0.44	0.00	0.33	0.5767	0.0401	-1.1	1.0	0.6	1.0
Math	7	561394	186	0	9	B.2.1.3	1	2814	0.55	0.01	0.31	0.0943	0.0401	1.0	1.0	1.1	1.0
Math	7	561353	187	0	10	B.2.2.1	2	2814	0.28	0.00	0.33	1.4522	0.0446	-1.0	1.0	-0.9	1.0
Math	7	561410	188	0	11	C.1.1.1	1	2814	0.52	0.00	0.35	0.2119	0.0399	-2.7	1.0	-2.6	1.0
Math	7	561405	189	0	12	C.1.1.2	1	2814	0.58	0.00	0.33	-0.0704	0.0404	-1.1	1.0	-1.6	1.0
Math	7	561413	190	0	13	C.1.1.3	1	2814	0.82	0.00	0.30	-1.3639	0.0504	-1.6	1.0	-1.3	0.9
Math	7	561338	191	0	14	C.1.2.1	1	2814	0.30	0.00	0.22	1.2977	0.0434	2.9	1.1	4.2	1.1
Math	7	560759	192	0	15	C.1.2.2	1	2814	0.58	0.00	0.36	-0.0824	0.0404	-2.6	1.0	-2.8	0.9
Math	7	561425	194	0	18	C.3.1.1	1	2814	0.63	0.00	0.34	-0.2933	0.0412	-1.8	1.0	-1.7	1.0
Math	7	561347	195	0	19	C.3.1.2	1	2814	0.49	0.00	0.28	0.3359	0.0399	1.8	1.0	1.6	1.0
Math	7	560760	196	0	20	D.1.1.1	2	2814	0.58	0.00	0.17	-0.0448	0.0403	7.6	1.1	6.1	1.1
Math	7	561725	197	0	21	D.1.1.1	2	2814	0.34	0.00	0.21	1.0737	0.0419	4.5	1.1	4.9	1.1
Math	7	561901	198	0	22	D.1.1.1	2	2814	0.56	0.00	0.27	0.0385	0.0402	2.2	1.0	1.1	1.0
Math	7	561715	199	0	23	D.2.1.1	1	2814	0.66	0.01	0.44	-0.4671	0.0420	-7.1	0.9	-6.6	0.8
Math	7	561886	200	0	24	D.2.1.2	1	2814	0.53	0.01	0.42	0.1801	0.0400	-6.8	0.9	-5.9	0.9
Math	7	561691	201	0	25	D.2.2.1	2	2814	0.55	0.00	0.25	0.0656	0.0401	3.8	1.1	3.4	1.1
Math	7	561916	202	0	26	D.3.1.1	2	2814	0.54	0.00	0.28	0.1583	0.0400	2.8	1.0	1.9	1.0
Math	7	561651	203	0	27	E.1.1.1	1	2814	0.46	0.00	0.30	0.4753	0.0400	0.8	1.0	1.3	1.0
Math	7	561670	204	0	28	E.2.1.1	1	2814	0.75	0.00	0.29	-0.9248	0.0454	-0.2	1.0	0.2	1.0
Math	7	561682	205	0	29	E.2.1.1	1	2814	0.70	0.00	0.32	-0.6902	0.0435	-1.2	1.0	-0.2	1.0
Math	7	560763	206	0	30	E.2.1.2	2	2814	0.54	0.01	0.29	0.1179	0.0400	1.5	1.0	1.4	1.0
Math	7	561941	207	0	31	E.3.1.1	2	2814	0.43	0.00	0.32	0.6483	0.0403	-1.2	1.0	-0.7	1.0
Math	7	560762	208	0	32	E.4.1.1	1	2814	0.86	0.00	0.32	-1.7379	0.0563	-1.7	0.9	-3.1	0.8
Math	7	597484	211	1	33	A.1.1.1	1	946	0.32	0.00	0.34	1.1406	0.0726	-0.9	1.0	-0.9	1.0
Math	7	597493	212	1	34	A.2.2.3	2	946	0.39	0.00	0.04	0.8115	0.0698	7.3	1.2	6.7	1.2
Math	7	597345	213	1	35	C.1.1.1	1	946	0.63	0.00	0.26	-0.2814	0.0705	0.9	1.0	0.2	1.0

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	7	597434	214	1	36	C.1.2.1	2	946	0.28	0.00	0.06	1.3917	0.0757	4.5	1.2	4.6	1.3
Math	7	597223	215	1	37	D.1.1.1	2	946	0.37	0.00	0.31	0.8916	0.0704	-0.9	1.0	0.0	1.0
Math	7	597273	216	1	38	D.3.1.1	2	946	0.18	0.01	0.06	2.0026	0.0873	1.8	1.1	5.0	1.4
Math	7	597175	217	1	39	E.3.1.2	2	946	0.36	0.00	0.13	0.9832	0.0711	4.1	1.1	4.7	1.2
Math	7	561877	218	1	40	E.4.1.1	2	946	0.37	0.00	0.19	0.8966	0.0704	2.6	1.1	2.3	1.1
Math	7	597464	220	2	33	A.1.2.1	1	943	0.64	0.00	0.49	-0.3816	0.0719	-5.6	0.9	-4.6	0.8
Math	7	561352	221	2	34	A.2.2.3	2	943	0.39	0.01	0.28	0.8205	0.0708	1.1	1.0	1.8	1.1
Math	7	597479	222	2	35	A.3.2.1	2	943	0.18	0.00	0.10	2.0328	0.0885	1.6	1.1	5.2	1.5
Math	7	561398	223	2	36	B.2.2.1	2	943	0.53	0.00	0.16	0.1752	0.0692	5.2	1.1	4.9	1.2
Math	7	597267	224	2	37	D.2.1.2	1	943	0.43	0.00	0.32	0.6049	0.0697	0.5	1.0	0.7	1.0
Math	7	597263	225	2	38	D.2.2.1	2	943	0.60	0.00	0.23	-0.1758	0.0705	2.6	1.1	2.4	1.1
Math	7	561933	226	2	39	D.3.1.1	2	943	0.45	0.00	0.28	0.5359	0.0695	1.6	1.0	1.6	1.1
Math	7	561869	227	2	40	E.1.1.1	2	943	0.76	0.00	0.31	-1.0347	0.0802	-1.0	1.0	-0.7	1.0
Math	7	597500	229	3	33	A.2.1.1	1	925	0.58	0.00	0.28	-0.0387	0.0706	1.5	1.0	1.1	1.0
Math	7	562192	230	3	34	A.2.2.4		925	0.63	0.00	0.32	-0.2918	0.0721	-0.4	1.0	0.1	1.0
Math	7	597482	231	3	35	A.3.2.2	1	925	0.56	0.00	0.35	0.0471	0.0703	-0.9	1.0	-0.7	1.0
Math	7	597311	232	3	36	B.2.1.1	2	925	0.15	0.00	0.14	2.2364	0.0941	0.6	1.0	3.9	1.4
Math	7	561367	233	3	37	C.3.1.2	1	925	0.42	0.00	0.24	0.6708	0.0706	2.7	1.1	2.6	1.1
Math	7	561924	234	3	38	D.2.1.1	1	925	0.54	0.00	0.45	0.1621	0.0700	-4.2	0.9	-3.3	0.9
Math	7	597235	235	3	39	D.3.1.1	2	925	0.29	0.00	0.20	1.3347	0.0761	2.1	1.1	3.4	1.2
Math	7	561674	236	3	40	D.3.1.2	2	925	0.15	0.00	0.09	2.3007	0.0959	1.3	1.1	4.2	1.5
Math	8	561346	237	0	1	A.1.1.1	1	3012	0.63	0.00	0.41	-0.3466	0.0398	-6.0	0.9	-5.8	0.9
Math	8	561356	238	0	2	A.1.1.2	1	3012	0.74	0.00	0.41	-0.9411	0.0436	-5.0	0.9	-5.8	0.8
Math	8	561358	239	0	3	A.2.1.1	1	3012	0.64	0.00	0.21	-0.3830	0.0400	5.0	1.1	3.9	1.1
Math	8	561409	240	0	4	A.3.3.1	2	3012	0.30	0.01	0.17	1.2843	0.0422	5.7	1.1	6.8	1.2
Math	8	561393	241	0	5	B.1.1.1	1	3012	0.53	0.00	0.28	0.1542	0.0387	2.1	1.0	2.0	1.0
Math	8	561412	242	0	6	B.1.1.2	1	3012	0.46	0.00	0.35	0.4485	0.0388	-2.5	1.0	-2.1	1.0
Math	8	560741	243	0	7	B.2.1.1	1	3012	0.62	0.00	0.32	-0.2828	0.0396	-0.5	1.0	-1.3	1.0
Math	8	561366	244	0	8	B.2.1.2	1	3012	0.48	0.00	0.27	0.3538	0.0387	2.7	1.0	2.2	1.0
Math	8	561635	245	0	9	B.2.2.2	1	3012	0.87	0.00	0.27	-1.8432	0.0554	-1.0	1.0	-2.0	0.9
Math	8	560742	246	0	10	C.1.1.1	2	3012	0.54	0.00	0.29	0.0984	0.0387	1.5	1.0	0.8	1.0
Math	8	561382	247	0	11	C.1.1.1	2	3012	0.71	0.00	0.22	-0.7082	0.0418	2.3	1.0	3.0	1.1
Math	8	561637	248	0	12	C.1.1.2	2	3012	0.52	0.00	0.28	0.1728	0.0387	2.4	1.0	1.8	1.0
Math	8	561418	249	0	13	C.1.1.3		3012	0.40	0.00	0.29	0.7619	0.0395	1.2	1.0	2.0	1.0
Math	8	561647	250	0	14	C.1.2.1	2	3012	0.41	0.00	0.33	0.7200	0.0394	-0.8	1.0	-0.1	1.0
Math	8	560744	251	0	15	C.1.2.1	2	3012	0.33	0.00	0.24	1.0955	0.0410	3.1	1.1	4.6	1.1
Math	8	561653	253	0	18	C.3.1.1	1	3012	0.67	0.00	0.30	-0.5749	0.0410	-0.7	1.0	0.9	1.0
Math	8	561368	254	0	19	D.1.1.1	2	3012	0.39	0.00	0.23	0.7749	0.0395	4.4	1.1	5.1	1.1
Math	8	561355	255	0	20	D.1.1.1	2	3012	0.71	0.01	0.33	-0.7595	0.0422	-1.8	1.0	-2.6	0.9

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	8	561582	256	0	21	D.1.1.3	2	3012	0.41	0.01	0.29	0.6720	0.0392	0.7	1.0	1.5	1.0
Math	8	560746	257	0	22	D.2.1.3	1	3012	0.58	0.00	0.37	-0.1145	0.0391	-4.2	0.9	-4.2	0.9
Math	8	561583	258	0	23	D.2.1.3	1	3012	0.25	0.00	0.21	1.5511	0.0445	3.1	1.1	3.9	1.1
Math	8	561554	259	0	24	D.4.1.1	1	3012	0.48	0.00	0.33	0.3802	0.0387	-1.6	1.0	-1.3	1.0
Math	8	561546	260	0	25	D.4.1.2	1	3012	0.51	0.00	0.24	0.2316	0.0386	4.8	1.1	4.4	1.1
Math	8	561519	261	0	26	E.1.1.1	2	3012	0.60	0.00	0.30	-0.1989	0.0393	0.4	1.0	0.1	1.0
Math	8	561594	262	0	27	E.1.1.2	2	3012	0.80	0.00	0.17	-1.2504	0.0467	2.0	1.1	4.0	1.2
Math	8	560748	263	0	28	E.1.1.3	1	3012	0.70	0.00	0.39	-0.7010	0.0418	-4.7	0.9	-5.2	0.9
Math	8	561575	264	0	29	E.1.1.3	1	3012	0.80	0.01	0.27	-1.2711	0.0470	-0.6	1.0	-1.2	1.0
Math	8	561561	265	0	30	E.3.1.1	2	3012	0.51	0.00	0.37	0.2347	0.0386	-4.1	1.0	-3.9	0.9
Math	8	561401	266	0	31	E.4.1.1	1	3012	0.40	0.00	0.27	0.7635	0.0395	2.0	1.0	2.9	1.1
Math	8	561502	267	0	32	E.4.1.2	2	3012	0.63	0.00	0.40	-0.3220	0.0397	-5.9	0.9	-4.6	0.9
Math	8	597611	270	1	33	A.1.1.1	1	1008	0.79	0.00	0.42	-1.1974	0.0805	-3.1	0.9	-3.9	0.7
Math	8	597622	271	1	34	A.2.2.1	2	1008	0.26	0.00	0.26	1.5135	0.0758	0.5	1.0	1.0	1.1
Math	8	597507	272	1	35	C.1.1.2	2	1008	0.49	0.00	0.19	0.3696	0.0668	4.7	1.1	3.8	1.1
Math	8	561415	273	1	36	C.3.1.1	1	1008	0.70	0.00	0.38	-0.6597	0.0721	-2.4	0.9	-2.8	0.9
Math	8	597344	274	1	37	D.1.1.1	2	1008	0.24	0.00	0.13	1.6272	0.0776	2.2	1.1	5.6	1.4
Math	8	597418	275	1	38	D.2.1.1	1	1008	0.68	0.01	0.40	-0.5762	0.0712	-3.1	0.9	-3.5	0.8
Math	8	561571	276	1	39	D.4.1.1	1	1008	0.41	0.00	0.29	0.7324	0.0679	0.6	1.0	1.1	1.0
Math	8	597315	277	1	40	E.4.1.2	2	1008	0.29	0.00	0.14	1.3329	0.0733	3.6	1.1	3.7	1.2
Math	8	597615	279	2	33	A.1.1.1	1	1005	0.52	0.00	0.27	0.1592	0.0666	1.2	1.0	0.7	1.0
Math	8	561406	280	2	34	A.3.3.1	2	1005	0.35	0.00	0.16	0.9495	0.0695	3.4	1.1	4.7	1.2
Math	8	597518	281	2	35	B.2.2.3	1	1005	0.50	0.00	0.21	0.2268	0.0665	3.1	1.1	2.9	1.1
Math	8	597360	282	2	36	D.1.1.2	1	1005	0.43	0.00	0.41	0.5436	0.0671	-3.9	0.9	-3.6	0.9
Math	8	597410	283	2	37	D.2.1.2	1	1005	0.25	0.00	0.11	1.4876	0.0760	3.0	1.1	4.8	1.3
Math	8	561632	284	2	38	D.2.2.1	2	1005	0.37	0.00	0.21	0.8523	0.0687	2.3	1.1	2.9	1.1
Math	8	561550	285	2	39	E.1.1.2	2	1005	0.81	0.00	0.35	-1.3844	0.0833	-1.8	0.9	-2.1	0.9
Math	8	597283	286	2	40	E.3.2.1	2	1005	0.33	0.00	0.13	1.0290	0.0702	4.2	1.1	4.9	1.2
Math	8	597587	288	3	33	A.2.1.1	1	999	0.50	0.01	0.47	0.2362	0.0675	-5.7	0.9	-5.0	0.9
Math	8	561389	289	3	34	A.3.1.1		999	0.48	0.01	0.16	0.3567	0.0676	6.1	1.2	5.4	1.2
Math	8	597550	290	3	35	B.1.1.4	1	999	0.55	0.01	0.36	0.0413	0.0677	-0.8	1.0	-1.3	1.0
Math	8	597564	291	3	36	C.1.1.3	1	999	0.55	0.01	0.29	0.0320	0.0677	1.6	1.0	1.6	1.1
Math	8	561601	292	3	37	D.1.1.3	2	999	0.46	0.00	0.33	0.4452	0.0677	0.1	1.0	0.3	1.0
Math	8	597373	293	3	38	D.2.2.1	2	999	0.29	0.00	0.14	1.2881	0.0739	3.7	1.1	5.2	1.3
Math	8	560747	294	3	39	D.2.2.2	1	999	0.56	0.00	0.44	-0.0381	0.0679	-4.9	0.9	-4.5	0.9
Math	8	597184	295	3	40	D.4.1.2	1	999	0.37	0.01	0.26	0.8757	0.0698	1.7	1.1	3.0	1.1
Math	11	561116	296	0	1	A.1.1.1		3532	0.86	0.00	0.31	-1.9019	0.0495	-2.8	0.9	-5.4	0.7
Math	11	561124	297	0	2	A.1.3.1		3532	0.66	0.00	0.32	-0.6527	0.0373	-2.4	1.0	-1.6	1.0
Math	11	560754	298	0	3	A.2.1.1	2	3532	0.50	0.00	0.20	0.1395	0.0356	7.1	1.1	6.4	1.1

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	11	561129	299	0	4	A.2.1.2		3532	0.48	0.00	0.33	0.1884	0.0356	-1.5	1.0	-1.9	1.0
Math	11	561137	300	0	5	A.3.1.1		3532	0.46	0.00	0.35	0.2586	0.0357	-2.9	1.0	-2.7	1.0
Math	11	561138	301	0	6	B.2.1.1		3532	0.72	0.00	0.24	-0.9624	0.0391	0.9	1.0	0.0	1.0
Math	11	561163	302	0	7	C.1.1.1	2	3532	0.47	0.00	0.18	0.2679	0.0357	9.0	1.1	8.9	1.2
Math	11	561166	303	0	8	C.1.1.2	1	3532	0.39	0.00	0.29	0.6302	0.0365	0.9	1.0	1.3	1.0
Math	11	561118	304	0	9	C.1.2.3		3532	0.53	0.00	0.27	-0.0190	0.0356	2.0	1.0	2.3	1.0
Math	11	561181	305	0	10	C.1.3.1		3532	0.84	0.00	0.06	-1.7270	0.0469	2.5	1.1	7.1	1.4
Math	11	561164	306	0	11	C.1.4.1	1	3532	0.54	0.00	0.26	-0.0866	0.0357	3.1	1.0	3.1	1.1
Math	11	560770	307	0	12	C.3.1.1	1	3532	0.53	0.00	0.44	-0.0548	0.0356	-9.6	0.9	-8.6	0.9
Math	11	561190	308	0	13	D.1.1.1	2	3532	0.45	0.01	0.39	0.3464	0.0358	-5.6	0.9	-5.5	0.9
Math	11	560753	309	0	14	D.2.1.3		3532	0.54	0.01	0.33	-0.0641	0.0356	-1.2	1.0	-1.5	1.0
Math	11	561229	310	0	15	D.2.1.4	1	3532	0.49	0.00	0.33	0.1646	0.0356	-0.8	1.0	-0.9	1.0
Math	11	561133	312	0	18	D.2.1.5		3532	0.45	0.01	0.30	0.3517	0.0358	0.0	1.0	0.4	1.0
Math	11	561139	313	0	19	D.2.2.1		3532	0.40	0.00	0.39	0.5626	0.0363	-5.9	0.9	-4.8	0.9
Math	11	561414	314	0	20	D.2.2.2		3532	0.44	0.01	0.24	0.3932	0.0359	3.7	1.1	3.7	1.1
Math	11	561143	315	0	21	D.3.1.1		3532	0.38	0.00	0.28	0.6636	0.0367	1.7	1.0	1.8	1.0
Math	11	561144	316	0	22	D.3.1.2		3532	0.43	0.01	0.20	0.4106	0.0359	6.5	1.1	6.1	1.1
Math	11	561147	317	0	23	D.3.2.1	1	3532	0.53	0.01	0.45	-0.0376	0.0356	-9.9	0.9	-9.9	0.9
Math	11	561150	318	0	24	D.3.2.2	1	3532	0.52	0.01	0.31	0.0021	0.0356	-0.6	1.0	-0.7	1.0
Math	11	561154	319	0	25	D.4.1.1		3532	0.68	0.00	0.42	-0.7397	0.0377	-7.8	0.9	-7.2	0.8
Math	11	561430	320	0	26	D.4.1.1		3532	0.37	0.00	0.23	0.7029	0.0368	4.0	1.1	4.2	1.1
Math	11	561156	321	0	27	E.1.1.2		3532	0.45	0.01	0.33	0.3157	0.0358	-1.7	1.0	-2.0	1.0
Math	11	561141	322	0	28	E.2.1.1		3532	0.71	0.01	0.21	-0.8878	0.0386	2.7	1.1	2.0	1.1
Math	11	561158	323	0	29	E.2.1.2		3532	0.29	0.00	0.30	1.1182	0.0390	-0.5	1.0	0.4	1.0
Math	11	561159	324	0	30	E.4.1.1		3532	0.64	0.00	0.27	-0.5620	0.0369	0.4	1.0	0.8	1.0
Math	11	561161	325	0	31	E.4.2.1		3532	0.29	0.01	0.30	1.1325	0.0391	-0.8	1.0	0.6	1.0
Math	11	561162	326	0	32	E.4.2.2		3532	0.51	0.00	0.31	0.0471	0.0356	-0.3	1.0	0.5	1.0
Math	11	597424	329	1	33	A.3.1.1	1	1177	0.23	0.00	0.10	1.5171	0.0730	3.4	1.1	5.1	1.3
Math	11	561153	330	1	34	B.2.2.4	2	1177	0.70	0.00	0.45	-0.8252	0.0665	-4.9	0.9	-5.3	0.8
Math	11	597222	331	1	35	C.1.2.2	2	1177	0.27	0.01	0.23	1.2411	0.0689	1.3	1.0	2.9	1.2
Math	11	597286	332	1	36	D.1.1.1	2	1177	0.46	0.01	0.32	0.3223	0.0621	-0.2	1.0	0.1	1.0
Math	11	597099	333	1	37	D.2.1.2	2	1177	0.31	0.00	0.19	1.0088	0.0663	2.9	1.1	3.6	1.2
Math	11	597292	334	1	38	D.2.2.3	1	1177	0.29	0.01	0.03	1.1273	0.0676	6.9	1.2	6.4	1.3
Math	11	597233	335	1	39	D.3.2.2	2	1177	0.38	0.01	0.32	0.6997	0.0638	-0.6	1.0	0.0	1.0
Math	11	597186	336	1	40	D.4.1.1	1	1177	0.32	0.01	0.31	0.9599	0.0658	-0.4	1.0	0.1	1.0
Math	11	561117	338	2	33	A.1.1.3		1160	0.35	0.00	0.43	0.7989	0.0647	-4.2	0.9	-4.0	0.9
Math	11	597371	339	2	34	B.2.1.1	1	1160	0.55	0.00	0.29	-0.1131	0.0620	-0.3	1.0	0.2	1.0
Math	11	597291	340	2	35	B.2.3.1	2	1160	0.54	0.01	0.13	-0.0624	0.0619	6.2	1.1	6.9	1.2
Math	11	597330	341	2	36	D.1.1.2	1	1160	0.20	0.01	-0.11	1.6512	0.0761	5.3	1.3	8.7	1.7

Item Information								Classical				Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	PVal	P(-)	PtBis	Meas	MeasSE	t	MS	t	MS
Math	11	597101	342	2	37	D.2.1.4	2	1160	0.20	0.00	0.10	1.6630	0.0763	2.1	1.1	4.5	1.3
Math	11	597261	343	2	38	D.3.2.1	2	1160	0.47	0.00	0.23	0.2569	0.0620	2.9	1.1	2.9	1.1
Math	11	597275	344	2	39	D.3.1.1	2	1160	0.08	0.00	-0.14	2.8546	0.1128	1.6	1.2	7.4	2.3
Math	11	597163	345	2	40	E.3.1.2	2	1160	0.12	0.00	0.05	2.2650	0.0912	1.8	1.1	4.2	1.4
Math	11	561132	347	3	33	A.2.1.3		1195	0.45	0.00	0.40	0.2931	0.0615	-3.3	0.9	-3.7	0.9
Math	11	597137	348	3	34	A.2.2.1	1	1195	0.31	0.00	0.31	0.9936	0.0660	-0.4	1.0	-0.2	1.0
Math	11	561140	349	3	35	B.2.2.2		1195	0.50	0.00	0.20	0.0903	0.0612	4.6	1.1	3.6	1.1
Math	11	597284	350	3	36	C.3.1.2	1	1195	0.25	0.00	0.14	1.3569	0.0705	2.6	1.1	4.2	1.2
Math	11	561208	351	3	37	D.1.1.3		1195	0.41	0.00	0.15	0.5034	0.0623	5.5	1.1	5.6	1.2
Math	11	561142	352	3	38	D.2.2.2	1	1195	0.35	0.00	0.37	0.7664	0.0639	-2.0	1.0	-2.1	0.9
Math	11	597207	353	3	39	D.4.1.1	2	1195	0.17	0.00	0.13	1.8821	0.0802	1.7	1.1	3.0	1.2
Math	11	560774	354	3	40	E.2.1.2		1195	0.19	0.00	0.20	1.7192	0.0767	0.6	1.0	2.4	1.2

Item Information								Classical														Rasch		Infit		Outfit	
Cont	Grade	ID	PubID	Form	Seq	Std	DOK	N	Mean	P(0)	P(1)	P(2)	P(3)	P(4)	P(B)	PtBis	PT(0)	PT(1)	PT(2)	PT(3)	PT(4)	Meas	MeasSE	t	MS	t	MS
Math	4	561835	16	0	16	A.2	3	2169	1.39	0.36	0.30	0.09	0.12	0.14	0.00	0.53	-0.51	0.07	0.11	0.21	0.33	1.5309	0.0210	1.7	1.1	2.1	1.1
Math	4	561881	32	0	41	B.1	2	2169	1.26	0.31	0.32	0.21	0.09	0.06	0.00	0.53	-0.44	-0.02	0.20	0.25	0.27	1.7937	0.0239	-0.4	1.0	-1.0	1.0
Math	4	597111	33	1	17	A.1	2	735	1.43	0.21	0.37	0.23	0.19	0.01	0.00	0.54	-0.41	-0.11	0.14	0.40	0.10	2.1505	0.0444	-0.3	1.0	0.0	1.0
Math	4	560710	42	2	17	C.1	3	716	0.57	0.57	0.31	0.11	0.01	0.00	0.00	0.30	-0.30	0.18	0.18	0.03	0.07	3.1698	0.0581	3.7	1.2	4.7	1.4
Math	4	597741	51	3	17	D.2	2	718	1.76	0.16	0.34	0.23	0.13	0.14	0.01	0.57	-0.41	-0.20	0.10	0.23	0.36	1.1950	0.0387	0.1	1.0	0.1	1.0
Math	5	561854	75	0	16	A.2		2551	0.66	0.62	0.21	0.08	0.07	0.02	0.00	0.41	-0.33	0.03	0.20	0.26	0.18	2.4570	0.0251	1.6	1.1	3.5	1.2
Math	5	560729	91	0	41	D.1		2551	1.71	0.31	0.24	0.11	0.12	0.23	0.00	0.55	-0.48	-0.03	0.06	0.11	0.44	1.0408	0.0180	0.6	1.0	0.6	1.0
Math	5	597819	92	1	17	A.2	3	878	1.42	0.37	0.14	0.23	0.21	0.04	0.01	0.58	-0.51	-0.06	0.15	0.40	0.19	1.6849	0.0345	-1.0	1.0	-1.4	0.9
Math	5	597744	101	2	17	C.2	3	836	0.88	0.33	0.49	0.16	0.02	0.00	0.00	0.28	-0.28	0.14	0.13	0.08	0.05	2.9234	0.0527	3.8	1.2	4.3	1.2
Math	5	597548	110	3	17	E.2	3	835	1.14	0.20	0.57	0.15	0.06	0.02	0.01	0.47	-0.27	-0.15	0.27	0.25	0.19	1.8311	0.0468	0.1	1.0	-1.0	0.9
Math	6	560738	134	0	16	A.3		2698	0.87	0.42	0.40	0.12	0.04	0.03	0.00	0.47	-0.40	0.09	0.24	0.22	0.19	1.4602	0.0244	-0.8	1.0	-1.5	1.0
Math	6	561740	150	0	41	E.2		2698	1.20	0.17	0.55	0.21	0.06	0.01	0.00	0.50	-0.33	-0.16	0.30	0.26	0.15	1.3380	0.0272	-2.9	0.9	-3.6	0.9
Math	6	597805	151	1	17	A.2	2	914	0.63	0.56	0.33	0.04	0.06	0.01	0.01	0.38	-0.30	0.10	0.16	0.24	0.13	1.8770	0.0448	1.6	1.1	3.1	1.2
Math	6	597798	160	2	17	B.2	3	899	1.45	0.11	0.47	0.32	0.08	0.03	0.01	0.50	-0.31	-0.23	0.23	0.27	0.20	0.8679	0.0438	-1.6	0.9	-2.0	0.9
Math	6	597393	169	3	17	D.2	2	881	1.38	0.23	0.40	0.17	0.16	0.04	0.01	0.57	-0.44	-0.10	0.15	0.36	0.23	0.9107	0.0372	-2.6	0.9	-2.9	0.9
Math	7	561437	193	0	16	A.2	2	2814	1.52	0.27	0.31	0.15	0.16	0.11	0.01	0.61	-0.50	-0.08	0.13	0.27	0.36	0.7503	0.0187	-4.4	0.9	-4.7	0.9
Math	7	561447	209	0	41	D.3	2	2814	0.69	0.75	0.04	0.06	0.07	0.08	0.01	0.41	-0.41	0.07	0.15	0.22	0.26	1.4389	0.0197	6.6	1.3	3.8	1.4
Math	7	597316	210	1	17	B.1	2	945	1.37	0.29	0.34	0.20	0.07	0.11	0.02	0.48	-0.29	-0.19	0.16	0.22	0.32	0.8001	0.0326	1.0	1.1	1.0	1.1
Math	7	597298	219	2	17	C.1	2	943	1.03	0.50	0.12	0.25	0.12	0.02	0.01	0.57	-0.50	-0.07	0.32	0.33	0.17	1.5304	0.0352	-1.3	0.9	-1.9	0.9
Math	7	560771	228	3	17	E.2	2	925	1.53	0.16	0.45	0.15	0.16	0.07	0.01	0.56	-0.35	-0.24	0.15	0.33	0.28	0.7367	0.0359	-0.3	1.0	-0.4	1.0
Math	8	561664	252	0	16	A.2	2	3012	1.23	0.25	0.40	0.25	0.06	0.04	0.00	0.49	-0.41	-0.02	0.23	0.23	0.19	1.1307	0.0221	0.2	1.0	0.3	1.0
Math	8	561438	268	0	41	D.2	2	3012	1.10	0.34	0.38	0.16	0.10	0.02	0.01	0.54	-0.50	0.07	0.26	0.29	0.15	1.3549	0.0215	-2.8	0.9	-3.8	0.9
Math	8	597582	269	1	17	B.2	3	1006	0.60	0.60	0.28	0.07	0.02	0.03	0.02	0.54	-0.52	0.28	0.21	0.19	0.28	1.7535	0.0425	-1.9	0.9	-3.2	0.8
Math	8	560765	278	2	17	C.1	2	1005	0.68	0.55	0.28	0.12	0.04	0.01	0.02	0.46	-0.39	0.11	0.27	0.24	0.13	1.8210	0.0410	-0.3	1.0	-0.5	1.0
Math	8	561585	287	3	17	E.3	2	999	0.44	0.60	0.36	0.03	0.00	0.00	0.01	0.47	-0.46	0.38	0.17	0.12	0.08	2.6896	0.0588	-1.3	0.9	-2.8	0.9
Math	11	560782	311	0	16	B.2		3532	0.66	0.53	0.35	0.07	0.02	0.03	0.03	0.52	-0.48	0.22	0.27	0.14	0.26	1.4930	0.0229	-2.8	0.9	-4.8	0.9
Math	11	561120	327	0	41	D.1		3532	0.73	0.41	0.50	0.05	0.04	0.00	0.04	0.50	-0.46	0.26	0.17	0.26	0.10	2.0356	0.0255	-2.5	0.9	-3.3	0.9
Math	11	561114	328	1	17	E.3		1177	0.57	0.52	0.39	0.09	0.00	0.00	0.03	0.43	-0.38	0.21	0.30			1.2325	0.0495	-0.6	1.0	-1.6	0.9
Math	11	597133	337	2	17	C.3	3	1160	1.00	0.45	0.24	0.16	0.14	0.00	0.05	0.68	-0.59	0.02	0.32	0.47		0.7008	0.0333	-8.4	0.7	-7.9	0.7
Math	11	597107	346	3	17	D.2	3	1195	0.33	0.75	0.21	0.03	0.01	0.01	0.09	0.41	-0.41	0.31	0.15	0.18	0.11	2.2177	0.0511	-0.3	1.0	-0.2	1.0

Appendix J:

Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	2169	22.9	6.55	0.81	2.8	MC/OE
	A	All	18	15	2169	9.4	3.82	0.69	2.1	MC/OE
	B	All	5	2	2169	1.8	1.37	0.26	1.2	MC/OE
	C	All	5	5	2169	3.7	1.15	0.41	0.9	MC
	D	All	5	5	2169	3.8	1.08	0.45	0.8	MC
	E	All	5	5	2169	4.2	1.05	0.55	0.7	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	1249	23.2	6.65	0.82	2.8	MC/OE
		Female	38	32	912	22.5	6.42	0.81	2.8	MC/OE
	A	Male	18	15	1249	9.5	3.82	0.69	2.1	MC/OE
		Female	18	15	912	9.3	3.82	0.69	2.1	MC/OE
	B	Male	5	2	1249	1.9	1.40	0.29	1.2	MC/OE
		Female	5	2	912	1.6	1.30	0.22	1.2	MC/OE
	C	Male	5	5	1249	3.7	1.17	0.43	0.9	MC
		Female	5	5	912	3.7	1.13	0.37	0.9	MC
	D	Male	5	5	1249	3.8	1.07	0.46	0.8	MC
		Female	5	5	912	3.7	1.09	0.43	0.8	MC
	E	Male	5	5	1249	4.2	1.05	0.55	0.7	MC
		Female	5	5	912	4.2	1.06	0.56	0.7	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	1492	23.5	6.39	0.81	2.8	MC/OE
		Af. Amer.	38	32	415	21.1	6.57	0.82	2.8	MC/OE
		Hispanic	38	32	193	22.0	6.37	0.80	2.8	MC/OE
		Asian	38	32	27	23.0	7.66	0.85	3.0	MC/OE
		Am. Indian	38	32	4	21.3	7.41	0.86	2.7	MC/OE
		Multi	38	32	28	21.4	9.77	0.92	2.8	MC/OE
	A	White	18	15	1492	9.7	3.79	0.68	2.1	MC/OE
		Af. Amer.	18	15	415	8.5	3.76	0.69	2.1	MC/OE
		Hispanic	18	15	193	9.1	3.60	0.66	2.1	MC/OE
		Asian	18	15	27	9.3	4.10	0.71	2.2	MC/OE
		Am. Indian	18	15	4	8.0	3.74	0.70	2.0	MC/OE
		Multi	18	15	28	9.2	4.93	0.83	2.0	MC/OE
	B	White	5	2	1492	1.9	1.36	0.24	1.2	MC/OE
		Af. Amer.	5	2	415	1.5	1.34	0.34	1.1	MC/OE
		Hispanic	5	2	193	1.6	1.33	0.19	1.2	MC/OE
		Asian	5	2	27	2.3	1.54	0.20	1.4	MC/OE
		Am. Indian	5	2	4	1.5	1.29	0.00	1.3	MC/OE
		Multi	5	2	28	1.8	1.75	0.48	1.3	MC/OE
	C	White	5	5	1492	3.8	1.14	0.42	0.9	MC
		Af. Amer.	5	5	415	3.6	1.19	0.38	0.9	MC
		Hispanic	5	5	193	3.7	1.15	0.38	0.9	MC
		Asian	5	5	27	3.9	1.01	0.19	0.9	MC
		Am. Indian	5	5	4	3.0	1.15	0.00	1.2	MC
		Multi	5	5	28	3.5	1.35	0.52	0.9	MC
	D	White	5	5	1492	3.8	1.05	0.44	0.8	MC
		Af. Amer.	5	5	415	3.6	1.09	0.43	0.8	MC
		Hispanic	5	5	193	3.6	1.10	0.48	0.8	MC
		Asian	5	5	27	3.5	1.31	0.62	0.8	MC
		Am. Indian	5	5	4	4.3	0.96	0.23	0.8	MC
		Multi	5	5	28	3.3	1.44	0.69	0.8	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
	E	White	5	5	1492	4.3	0.97	0.51	0.7	MC
		Af. Amer.	5	5	415	3.9	1.16	0.58	0.8	MC
		Hispanic	5	5	193	4.0	1.20	0.63	0.7	MC
		Asian	5	5	27	4.0	1.39	0.71	0.7	MC
		Am. Indian	5	5	4	4.5	0.58	0.00	0.6	MC
		Multi	5	5	28	3.6	1.34	0.62	0.8	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
ELL	Tot.	All	38	32	90	22.7	6.79	0.82	2.9	MC/OE
	A	All	18	15	90	9.5	3.78	0.67	2.2	MC/OE
	B	All	5	2	90	1.7	1.37	0.11	1.3	MC/OE
	C	All	5	5	90	3.7	1.14	0.38	0.9	MC
	D	All	5	5	90	3.7	1.08	0.44	0.8	MC
	E	All	5	5	90	4.0	1.21	0.64	0.7	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Eco. Disadv.	Tot.	All	38	32	1291	22.5	6.57	0.81	2.8	MC/OE
	A	All	18	15	1291	9.2	3.81	0.69	2.1	MC/OE
	B	All	5	2	1291	1.7	1.36	0.26	1.2	MC/OE
	C	All	5	5	1291	3.7	1.16	0.40	0.9	MC
	D	All	5	5	1291	3.7	1.08	0.45	0.8	MC
	E	All	5	5	1291	4.1	1.08	0.56	0.7	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	2552	21.9	6.43	0.80	2.9	MC/OE
	A	All	17	14	2552	8.4	2.94	0.63	1.8	MC/OE
	B	All	5	5	2552	3.3	1.28	0.43	1.0	MC
	C	All	5	5	2552	3.5	1.25	0.45	0.9	MC
	D	All	6	3	2552	3.2	1.88	0.32	1.6	MC/OE
	E	All	5	5	2552	3.4	1.17	0.43	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	1471	22.2	6.40	0.80	2.9	MC/OE
		Female	38	32	1069	21.5	6.45	0.80	2.9	MC/OE
	A	Male	17	14	1471	8.7	2.93	0.63	1.8	MC/OE
		Female	17	14	1069	8.1	2.92	0.61	1.8	MC/OE
	B	Male	5	5	1471	3.4	1.26	0.44	0.9	MC
		Female	5	5	1069	3.3	1.29	0.43	1.0	MC
	C	Male	5	5	1471	3.5	1.23	0.44	0.9	MC
		Female	5	5	1069	3.5	1.27	0.49	0.9	MC
	D	Male	6	3	1471	3.3	1.86	0.30	1.6	MC/OE
		Female	6	3	1069	3.1	1.90	0.33	1.5	MC/OE
	E	Male	5	5	1471	3.4	1.17	0.43	0.9	MC
		Female	5	5	1069	3.5	1.16	0.43	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	1783	22.1	6.30	0.80	2.9	MC/OE
		Af. Amer.	38	32	465	21.1	6.72	0.81	3.0	MC/OE
		Hispanic	38	32	223	21.9	6.65	0.81	2.9	MC/OE
		Asian	38	32	19	21.4	6.72	0.81	2.9	MC/OE
		Am. Indian	38	32	6	23.5	8.50	0.89	2.8	MC/OE
		Multi	38	32	42	20.8	6.50	0.82	2.8	MC/OE
	A	White	17	14	1783	8.5	2.89	0.62	1.8	MC/OE
		Af. Amer.	17	14	465	8.2	3.01	0.63	1.8	MC/OE
		Hispanic	17	14	223	8.5	3.13	0.67	1.8	MC/OE
		Asian	17	14	19	7.8	2.76	0.61	1.7	MC/OE
		Am. Indian	17	14	6	9.0	3.74	0.77	1.8	MC/OE
		Multi	17	14	42	7.7	2.91	0.64	1.7	MC/OE
	B	White	5	5	1783	3.4	1.26	0.43	1.0	MC
		Af. Amer.	5	5	465	3.2	1.30	0.43	1.0	MC
		Hispanic	5	5	223	3.3	1.36	0.52	0.9	MC
		Asian	5	5	19	3.3	1.28	0.38	1.0	MC
		Am. Indian	5	5	6	3.5	1.64	0.71	0.9	MC
		Multi	5	5	42	3.4	1.10	0.17	1.0	MC
	C	White	5	5	1783	3.5	1.23	0.45	0.9	MC
		Af. Amer.	5	5	465	3.3	1.28	0.45	0.9	MC
		Hispanic	5	5	223	3.3	1.27	0.46	0.9	MC
		Asian	5	5	19	3.3	1.38	0.61	0.9	MC
		Am. Indian	5	5	6	3.7	1.21	0.51	0.8	MC
		Multi	5	5	42	3.3	1.24	0.44	0.9	MC
	D	White	6	3	1783	3.2	1.87	0.33	1.5	MC/OE
		Af. Amer.	6	3	465	3.2	1.92	0.29	1.6	MC/OE
		Hispanic	6	3	223	3.3	1.88	0.28	1.6	MC/OE
		Asian	6	3	19	3.6	1.92	0.17	1.8	MC/OE
		Am. Indian	6	3	6	3.7	1.86	0.17	1.7	MC/OE
		Multi	6	3	42	3.1	1.75	0.32	1.4	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
	E	White	5	5	1783	3.5	1.13	0.42	0.9	MC
		Af. Amer.	5	5	465	3.3	1.26	0.47	0.9	MC
		Hispanic	5	5	223	3.4	1.17	0.40	0.9	MC
		Asian	5	5	19	3.4	0.96	0.02	0.9	MC
		Am. Indian	5	5	6	3.7	1.37	0.67	0.8	MC
		Multi	5	5	42	3.3	1.33	0.56	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
ELL	Tot.	All	38	32	95	21.8	6.58	0.81	2.9	MC/OE
	A	All	17	14	95	8.3	3.04	0.67	1.8	MC/OE
	B	All	5	5	95	3.3	1.34	0.49	1.0	MC
	C	All	5	5	95	3.4	1.24	0.46	0.9	MC
	D	All	6	3	95	3.2	1.85	0.26	1.6	MC/OE
	E	All	5	5	95	3.4	1.23	0.46	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Eco. Disadv.	Tot.	All	38	32	1526	21.8	6.52	0.80	2.9	MC/OE
	A	All	17	14	1526	8.4	2.97	0.63	1.8	MC/OE
	B	All	5	5	1526	3.3	1.29	0.45	1.0	MC
	C	All	5	5	1526	3.4	1.24	0.44	0.9	MC
	D	All	6	3	1526	3.2	1.91	0.33	1.6	MC/OE
	E	All	5	5	1526	3.4	1.19	0.45	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	2700	18.3	6.01	0.80	2.7	MC/OE
	A	All	12	9	2700	5.2	2.44	0.60	1.5	MC/OE
	B	All	5	5	2700	3.1	1.14	0.30	1.0	MC
	C	All	7	7	2700	3.9	1.56	0.42	1.2	MC
	D	All	7	7	2700	3.2	1.59	0.44	1.2	MC
	E	All	7	4	2700	2.9	1.41	0.45	1.0	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	1616	18.4	6.04	0.80	2.7	MC/OE
		Female	38	32	1076	18.1	5.95	0.80	2.7	MC/OE
	A	Male	12	9	1616	5.3	2.43	0.59	1.6	MC/OE
		Female	12	9	1076	5.1	2.46	0.62	1.5	MC/OE
	B	Male	5	5	1616	3.1	1.17	0.34	1.0	MC
		Female	5	5	1076	3.0	1.09	0.23	1.0	MC
	C	Male	7	7	1616	3.9	1.56	0.42	1.2	MC
		Female	7	7	1076	3.8	1.57	0.43	1.2	MC
	D	Male	7	7	1616	3.2	1.63	0.48	1.2	MC
		Female	7	7	1076	3.2	1.54	0.38	1.2	MC
	E	Male	7	4	1616	2.9	1.41	0.44	1.1	MC/OE
		Female	7	4	1076	2.9	1.42	0.46	1.0	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	1874	18.4	5.92	0.79	2.7	MC/OE
		Af. Amer.	38	32	511	17.9	6.20	0.81	2.7	MC/OE
		Hispanic	38	32	233	17.5	5.99	0.80	2.7	MC/OE
		Asian	38	32	31	19.8	5.96	0.78	2.8	MC/OE
		Am. Indian	38	32	7	18.6	6.08	0.78	2.8	MC/OE
		Multi	38	32	36	17.5	7.08	0.87	2.6	MC/OE
	A	White	12	9	1874	5.3	2.45	0.60	1.6	MC/OE
		Af. Amer.	12	9	511	5.2	2.43	0.62	1.5	MC/OE
		Hispanic	12	9	233	5.1	2.41	0.60	1.5	MC/OE
		Asian	12	9	31	5.6	2.06	0.39	1.6	MC/OE
		Am. Indian	12	9	7	4.7	1.98	0.15	1.8	MC/OE
		Multi	12	9	36	5.1	2.69	0.73	1.4	MC/OE
	B	White	5	5	1874	3.1	1.12	0.28	0.9	MC
		Af. Amer.	5	5	511	3.0	1.16	0.30	1.0	MC
		Hispanic	5	5	233	2.9	1.21	0.40	0.9	MC
		Asian	5	5	31	3.2	1.08	0.21	1.0	MC
		Am. Indian	5	5	7	3.6	1.27	0.48	0.9	MC
		Multi	5	5	36	3.0	1.20	0.28	1.0	MC
	C	White	7	7	1874	4.0	1.54	0.42	1.2	MC
		Af. Amer.	7	7	511	3.7	1.61	0.45	1.2	MC
		Hispanic	7	7	233	3.6	1.57	0.40	1.2	MC
		Asian	7	7	31	4.1	1.48	0.37	1.2	MC
		Am. Indian	7	7	7	3.9	1.35	0.03	1.3	MC
		Multi	7	7	36	3.9	1.63	0.45	1.2	MC
	D	White	7	7	1874	3.2	1.59	0.44	1.2	MC
		Af. Amer.	7	7	511	3.2	1.57	0.42	1.2	MC
		Hispanic	7	7	233	3.2	1.65	0.49	1.2	MC
		Asian	7	7	31	3.8	1.56	0.40	1.2	MC
		Am. Indian	7	7	7	3.4	2.07	0.66	1.2	MC
		Multi	7	7	36	2.8	1.70	0.54	1.2	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
	E	White	7	4	1874	2.9	1.39	0.44	1.0	MC/OE
		Af. Amer.	7	4	511	2.9	1.52	0.50	1.1	MC/OE
		Hispanic	7	4	233	2.7	1.35	0.39	1.1	MC/OE
		Asian	7	4	31	3.2	1.68	0.55	1.1	MC/OE
		Am. Indian	7	4	7	3.0	0.58	-2.10	1.0	MC/OE
		Multi	7	4	36	2.8	1.48	0.64	0.9	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
ELL	Tot.	All	38	32	99	18.1	5.98	0.80	2.7	MC/OE
	A	All	12	9	99	5.5	2.49	0.62	1.5	MC/OE
	B	All	5	5	99	3.0	1.08	0.21	1.0	MC
	C	All	7	7	99	3.5	1.48	0.31	1.2	MC
	D	All	7	7	99	3.4	1.65	0.50	1.2	MC
	E	All	7	4	99	2.7	1.27	0.30	1.1	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Eco. Disadv.	Tot.	All	38	32	1499	18.0	6.16	0.81	2.7	MC/OE
	A	All	12	9	1499	5.2	2.46	0.62	1.5	MC/OE
	B	All	5	5	1499	3.0	1.16	0.32	1.0	MC
	C	All	7	7	1499	3.8	1.58	0.43	1.2	MC
	D	All	7	7	1499	3.1	1.63	0.47	1.2	MC
	E	All	7	4	1499	2.9	1.43	0.46	1.1	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	2817	19.1	7.00	0.82	3.0	MC/OE
	A	All	9	6	2817	4.5	2.28	0.54	1.6	MC/OE
	B	All	5	5	2817	2.4	1.35	0.44	1.0	MC
	C	All	7	7	2817	3.9	1.66	0.50	1.2	MC
	D	All	11	8	2817	4.5	2.47	0.51	1.7	MC/OE
	E	All	6	6	2817	3.7	1.48	0.51	1.0	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	1647	19.3	7.12	0.82	3.0	MC/OE
		Female	38	32	1156	18.8	6.81	0.80	3.0	MC/OE
	A	Male	9	6	1647	4.7	2.25	0.52	1.6	MC/OE
		Female	9	6	1156	4.3	2.30	0.55	1.6	MC/OE
	B	Male	5	5	1647	2.4	1.36	0.44	1.0	MC
		Female	5	5	1156	2.3	1.34	0.45	1.0	MC
	C	Male	7	7	1647	4.0	1.68	0.52	1.2	MC
		Female	7	7	1156	3.9	1.63	0.48	1.2	MC
	D	Male	11	8	1647	4.5	2.54	0.53	1.7	MC/OE
		Female	11	8	1156	4.5	2.37	0.46	1.7	MC/OE
	E	Male	6	6	1647	3.7	1.49	0.50	1.0	MC
		Female	6	6	1156	3.8	1.46	0.51	1.0	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	1987	19.1	6.94	0.81	3.0	MC/OE
		Af. Amer.	38	32	511	18.5	6.96	0.81	3.0	MC/OE
		Hispanic	38	32	236	20.2	7.43	0.83	3.1	MC/OE
		Asian	38	32	36	19.7	6.27	0.78	3.0	MC/OE
		Am. Indian	38	32	3	22.7	5.03	0.67	2.9	MC/OE
		Multi	38	32	27	18.5	8.03	0.87	2.8	MC/OE
	A	White	9	6	1987	4.6	2.29	0.54	1.6	MC/OE
		Af. Amer.	9	6	511	4.4	2.22	0.52	1.5	MC/OE
		Hispanic	9	6	236	4.8	2.33	0.53	1.6	MC/OE
		Asian	9	6	36	4.7	2.30	0.66	1.3	MC/OE
		Am. Indian	9	6	3	5.7	1.53	0.34	1.2	MC/OE
		Multi	9	6	27	4.2	2.50	0.66	1.5	MC/OE
	B	White	5	5	1987	2.4	1.36	0.44	1.0	MC
		Af. Amer.	5	5	511	2.2	1.31	0.39	1.0	MC
		Hispanic	5	5	236	2.5	1.41	0.50	1.0	MC
		Asian	5	5	36	2.5	1.25	0.29	1.1	MC
		Am. Indian	5	5	3	3.0	1.00	-0.42	1.2	MC
		Multi	5	5	27	2.2	1.45	0.52	1.0	MC
	C	White	7	7	1987	4.0	1.64	0.49	1.2	MC
		Af. Amer.	7	7	511	3.7	1.71	0.52	1.2	MC
		Hispanic	7	7	236	4.1	1.74	0.56	1.2	MC
		Asian	7	7	36	3.8	1.68	0.49	1.2	MC
		Am. Indian	7	7	3	4.7	0.58	0.00	0.6	MC
		Multi	7	7	27	4.1	1.71	0.52	1.2	MC
	D	White	11	8	1987	4.4	2.44	0.50	1.7	MC/OE
		Af. Amer.	11	8	511	4.6	2.52	0.52	1.7	MC/OE
		Hispanic	11	8	236	4.8	2.62	0.51	1.8	MC/OE
		Asian	11	8	36	5.0	2.37	0.46	1.7	MC/OE
		Am. Indian	11	8	3	5.3	3.51	0.74	1.8	MC/OE
		Multi	11	8	27	4.4	2.37	0.56	1.6	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
	E	White	6	6	1987	3.8	1.47	0.51	1.0	MC
		Af. Amer.	6	6	511	3.5	1.50	0.49	1.1	MC
		Hispanic	6	6	236	4.0	1.47	0.53	1.0	MC
		Asian	6	6	36	3.7	1.33	0.41	1.0	MC
		Am. Indian	6	6	3	4.0	1.00	-0.40	1.2	MC
		Multi	6	6	27	3.7	1.69	0.65	1.0	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
ELL	Tot.	All	38	32	74	19.6	7.76	0.85	3.0	MC/OE
	A	All	9	6	74	4.8	2.41	0.55	1.6	MC/OE
	B	All	5	5	74	2.3	1.40	0.49	1.0	MC
	C	All	7	7	74	3.9	1.86	0.62	1.1	MC
	D	All	11	8	74	4.6	2.69	0.56	1.8	MC/OE
	E	All	6	6	74	4.0	1.53	0.58	1.0	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Eco. Disadv.	Tot.	All	38	32	1584	18.9	7.07	0.82	3.0	MC/OE
	A	All	9	6	1584	4.5	2.29	0.53	1.6	MC/OE
	B	All	5	5	1584	2.3	1.36	0.44	1.0	MC
	C	All	7	7	1584	3.9	1.70	0.52	1.2	MC
	D	All	11	8	1584	4.5	2.48	0.52	1.7	MC/OE
	E	All	6	6	1584	3.7	1.49	0.51	1.0	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	3019	18.9	6.41	0.81	2.8	MC/OE
	A	All	8	5	3019	3.5	1.73	0.45	1.3	MC/OE
	B	All	5	5	3019	3.0	1.28	0.41	1.0	MC
	C	All	7	7	3019	3.6	1.70	0.51	1.2	MC
	D	All	11	8	3019	4.4	2.26	0.53	1.5	MC/OE
	E	All	7	7	3019	4.4	1.65	0.52	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	1803	18.9	6.53	0.82	2.8	MC/OE
		Female	38	32	1196	19.0	6.22	0.80	2.8	MC/OE
	A	Male	8	5	1803	3.5	1.77	0.46	1.3	MC/OE
		Female	8	5	1196	3.5	1.67	0.45	1.2	MC/OE
	B	Male	5	5	1803	3.0	1.30	0.43	1.0	MC
		Female	5	5	1196	3.0	1.25	0.39	1.0	MC
	C	Male	7	7	1803	3.6	1.73	0.53	1.2	MC
		Female	7	7	1196	3.6	1.65	0.48	1.2	MC
	D	Male	11	8	1803	4.4	2.26	0.53	1.5	MC/OE
		Female	11	8	1196	4.5	2.26	0.54	1.5	MC/OE
	E	Male	7	7	1803	4.4	1.68	0.54	1.1	MC
		Female	7	7	1196	4.4	1.61	0.51	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	2138	19.1	6.38	0.81	2.8	MC/OE
		Af. Amer.	38	32	554	18.0	6.36	0.81	2.8	MC/OE
		Hispanic	38	32	251	19.8	6.39	0.80	2.9	MC/OE
		Asian	38	32	22	18.0	7.04	0.86	2.6	MC/OE
		Am. Indian	38	32	8	20.9	7.45	0.82	3.1	MC/OE
		Multi	38	32	26	17.8	6.88	0.83	2.9	MC/OE
	A	White	8	5	2138	3.5	1.73	0.47	1.3	MC/OE
		Af. Amer.	8	5	554	3.5	1.67	0.40	1.3	MC/OE
		Hispanic	8	5	251	3.8	1.78	0.42	1.4	MC/OE
		Asian	8	5	22	3.0	1.46	0.31	1.2	MC/OE
		Am. Indian	8	5	8	3.9	2.36	0.64	1.4	MC/OE
		Multi	8	5	26	3.8	1.92	0.49	1.4	MC/OE
	B	White	5	5	2138	3.0	1.28	0.41	1.0	MC
		Af. Amer.	5	5	554	2.9	1.29	0.42	1.0	MC
		Hispanic	5	5	251	3.0	1.28	0.43	1.0	MC
		Asian	5	5	22	2.9	1.32	0.51	0.9	MC
		Am. Indian	5	5	8	3.0	1.20	0.22	1.1	MC
		Multi	5	5	26	3.0	1.33	0.44	1.0	MC
	C	White	7	7	2138	3.6	1.72	0.53	1.2	MC
		Af. Amer.	7	7	554	3.4	1.66	0.47	1.2	MC
		Hispanic	7	7	251	3.7	1.62	0.43	1.2	MC
		Asian	7	7	22	3.8	1.63	0.45	1.2	MC
		Am. Indian	7	7	8	4.4	2.13	0.77	1.0	MC
		Multi	7	7	26	3.0	1.46	0.29	1.2	MC
	D	White	11	8	2138	4.5	2.26	0.53	1.5	MC/OE
		Af. Amer.	11	8	554	4.0	2.19	0.52	1.5	MC/OE
		Hispanic	11	8	251	4.8	2.30	0.52	1.6	MC/OE
		Asian	11	8	22	4.3	2.34	0.64	1.4	MC/OE
		Am. Indian	11	8	8	4.6	2.56	0.37	2.0	MC/OE
		Multi	11	8	26	4.1	2.25	0.49	1.6	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
	E	White	7	7	2138	4.5	1.63	0.51	1.1	MC
		Af. Amer.	7	7	554	4.2	1.70	0.55	1.1	MC
		Hispanic	7	7	251	4.5	1.69	0.57	1.1	MC
		Asian	7	7	22	4.1	1.80	0.60	1.1	MC
		Am. Indian	7	7	8	5.0	1.41	0.42	1.1	MC
		Multi	7	7	26	4.0	1.70	0.55	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
ELL	Tot.	All	38	32	76	20.2	5.78	0.74	3.0	MC/OE
	A	All	8	5	76	4.0	1.64	0.28	1.4	MC/OE
	B	All	5	5	76	3.1	1.30	0.45	1.0	MC
	C	All	7	7	76	3.7	1.55	0.38	1.2	MC
	D	All	11	8	76	4.9	2.37	0.48	1.7	MC/OE
	E	All	7	7	76	4.5	1.65	0.54	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Eco. Disadv.	Tot.	All	38	32	1621	18.8	6.49	0.81	2.8	MC/OE
	A	All	8	5	1621	3.6	1.74	0.45	1.3	MC/OE
	B	All	5	5	1621	2.9	1.27	0.40	1.0	MC
	C	All	7	7	1621	3.5	1.69	0.51	1.2	MC
	D	All	11	8	1621	4.4	2.28	0.54	1.5	MC/OE
	E	All	7	7	1621	4.4	1.67	0.54	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Overall	Tot.	All	38	32	3536	17.0	6.27	0.81	2.7	MC/OE
	A	All	5	5	3536	3.0	1.27	0.40	1.0	MC
	B	All	5	2	3536	1.4	1.08	0.23	1.0	MC/OE
	C	All	6	6	3536	3.3	1.41	0.37	1.1	MC
	D	All	16	13	3536	6.4	3.11	0.69	1.7	MC/OE
	E	All	6	6	3536	2.9	1.45	0.43	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Gender	Tot.	Male	38	32	2163	17.1	6.24	0.81	2.7	MC/OE
		Female	38	32	1332	16.8	6.31	0.82	2.7	MC/OE
	A	Male	5	5	2163	3.0	1.27	0.41	1.0	MC
		Female	5	5	1332	2.9	1.26	0.40	1.0	MC
	B	Male	5	2	2163	1.4	1.07	0.24	0.9	MC/OE
		Female	5	2	1332	1.4	1.11	0.24	1.0	MC/OE
	C	Male	6	6	2163	3.4	1.41	0.37	1.1	MC
		Female	6	6	1332	3.2	1.39	0.36	1.1	MC
	D	Male	16	13	2163	6.3	3.10	0.69	1.7	MC/OE
		Female	16	13	1332	6.5	3.12	0.70	1.7	MC/OE
	E	Male	6	6	2163	2.9	1.44	0.42	1.1	MC
		Female	6	6	1332	2.8	1.45	0.44	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Ethnicity	Tot.	White	38	32	2496	17.5	6.30	0.81	2.7	MC/OE
		Af. Amer.	38	32	700	15.2	5.99	0.80	2.7	MC/OE
		Hispanic	38	32	232	16.9	5.89	0.79	2.7	MC/OE
		Asian	38	32	27	18.6	6.07	0.80	2.7	MC/OE
		Am. Indian	38	32	5	15.2	5.36	0.78	2.5	MC/OE
		Multi	38	32	36	14.7	5.03	0.71	2.7	MC/OE
	A	White	5	5	2496	3.0	1.28	0.42	1.0	MC
		Af. Amer.	5	5	700	2.7	1.24	0.35	1.0	MC
		Hispanic	5	5	232	3.0	1.22	0.33	1.0	MC
		Asian	5	5	27	3.0	1.40	0.54	0.9	MC
		Am. Indian	5	5	5	2.2	1.48	0.63	0.9	MC
		Multi	5	5	36	2.6	1.08	0.01	1.1	MC
	B	White	5	2	2496	1.4	1.12	0.24	1.0	MC/OE
		Af. Amer.	5	2	700	1.2	1.01	0.25	0.9	MC/OE
		Hispanic	5	2	232	1.3	0.97	0.12	0.9	MC/OE
		Asian	5	2	27	1.6	0.70	-0.09	0.7	MC/OE
		Am. Indian	5	2	5	1.6	0.55	-0.67	0.7	MC/OE
		Multi	5	2	36	1.3	0.94	0.07	0.9	MC/OE
	C	White	6	6	2496	3.4	1.40	0.37	1.1	MC
		Af. Amer.	6	6	700	2.9	1.37	0.33	1.1	MC
		Hispanic	6	6	232	3.2	1.37	0.34	1.1	MC
		Asian	6	6	27	3.0	1.53	0.48	1.1	MC
		Am. Indian	6	6	5	2.6	0.55	-3.60	1.2	MC
		Multi	6	6	36	2.9	1.35	0.32	1.1	MC
	D	White	16	13	2496	6.6	3.11	0.69	1.7	MC/OE
		Af. Amer.	16	13	700	5.7	3.04	0.68	1.7	MC/OE
		Hispanic	16	13	232	6.6	3.07	0.69	1.7	MC/OE
		Asian	16	13	27	7.4	3.15	0.67	1.8	MC/OE
		Am. Indian	16	13	5	5.0	3.74	0.83	1.6	MC/OE
		Multi	16	13	36	5.4	2.61	0.58	1.7	MC/OE

Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
E	White	6	6	2496	3.0	1.46	0.44	1.1	MC
	Af. Amer.	6	6	700	2.6	1.38	0.37	1.1	MC
	Hispanic	6	6	232	2.8	1.39	0.39	1.1	MC
	Asian	6	6	27	3.6	1.25	0.07	1.2	MC
	Am. Indian	6	6	5	3.8	0.45	-4.20	1.0	MC
	Multi	6	6	36	2.6	1.20	0.07	1.2	MC

ELL

Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Tot.	All	38	32	25	14.6	6.74	0.85	2.6	MC/OE
A	All	5	5	25	2.8	1.36	0.48	1.0	MC
B	All	5	2	25	1.2	1.08	0.17	1.0	MC/OE
C	All	6	6	25	2.9	1.38	0.36	1.1	MC
D	All	16	13	25	5.2	3.16	0.73	1.6	MC/OE
E	All	6	6	25	2.5	1.50	0.51	1.1	MC

Eco. Disadv.

Strand	Group	Pts.	Len.	N	Mean	*SD*	r	*SEM*	Items
Tot.	All	38	32	1687	16.2	6.17	0.81	2.7	MC/OE
A	All	5	5	1687	2.9	1.28	0.40	1.0	MC
B	All	5	2	1687	1.3	1.05	0.25	0.9	MC/OE
C	All	6	6	1687	3.2	1.42	0.39	1.1	MC
D	All	16	13	1687	6.1	3.07	0.68	1.7	MC/OE
E	All	6	6	1687	2.7	1.40	0.39	1.1	MC

Appendix K:

Cut Scores and Transformations

	Grade	Scaling	LOSS	Scaled Score Cuts			Logit Cuts		
				Basic	Prof.	Adv.	Basic	Prof.	Adv.
Mathematics	4	84.31X + 1199.67	1075	1150	1275	1356	-0.5891	0.8935	1.8540
	5	89.34X + 1197.81	1075	1150	1275	1374	-0.5352	0.8640	1.9734
	6	95.81X + 1242.03	1075	1150	1275	1381	-0.9606	0.3441	1.4543
	7	87.29X + 1223.69	1075	1150	1275	1364	-0.8442	0.5878	1.6086
	8	94.23X + 1224.05	1075	1150	1275	1395	-0.7858	0.5407	1.8139
	11	115.64X + 1213.51	1075	1150	1275	1403	-0.5492	0.5317	1.6389

Appendix L:

Raw-to-Scaled Scores

Appendix L: Raw-to-Scaled Scores
Modified Mathematics Grade 4

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.1572	1.8441	1075	155	0	0.0	0	0.0	0
1	-3.9059	1.0335	1075	87	0	0.0	0	0.0	0
2	-3.1438	0.7548	1075	64	0	0.0	0	0.0	0
3	-2.6680	0.6360	1075	54	0	0.0	0	0.0	0
4	-2.3084	0.5679	1075	48	4	0.2	4	0.2	1
5	-2.0121	0.5233	1075	44	1	0.0	5	0.2	1
6	-1.7552	0.4918	1075	41	4	0.2	9	0.4	1
7	-1.5252	0.4683	1075	39	9	0.4	18	0.8	1
8	-1.3146	0.4502	1089	38	14	0.6	32	1.5	1
9	-1.1185	0.4359	1105	37	19	0.9	51	2.4	2
10	-0.9337	0.4243	1121	36	23	1.1	74	3.4	3
11	-0.7579	0.4148	1136	35	30	1.4	104	4.8	4
12	-0.5891	0.4069	1150	34	33	1.5	137	6.3	6
13	-0.4263	0.4003	1164	34	50	2.3	187	8.6	7
14	-0.2684	0.3948	1177	33	52	2.4	239	11.0	10
15	-0.1143	0.3902	1190	33	73	3.4	312	14.4	13
16	0.0363	0.3863	1203	33	73	3.4	385	17.8	16
17	0.1842	0.3829	1215	32	81	3.7	466	21.5	20
18	0.3297	0.3801	1227	32	116	5.3	582	26.8	24
19	0.4732	0.3776	1240	32	91	4.2	673	31.0	29
20	0.6149	0.3753	1252	32	108	5.0	781	36.0	34
21	0.7550	0.3732	1263	31	99	4.6	880	40.6	38
22	0.8935	0.3712	1275	31	116	5.3	996	45.9	43
23	1.0306	0.3693	1287	31	121	5.6	1117	51.5	49
24	1.1664	0.3678	1298	31	124	5.7	1241	57.2	54
25	1.3014	0.3671	1309	31	128	5.9	1369	63.1	60
26	1.4363	0.3676	1321	31	118	5.4	1487	68.6	66
27	1.5721	0.3700	1332	31	113	5.2	1600	73.8	71
28	1.7107	0.3749	1344	32	110	5.1	1710	78.8	76
29	1.8540	0.3829	1356	32	97	4.5	1807	83.3	81
30	2.0050	0.3951	1369	33	84	3.9	1891	87.2	85
31	2.1676	0.4122	1382	35	59	2.7	1950	89.9	89
32	2.3470	0.4359	1398	37	70	3.2	2020	93.1	92
33	2.5508	0.4687	1415	40	51	2.4	2071	95.5	94
34	2.7916	0.5154	1435	43	40	1.8	2111	97.3	96
35	3.0926	0.5867	1460	49	30	1.4	2141	98.7	98
36	3.5062	0.7109	1495	60	17	0.8	2158	99.5	99
37	4.2018	0.9998	1554	84	9	0.4	2167	99.9	99
38	5.4042	1.8247	1655	154	2	0.1	2169	100.0	99

Appendix L: Raw-to-Scaled Scores
Modified Mathematics Grade 5

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.3945	1.8838	1075	168	0	0.0	0	0.0	0
1	-4.0515	1.0871	1075	97	0	0.0	0	0.0	0
2	-3.1983	0.8004	1075	72	0	0.0	0	0.0	0
3	-2.6665	0.6689	1075	60	0	0.0	0	0.0	0
4	-2.2733	0.5901	1075	53	0	0.0	0	0.0	0
5	-1.9571	0.5372	1075	48	1	0.0	1	0.0	1
6	-1.6895	0.4993	1075	45	6	0.2	7	0.3	1
7	-1.4546	0.4711	1075	42	8	0.3	15	0.6	1
8	-1.2432	0.4493	1087	40	13	0.5	28	1.1	1
9	-1.0491	0.4323	1104	39	29	1.1	57	2.2	2
10	-0.8683	0.4187	1120	37	30	1.2	87	3.4	3
11	-0.6977	0.4078	1135	36	53	2.1	140	5.5	4
12	-0.5352	0.3989	1150	36	52	2.0	192	7.5	7
13	-0.3790	0.3917	1164	35	75	2.9	267	10.5	9
14	-0.2280	0.3857	1177	34	86	3.4	353	13.8	12
15	-0.0813	0.3806	1191	34	101	4.0	454	17.8	16
16	0.0619	0.3762	1203	34	99	3.9	553	21.7	20
17	0.2019	0.3722	1216	33	120	4.7	673	26.4	24
18	0.3391	0.3686	1228	33	114	4.5	787	30.8	29
19	0.4737	0.3651	1240	33	140	5.5	927	36.3	34
20	0.6058	0.3619	1252	32	165	6.5	1092	42.8	40
21	0.7357	0.3592	1264	32	156	6.1	1248	48.9	46
22	0.8640	0.3573	1275	32	142	5.6	1390	54.5	52
23	0.9914	0.3567	1286	32	137	5.4	1527	59.8	57
24	1.1188	0.3576	1298	32	118	4.6	1645	64.5	62
25	1.2476	0.3606	1309	32	111	4.3	1756	68.8	67
26	1.3794	0.3658	1321	33	129	5.1	1885	73.9	71
27	1.5159	0.3734	1333	33	122	4.8	2007	78.6	76
28	1.6589	0.3834	1346	34	107	4.2	2114	82.8	81
29	1.8106	0.3960	1360	35	105	4.1	2219	87.0	85
30	1.9734	0.4116	1374	37	82	3.2	2301	90.2	89
31	2.1508	0.4314	1390	39	67	2.6	2368	92.8	91
32	2.3478	0.4575	1408	41	59	2.3	2427	95.1	94
33	2.5731	0.4937	1428	44	36	1.4	2463	96.5	96
34	2.8419	0.5464	1452	49	36	1.4	2499	97.9	97
35	3.1832	0.6274	1482	56	31	1.2	2530	99.1	99
36	3.6594	0.7639	1525	68	14	0.5	2544	99.7	99
37	4.4539	1.0590	1596	95	8	0.3	2552	100.0	99
38	5.7542	1.8676	1712	167	0	0.0	2552	100.0	100

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.1975	1.8528	1075	178	0	0.0	0	0.0	0
1	-3.9264	1.0451	1075	100	0	0.0	0	0.0	0
2	-3.1453	0.7642	1075	73	0	0.0	0	0.0	0
3	-2.6589	0.6415	1075	61	0	0.0	0	0.0	0
4	-2.2952	0.5694	1075	55	6	0.2	6	0.2	1
5	-1.9992	0.5212	1075	50	6	0.2	12	0.4	1
6	-1.7462	0.4864	1075	47	19	0.7	31	1.1	1
7	-1.5227	0.4602	1096	44	33	1.2	64	2.4	2
8	-1.3205	0.4398	1116	42	51	1.9	115	4.3	3
9	-1.1344	0.4236	1133	41	82	3.0	197	7.3	6
10	-0.9606	0.4106	1150	39	87	3.2	284	10.5	9
11	-0.7964	0.4000	1166	38	90	3.3	374	13.9	12
12	-0.6399	0.3915	1181	38	129	4.8	503	18.6	16
13	-0.4895	0.3845	1195	37	136	5.0	639	23.7	21
14	-0.3438	0.3789	1209	36	151	5.6	790	29.3	26
15	-0.2020	0.3746	1223	36	132	4.9	922	34.1	32
16	-0.0629	0.3713	1236	36	144	5.3	1066	39.5	37
17	0.0740	0.3689	1249	35	181	6.7	1247	46.2	43
18	0.2095	0.3674	1262	35	154	5.7	1401	51.9	49
19	0.3441	0.3667	1275	35	170	6.3	1571	58.2	55
20	0.4785	0.3667	1288	35	164	6.1	1735	64.3	61
21	0.6132	0.3674	1301	35	157	5.8	1892	70.1	67
22	0.7486	0.3687	1314	35	132	4.9	2024	75.0	73
23	0.8852	0.3707	1327	36	130	4.8	2154	79.8	77
24	1.0235	0.3733	1340	36	117	4.3	2271	84.1	82
25	1.1641	0.3766	1354	36	93	3.4	2364	87.6	86
26	1.3074	0.3807	1367	36	88	3.3	2452	90.8	89
27	1.4543	0.3860	1381	37	69	2.6	2521	93.4	92
28	1.6057	0.3928	1396	38	51	1.9	2572	95.3	94
29	1.7634	0.4017	1411	38	33	1.2	2605	96.5	96
30	1.9295	0.4139	1427	40	37	1.4	2642	97.9	97
31	2.1075	0.4307	1444	41	18	0.7	2660	98.5	98
32	2.3026	0.4540	1463	43	15	0.6	2675	99.1	99
33	2.5232	0.4872	1484	47	9	0.3	2684	99.4	99
34	2.7832	0.5354	1509	51	9	0.3	2693	99.7	99
35	3.1080	0.6094	1540	58	4	0.1	2697	99.9	99
36	3.5534	0.7364	1582	71	2	0.1	2699	100.0	99
37	4.2925	1.0248	1653	98	1	0.0	2700	100.0	99
38	5.5343	1.8417	1772	176	0	0.0	2700	100.0	100

Appendix L: Raw-to-Scaled Scores
Modified Mathematics Grade 7

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9183	1.8423	1075	161	0	0.0	0	0.0	0
1	-3.6721	1.0296	1075	90	0	0.0	0	0.0	0
2	-2.9193	0.7479	1075	65	0	0.0	0	0.0	0
3	-2.4548	0.6261	1075	55	4	0.1	4	0.1	1
4	-2.1086	0.5552	1075	48	5	0.2	9	0.3	1
5	-1.8274	0.5079	1075	44	14	0.5	23	0.8	1
6	-1.5872	0.4737	1085	41	30	1.1	53	1.9	1
7	-1.3754	0.4479	1104	39	52	1.8	105	3.7	3
8	-1.1840	0.4277	1120	37	61	2.2	166	5.9	5
9	-1.0081	0.4116	1136	36	69	2.4	235	8.3	7
10	-0.8442	0.3985	1150	35	93	3.3	328	11.6	10
11	-0.6898	0.3876	1163	34	112	4.0	440	15.6	14
12	-0.5431	0.3787	1176	33	99	3.5	539	19.1	17
13	-0.4026	0.3711	1189	32	154	5.5	693	24.6	22
14	-0.2673	0.3648	1200	32	129	4.6	822	29.2	27
15	-0.1363	0.3594	1212	31	141	5.0	963	34.2	32
16	-0.0088	0.3547	1223	31	116	4.1	1079	38.3	36
17	0.1155	0.3506	1234	31	158	5.6	1237	43.9	41
18	0.2371	0.3469	1244	30	147	5.2	1384	49.1	47
19	0.3562	0.3435	1255	30	137	4.9	1521	54.0	52
20	0.4731	0.3403	1265	30	135	4.8	1656	58.8	56
21	0.5878	0.3372	1275	29	133	4.7	1789	63.5	61
22	0.7006	0.3344	1285	29	130	4.6	1919	68.1	66
23	0.8115	0.3318	1295	29	119	4.2	2038	72.3	70
24	0.9210	0.3300	1304	29	106	3.8	2144	76.1	74
25	1.0296	0.3292	1314	29	108	3.8	2252	79.9	78
26	1.1381	0.3301	1323	29	99	3.5	2351	83.5	82
27	1.2481	0.3335	1333	29	92	3.3	2443	86.7	85
28	1.3613	0.3400	1343	30	92	3.3	2535	90.0	88
29	1.4803	0.3507	1353	31	55	2.0	2590	91.9	91
30	1.6086	0.3666	1364	32	55	2.0	2645	93.9	93
31	1.7509	0.3890	1377	34	55	2.0	2700	95.8	95
32	1.9138	0.4197	1391	37	40	1.4	2740	97.3	97
33	2.1070	0.4610	1408	40	27	1.0	2767	98.2	98
34	2.3450	0.5170	1428	45	12	0.4	2779	98.7	98
35	2.6528	0.5969	1455	52	23	0.8	2802	99.5	99
36	3.0838	0.7267	1493	63	8	0.3	2810	99.8	99
37	3.8072	1.0159	1556	89	4	0.1	2814	99.9	99
38	5.0341	1.8350	1663	160	3	0.1	2817	100.0	99

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9775	1.8405	1075	173	0	0.0	0	0.0	0
1	-3.7357	1.0266	1075	97	0	0.0	0	0.0	0
2	-2.9888	0.7441	1075	70	1	0.0	1	0.0	1
3	-2.5296	0.6221	1075	59	3	0.1	4	0.1	1
4	-2.1882	0.5512	1075	52	2	0.1	6	0.2	1
5	-1.9112	0.5041	1075	48	13	0.4	19	0.6	1
6	-1.6745	0.4703	1075	44	25	0.8	44	1.5	1
7	-1.4655	0.4450	1086	42	40	1.3	84	2.8	2
8	-1.2764	0.4253	1104	40	46	1.5	130	4.3	4
9	-1.1023	0.4098	1120	39	71	2.4	201	6.7	5
10	-0.9396	0.3973	1136	37	102	3.4	303	10.0	8
11	-0.7858	0.3872	1150	36	113	3.7	416	13.8	12
12	-0.6391	0.3791	1164	36	119	3.9	535	17.7	16
13	-0.4980	0.3725	1177	35	135	4.5	670	22.2	20
14	-0.3612	0.3673	1190	35	142	4.7	812	26.9	25
15	-0.2279	0.3632	1203	34	162	5.4	974	32.3	30
16	-0.0971	0.3602	1215	34	160	5.3	1134	37.6	35
17	0.0318	0.3580	1227	34	148	4.9	1282	42.5	40
18	0.1594	0.3567	1239	34	179	5.9	1461	48.4	45
19	0.2865	0.3562	1251	34	158	5.2	1619	53.6	51
20	0.4133	0.3564	1263	34	169	5.6	1788	59.2	56
21	0.5407	0.3573	1275	34	162	5.4	1950	64.6	62
22	0.6689	0.3590	1287	34	172	5.7	2122	70.3	67
23	0.7987	0.3615	1299	34	156	5.2	2278	75.5	73
24	0.9305	0.3648	1312	34	145	4.8	2423	80.3	78
25	1.0650	0.3690	1324	35	94	3.1	2517	83.4	82
26	1.2030	0.3743	1337	35	111	3.7	2628	87.0	85
27	1.3456	0.3809	1351	36	93	3.1	2721	90.1	89
28	1.4937	0.3892	1365	37	77	2.6	2798	92.7	91
29	1.6491	0.3996	1379	38	52	1.7	2850	94.4	94
30	1.8139	0.4128	1395	39	47	1.6	2897	96.0	95
31	1.9911	0.4300	1412	41	43	1.4	2940	97.4	97
32	2.1855	0.4529	1430	43	35	1.2	2975	98.5	98
33	2.4045	0.4847	1451	46	17	0.6	2992	99.1	99
34	2.6610	0.5307	1475	50	11	0.4	3003	99.5	99
35	2.9790	0.6022	1505	57	9	0.3	3012	99.8	99
36	3.4134	0.7272	1546	69	6	0.2	3018	100.0	99
37	4.1366	1.0159	1614	96	1	0.0	3019	100.0	99
38	5.3649	1.8361	1730	173	0	0.0	3019	100.0	100

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9393	1.8430	1075	213	2	0.1	2	0.1	1
1	-3.6912	1.0307	1075	119	0	0.0	2	0.1	1
2	-2.9364	0.7490	1075	87	1	0.0	3	0.1	1
3	-2.4705	0.6271	1075	73	6	0.2	9	0.3	1
4	-2.1233	0.5560	1075	64	7	0.2	16	0.5	1
5	-1.8414	0.5085	1075	59	21	0.6	37	1.0	1
6	-1.6006	0.4744	1075	55	49	1.4	86	2.4	2
7	-1.3881	0.4486	1075	52	71	2.0	157	4.4	3
8	-1.1960	0.4287	1075	50	130	3.7	287	8.1	6
9	-1.0192	0.4128	1096	48	142	4.0	429	12.1	10
10	-0.8541	0.4002	1115	46	144	4.1	573	16.2	14
11	-0.6981	0.3901	1133	45	188	5.3	761	21.5	19
12	-0.5492	0.3820	1150	44	182	5.1	943	26.7	24
13	-0.4058	0.3756	1167	43	209	5.9	1152	32.6	30
14	-0.2666	0.3708	1183	43	209	5.9	1361	38.5	36
15	-0.1305	0.3672	1198	42	222	6.3	1583	44.8	42
16	0.0033	0.3647	1214	42	188	5.3	1771	50.1	47
17	0.1358	0.3633	1229	42	190	5.4	1961	55.5	53
18	0.2676	0.3629	1244	42	193	5.5	2154	60.9	58
19	0.3994	0.3632	1260	42	208	5.9	2362	66.8	64
20	0.5317	0.3643	1275	42	164	4.6	2526	71.4	69
21	0.6649	0.3659	1290	42	166	4.7	2692	76.1	74
22	0.7995	0.3679	1306	43	141	4.0	2833	80.1	78
23	0.9357	0.3701	1322	43	112	3.2	2945	83.3	82
24	1.0734	0.3721	1338	43	115	3.3	3060	86.5	85
25	1.2127	0.3740	1354	43	115	3.3	3175	89.8	88
26	1.3532	0.3758	1370	43	95	2.7	3270	92.5	91
27	1.4951	0.3778	1386	44	67	1.9	3337	94.4	93
28	1.6389	0.3808	1403	44	52	1.5	3389	95.8	95
29	1.7857	0.3860	1420	45	35	1.0	3424	96.8	96
30	1.9378	0.3949	1438	46	38	1.1	3462	97.9	97
31	2.0993	0.4097	1456	47	27	0.8	3489	98.7	98
32	2.2761	0.4331	1477	50	17	0.5	3506	99.2	99
33	2.4787	0.4695	1500	54	15	0.4	3521	99.6	99
34	2.7250	0.5268	1529	61	8	0.2	3529	99.8	99
35	3.0500	0.6202	1566	72	4	0.1	3533	99.9	99
36	3.5344	0.7855	1622	91	2	0.1	3535	100.0	99
37	4.4122	1.1261	1724	130	1	0.0	3536	100.0	99
38	5.8476	1.9280	1890	223	0	0.0	3536	100.0	100