

TECHNICAL REPORT



**for the
2011 Modified Pennsylvania
System of School Assessment**

**Provided by
Data Recognition Corporation**

Table of Contents

Glossary of Common Terms	<i>i</i>
Preface: An Overview of Modified Assessments from 2008 to the Present.....	<i>ix</i>
Assessment Activities Occurring in the 2008–09 School Year	<i>ix</i>
Assessment Activities Occurring in the 2009–10 School Year	<i>x</i>
Assessment Activities Planned for the 2010–11 School Year	<i>xi</i>
Assessment Activities Planned for the 2011–12 School Year	<i>xii</i>
Chapter One: Background of the Modified Pennsylvania System of School Assessment.....	<i>1</i>
State and Federal Regulations Affecting the PSSA.....	<i>1</i>
Purposes of the PSSA.....	<i>1</i>
Changes in 2005 and Beyond.....	<i>2</i>
Students with Complex Support Needs: Alternate Assessment	<i>2</i>
Students with Disabilities Needing a Modified Approach: Modified Assessment	<i>3</i>
Chapter Two: Test Development Overview of the Modified PSSA.....	<i>5</i>
Overview of the Development Process	<i>5</i>
Academic Standards, Assessment Anchor Content Standards, and Eligible Content	<i>5</i>
Chapter Three: Item Development Process	<i>13</i>
Steps in the Development Process.....	<i>13</i>
Summary of Revision and/or Enhancement Guidelines.....	<i>19</i>
Item Authoring and Tracking	<i>20</i>
Internal Reviews and PDE Reviews.....	<i>20</i>
Reading Pilot	<i>23</i>
Cognitive Interviews	<i>25</i>
Test Content Blueprint for 2011 PSSA-M Assessments.....	<i>28</i>
Test Development Considerations for the PSSA-M.....	<i>37</i>
Test Development Process	<i>39</i>
Chapter Four: Universal Design Procedures Applied in the Modified PSSA	
Test Development Process.....	<i>41</i>
Elements of Universally Designed Assessments.....	<i>41</i>
Guidelines for Universally Designed Items	<i>43</i>
Item Development	<i>44</i>
Item Formatting	<i>45</i>
Assessment Accommodations	<i>46</i>
Chapter Five: Field Test Leading to the 2010 Core	<i>47</i>
Standalone Field Test Items	<i>47</i>
Statistical Analysis of Item Data.....	<i>48</i>
Review of Items with Data.....	<i>48</i>
Chapter Six: Operational Forms Construction for 2010	<i>51</i>
Final Selection of Items and 2011 PSSA-M Forms Construction.....	<i>51</i>
Linking the 2010 Operational Test to the 2011 Operational Test.....	<i>52</i>
Linking the 2011 Operational Test to the 2012 Operational Test.....	<i>52</i>
Special Forms Used in the 2010 PSSA-M	<i>53</i>

Table of Contents

Chapter Seven: Test Administration Procedures.....	55
Test Sessions, Test Sections, and Test Timing.....	55
Testing Window	58
Shipping, Packaging, and Delivery of Materials.....	58
Materials Returned	59
Test Security Measures	59
Sample Manuals	59
Testing Window Assessment Accommodations	59
Chapter Eight: Processing and Scoring.....	61
Receipt of Materials	61
Scanning of Materials.....	62
Materials Storage.....	65
Scoring Multiple-Choice Items	66
Rangefinding	66
Rater Recruitment/Qualifications.....	67
Leadership Recruitment/Qualifications.....	67
Training	68
Handscoring Process	69
Handscoring Validity Process	69
Quality Control.....	71
Chapter Nine: Description of Data Sources and Sampling Adequacy.....	77
Primary Student Filtering Criteria.....	77
Key Validation Data.....	78
Calibration Data	78
Final Data	78
Final N-Counts for all Data Sources.....	79
Chapter Ten: Summary Demographic, Program, and Accommodation Data for the 2011 PSSA Modified.....	81
Assessed Students.....	81
Composition of Sample Used in Subsequent Tables.....	84
Collection of Student Demographic Information	84
Demographic Characteristics.....	84
Test Accommodations Provided.....	89
Presentation Accommodations Received	89
Response Accommodations Received.....	89
Setting Accommodations Received.....	90
Timing Accommodations Received	90
Accommodation Rate	96
The Incidence of Accommodations and ELL Status	97
Glossary of Accommodations Terms	101

Table of Contents

Chapter Eleven: Classical Item Statistics	105
Item-Level Statistics	105
Item Difficulty	105
Item Discrimination	106
Discrimination on Difficulty Scatterplots	106
Observations and Interpretations	107
Item Omit Rates	113
Chapter Twelve: Rasch Item Calibration.....	115
Description of the Rasch Model	115
Checking Rasch Assumptions	116
Rasch Item Statistics	122
Visualizing the <i>P</i> -Value-Logit Relationship	124
Chapter Thirteen: Performance Level Setting.....	137
Summary	137
PSSA-M Cut Scores	139
Chapter Fourteen: Scaling.....	141
Scaled Scores.....	141
Raw-Score to Scaled-Score Tables	142
Strand (Reporting Category) Score Strength Profile	144
Chapter Fifteen: Linking.....	145
Forward	145
Introduction	145
Brief Summary of the PSSA-M Linking Procedure.....	145
PSSA-M Mathematics.....	147
Linking Method for PSSA-M Mathematics	148
Results Summary.....	148
Visualization Supplement.....	149
Chapter Sixteen: Scores and Score Reports.....	157
Scoring the PSSA-M	157
Description of Total Test Scores	157
Description of Strand (Reporting Category) Scores.....	160
Appropriate Score Uses.....	161
Cautions for Score Use.....	161
Reports.....	163
Chapter Seventeen: Operational Test Statistics.....	171
Performance Level Statistics	171
Scaled Scores.....	171
Raw Scores.....	172

Table of Contents

Chapter Eighteen: Reliability	177
Reliability Indices.....	178
Coefficient Alpha	178
Further Interpretations	180
Standard Error of Measurement	183
Rasch Conditional Standard Errors of Measurement	186
Decision Consistency	192
Rater Agreement.....	195
Chapter Nineteen: Validity	197
Purposes and Intended Uses of the PSSA-M.....	197
Evidence Based on Test Content.....	198
Evidence Based on Response Processes.....	200
Evidence Based on Internal Structure	200
Evidence Based on Consequences of Testing	212
Evidence Related to the Use of the Rasch Model	214
Validity Evidence Summary.....	214
References	215

Appendix A.	Assessment Anchor Explanations
Appendix B.	PSSA and PSSA-M General Scoring Guidelines
Appendix C.	2011 PSSA-M Tally Sheets; Comparing the 2011 PSSA Core with the 2011 PSSA-M Core
Appendix D.	Item and Test Development Process
Appendix E.	PSSA-M Item Review Cards
Appendix F.	Item Rating Sheet and Item Review Criteria Guidelines
Appendix G.	2011 Test Book Section Layout Plans
Appendix H.	Mean Raw Scores by Form
Appendix I.	Item Statistics
Appendix J.	Linking Item Statistics
Appendix K.	Reliabilities
Appendix L.	Cut Scores and Scale Transformations
Appendix M.	PSSA-M Historical Statistics
Appendix N.	Raw-to-Scaled Scores Conversion Tables
Appendix O.	PSSA and PSSA-M Demographics Comparison

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 Rasch scaling, ability is a generic term indicating the level of an individual on the construct measured by an exam. As an example for the PSSA, a student’s reading ability is measured by how the student performed on the PSSA Reading test. A student who answered more items correctly has a higher ability than a student who answered fewer items correctly.
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, etc.). 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 (see also 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.)

Table G–1 (continued). Glossary of Terms

Term	Common Definition
Core-Linking Item	Items that are utilized during the linking process (see also Linking). They are a subset of the PSSA 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 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 (e.g., 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. For the PSSA, three cut scores are used to place students into one of four performance levels (see also 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 who 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. In other words, students with the same ability level but different group memberships do not have the same probability of answering the item correctly (see also 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 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 also 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 (see also 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).

Table G–1 (continued). Glossary of Terms

Term	Common Definition
Field Test (FT) Items	The PSSA 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 (see also <i>P</i> -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 also Step Difficulty).
Key	The correct response option or answer to a test 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 (see also Equating).
Logit	In Rasch scaling, logits are units used to express both examinee ability and item difficulty. When expressing examinee ability, a student who answers more items correctly has a higher logit than a student who answers fewer items correctly. Logits are transformed into Scaled Scores through a linear transformation. When expressing item difficulty, logits are transformed <i>p</i> -value (see also <i>P</i> -value). The logit difficulty scale is inversely related to <i>p</i> -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, 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, 97} is 81. The value of a mean can be influenced by extreme values in a score distribution.

Table G–1 (continued). Glossary of Terms

Term	Common Definition
Measure	In Rasch scaling, measure generally refers to a specific estimate of an examinee’s ability (often expressed as logits) or an item’s difficulty (again, often expressed as logits). As an example for the PSSA, a student’s reading measure might be equal to 0.525 logits. Or, a PSSA Reading test item might have logit equal to -0.905.
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	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 (see also Open-Ended Item).
<i>N</i> -count	Sometimes designated as <i>N</i> or <i>n</i> , it is the number of 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 follow set {23, 32, 56, 65, 78, 87}, <i>n</i> = 6.
Open-Ended Item	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, etc.). 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 (MC) items 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 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.

Table G–1 (continued). Glossary of Terms

Term	Common Definition
Percent Correct	When referring to an individual item, the percent correct is the item’s <i>p</i> -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 possible points 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.
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, Percentiles 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 to Advanced, 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.

Table G–1 (continued). Glossary of Terms

Term	Common Definition
<i>P</i> -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. <i>P</i> -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. <i>P</i> -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 <i>p</i> -value-like scale can be estimated by dividing the item mean score by the maximum number of points possible for the item (see also Logit).
Raw Score	Sometimes abbreviated by 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, are 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 (see also 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.

Table G–1 (continued). Glossary of Terms

Term	Common Definition
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, and 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.
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)	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, who 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, are 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, implemented 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 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 provides information about 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 4–8 and 11. Item development for these new assessments took place during 2008.

Major assessment activities included the following:

- Spring standalone field test for mathematics at Grades 4–8 and 11

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

Subject	OP/FT	Grades	Assessment Schedule
Mathematics	FT (sa)	4–8, 11	Apr/May 2009

Note. FT (sa) refers to standalone field test.

ASSESSMENT ACTIVITIES OCCURRING IN THE 2009–10 SCHOOL YEAR

Table P–2 provides information about modified assessments during the 2009–10 school year. The mathematics modified assessments became operational for Grades 4–8 and 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. Consistent with the regular PSSA, a fall retest opportunity at Grade 12 was offered to students taking the mathematics modified assessment during the 2010 Fall Retest.

Standalone field tests in reading modified and science modified were conducted following the administration of the spring PSSA. Item development for these new assessments took place during 2009. Full implementation took place during the spring 2011 assessment.

Major assessment activities included the following:

- Spring operational assessment in mathematics for Grades 4–8 and 11 with embedded field testing
- Spring standalone field test for reading at Grades 4–8 and 11 and for science at Grades 8 and 11

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

Subject	OP/FT	Grades	Assessment Schedule
Mathematics	OP (eft)	4–8, 11	Apr/May 2010
Reading	FT (sa)	4–8, 11	May 2010
Science	FT (sa)	8, 11	May 2010

Note. OP (eft) refers to operational test with embedded field test.

FT (sa) refers to standalone field test.

ASSESSMENT ACTIVITIES PLANNED FOR THE 2010–11 SCHOOL YEAR

Table P–3 provides information about modified assessments during the 2010–11 school year. This was the second year for which the mathematics modified assessment was operational and the first year of implementation for the reading modified and science modified. Embedded field testing did not occur as part of the 2011 modified assessments.

A fall retest opportunity at Grade 12 was implemented for students taking the 2010 mathematics modified assessment. A retest opportunity will be available in the fall of 2011 for students failing to reach the Proficient level on the reading and/or science modified assessments.

Major assessment activities included the following:

- Spring operational assessment in mathematics and reading modified for Grades 4–8 and 11, and in science modified at Grades 8 and 11
- A retest opportunity for Grade 12 students who as 11th graders in the spring of 2010 failed to attain at least the Proficient level in mathematics modified

**Table P–3. Operational Assessment
During the 2010–11 School Year**

Subject	OP	Grades	Assessment Schedule
Mathematics	OP	4–8, 11	Mar 2011
Reading	OP	4–8, 11	Mar 2011
Science	OP	8, 11	May 2011
Retest for 2010 Mathematics	OP	12	Oct/Nov 2010

ASSESSMENT ACTIVITIES PLANNED FOR THE 2011–12 SCHOOL YEAR

Table P–4 shows the assessment plan for modified assessments during the 2011–12 school year. Assessment will begin in mid-March for mathematics and reading and late April for science. The make-up period for mathematics and reading will conclude in March; science will be complete in early May. This will be the third year of operational assessment for the mathematics modified and the second year of implementation for the reading and science modified. There is no embedded field testing as part of the operational modified assessments. A fall retest opportunity at Grade 12 will also be available in the fall of 2011.

Major planned assessment activities include the following:

- Spring operational assessment in mathematics and reading modified for Grades 4–8 and 11, and in science modified at Grades 8 and 11
- A retest opportunity will be offered to Grade 12 students who as 11th graders in the preceding spring assessment failed to attain at least the Proficient level in any of the subject areas

**Table P–4. Operational Assessment and Field Testing
During the 2011–12 School Year (Planned)**

Subject	OP	Grades	Assessment Schedule
Mathematics	OP	4–8, 11	Mar 2012
Reading	OP	4–8, 11	Mar 2012
Science	OP	8, 11	Apr/May 2012
Retest for 2011 Mathematics, Reading, Science	OP	12	Oct/Nov 2011

Chapter One: Background of the Modified Pennsylvania System of School Assessment

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 students with IEPs 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. Combined with the State Board adoption of *Science and Technology Standards* on July 12, 2001, and the *Environment and Ecology Standards* on January 5, 2002, the groundwork was laid 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 assess students at least once between Grades 10 and 12; students must be assessed in science at least once in each of the grade bands: Grades 3-5, Grades 6-9, and Grades 10-12.

PURPOSES OF THE PSSA

Chapter 4 regulations stipulated that the purposes of the PSSA include the following:

- 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 according to the 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 according to 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 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 determine more completely 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*. This report can be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” and then “Pennsylvania System of School Assessment (PSSA).” In the “Most Requested Content...” box, select “PSSA Technical Analysis.”

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, 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, science, and a school-administered alternate assessment for writing: 1) be enrolled in the assessed grade level for the subject area, 2) have a very severe cognitive disability, 3) require very intensive instruction, 4) require very extensive adaptation and support to perform or participate meaningfully, 5) require very substantial modification of the general education curriculum, and 6) participate in the general education curriculum differs very substantially in form and substance from that of other students. (See *The 2010–2011 PSSA Handbook for Assessment Coordinators (All Subjects)*, PDE, 2011, pp.10–11), which may be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Test Administration.”

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 students with IEPs 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.

Originally, PDE planned to develop modified assessments in reading for grades 3–8 and 11 and in science for grades 4, 8, and 11. However, the Pennsylvania PSSA-M Advisory Task Force met in January 2009 to discuss the criteria for the students for whom this test would be developed. The Task Force advised PDE to exclude third graders from the reading assessment and fourth graders from the science assessment, as the majority of these students could be properly assessed either with the general PSSA assessment at those grades or the PASA (Pennsylvania Alternate System of Assessment).

To address the unique needs of these students, and to 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 and 11. It was joined by operational modified assessments in reading at Grades 4–8 and 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 *2010–2011 PSSA Assessment Handbook* (see *Part Six: PSSA—Modified*). This handbook can be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “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, and 4) lacks academic progress. More detailed information on the PSSA-M eligibility criteria may be accessed at www.education.state.pa.us. On the left side, select “Programs,” “Programs S–Z,” and then “Special Education.” From the “Special Education” page select “Assessment” to access the relevant documents.

Chapter Two: Test Development Overview of the Modified PSSA

OVERVIEW OF THE DEVELOPMENT PROCESS

The Modified 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 were used to develop the general education assessment, Pennsylvania Student Assessment System (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 had 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 the items were, in fact, 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 acceptable for operational testing
- Reviewing the final operational forms prior to administering them 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, Reading, and Science

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 the 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 coding is read like an outline. The code includes the content, grade level, Reporting Category, Assessment Anchor, descriptor (Sub-Assessment Anchor), and Eligible Content. Thus, M4.A.1.1.1 would be: Mathematics, 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 while 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 better 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.edu. On the left, 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 4–8 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 usually generate fewer scoreable points in the same amount of testing time. They do, however, provide tasks that are more realistic and that 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 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, dropping 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 recall information from memory.

OPEN-ENDED TASKS FOR MATHEMATICS

For both the PSSA and the PSSA-M, open-ended tasks require students to read a problem description and to develop an appropriate solution. The PSSA-M open-ended items are designed to be scaffolded, which means that there are 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 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 at www.education.state.pa.us. On the left, select “Programs,” Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Resource Materials.”

The tables below 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 test content 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, has fewer items.

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
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%

Reading Assessment Measures

In keeping with the alignment of the PSSA, the PSSA-M reading assessment has two major reporting categories, Comprehension and Reading Skills and Interpretation and Analysis of Fictional and Nonfictional Text. These two reporting categories are derived from Reading Academic Standards 1.1, 1.2, and 1.3. As on the PSSA, Standards 1.6, 1.7, and 1.8 are not addressed on the PSSA-M because they are not specific to reading comprehension and can be more accurately evaluated at the school level. Standards 1.4 and 1.5 are addressed on the PSSA writing assessment. See Appendix A for more information about how the Academic Standards are linked to the Reporting Categories, Assessment Anchors, and Eligible Content.

The PSSA-M reading assessment, like the PSSA reading assessment, employs two types of test items: multiple-choice and open-ended. They are designed to measure students' comprehension of the information contained in the reading passages.

READING MULTIPLE-CHOICE ITEMS

Multiple-choice items measure such concepts as how well students comprehend the overall meaning of a passage or make basic inferences about it. At times, asking students to choose a preferred answer is the best way to determine whether they have gleaned certain important information from a story. Such information may include setting, central idea, or main events and their sequence.

Each reading multiple-choice item has four response options, only one of which is correct. The student is awarded one point for choosing the correct response. Incorrect response choices, or distractors, typically represent some kind of misinterpretation, predisposition, unsound reasoning, or casual reading. It is important to note that for the PSSA-M, dropping an answer option is not an allowable modification.

OPEN-ENDED TASKS FOR READING

Open-ended tasks are designed to address comprehension of text in ways that multiple-choice items cannot. A short written response, requiring about ten minutes per item, allows students to prepare an answer and summarize using supporting details or examples derived from the text.

The PSSA-M reading open-ended items, like the PSSA reading open-ended items, are scored on a 0–3 point scale with an item-specific scoring guideline. This scale is consistent with the scale used on the National Assessment of Educational Progress (NAEP). The change from the former 0–4 point scale improves the alignment with the types of tasks required. Each task is text-dependent and is carefully constructed with the scoring guide reflecting the task requirements. All item-specific scoring guidelines are based on the *General Scoring Guidelines for Open-Ended Reading Items*. The general guidelines describe a hierarchy of responses, which represent the four score levels. See Appendix B or the *Modified Reading Item and Scoring Samplers* available at www.education.state.pa.us. On the left, select “Programs,” Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Resource Materials.”

The following tables provide a high-level overview of the operational reading PSSA-M test plan as compared to the general education reading PSSA. In addition, a comparison of the reporting categories for the reading PSSA-M and the general education reading PSSA is also provided. The PSSA-M test content 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 2–3, has fewer items.

Table 2–3. Reading Operational Test Plan Summary: PSSA and PSSA-M

Program	Grades	Number of MC Items per PSSA	Number of 3-point OE Items per PSSA	Total Number of Points (MC + OE) per PSSA
PSSA	4, 5, 6, 7, 8, and 11	40	4	52
PSSA-M	4, 5, 6, 7, 8, and 11	30	2	36

Table 2–4. Reading Blueprint (percentage of total test points): PSSA and PSSA-M

Reporting Category	Program	Grade					
		4	5	6	7	8	11
Comprehension and Reading Skills	PSSA	60% - 80%	60% - 80%	50%-70%	50% -70%	40%-60%	40% - 60%
	PSSA-M	60% - 80%	60% - 80%	50%-70%	50% -70%	40%-60%	40% - 60%
Interpretation and Analysis of Fictional and Nonfictional Text	PSSA	20% - 40%	20% - 40%	30% - 50%	30% - 50%	40% - 60%	40% - 60%
	PSSA-M	20% - 40%	20% - 40%	30% - 50%	30% - 50%	40% - 60%	40% - 60%

Science Assessment Measures

The PSSA and the PSSA-M science assessments have four major reporting categories: The Nature of Science, Biological Science, Physical Science, and Earth and Space Sciences. These categories are similar to those used by the National Assessment of Educational Progress (NAEP) and The Third International Mathematics and Science Study (TIMSS). [However, the PSSA and the PSSA-M organize the categories differently.] The science assessment anchors cover seventeen major categories from two sets of standards: Science and Technology Standards (3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8) and Environment and Ecology Standards (4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, and 4.9). See Appendix A for more information about how the Academic Standards are linked to the Reporting Categories, Assessment Anchors, and Eligible Content.

The science assessment 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 science achievement. The design of the operational 2011 PSSA-M for science achieves a reasonable balance between the two item types.

SCIENCE MULTIPLE-CHOICE ITEMS

The majority of the science items included on the PSSA-M are multiple-choice items. This item type is especially efficient for measuring a broad range of content. In the PSSA-M science assessment, 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, or incorrect application of a scientific principle. It is important to note that for the PSSA-M, dropping 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 the application of science content. PSSA items involving application emphasize the requirement to utilize science content to find an answer, rather than simply recalling information from memory.

OPEN-ENDED ITEMS FOR SCIENCE

At all grades, standalone science open-ended items require students to read a description of a scientific problem and to develop an appropriate solution. Open-ended items require about five minutes per task.

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

The open-ended science items are scored on a 0–2-point scale with an item-specific scoring guideline, and each task is carefully constructed with a scoring guide reflecting the task requirements. The general guidelines describe a hierarchy of responses, which represent the three score levels. Each item-specific scoring guideline outlines the requirements at each score point, and each item-specific scoring guideline is based on the *Science Scoring Guidelines for Open-Ended Items*. See Appendix B or the *Modified Science Item and Scoring Samplers* available at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Resource Materials.”

The following tables provide a high-level overview of the operational science PSSA-M test plan as compared to the general education science PSSA. In addition, a comparison of the reporting categories for the science PSSA-M and the general education science PSSA is also provided. The PSSA-M test content 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 2–5, has fewer items.

Table 2–5. Science Operational Test Plan Summary: PSSA and PSSA-M

Program	Grades	Number of Standalone MC Items per PSSA	Number of Scenario-based MC Items per PSSA	Number of 2-point OE Items per PSSA	Number of Scenario-based 4-point OE Items per PSSA	Total Number of Points (MC + OE) per PSSA
PSSA	8	54	4	5	0	68
PSSA-M	8	30	0	2	0	34
PSSA	11	38	12	6	3	74
PSSA-M	11	30	0	2	0	34

Table 2–6. Science Blueprint (percentage of total test points): PSSA and PSSA-M

Reporting Category	Program	Grade	
		8	11
Nature of Science	PSSA	~50%	~50%
	PSSA-M	~50%	~50%
Biological Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%
Physical Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%
Earth and Space Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%

Chapter Three: Item Development Process

The core portion of the 2011 PSSA-M mathematics operational administration is primarily made up of items that were field tested in the 2010 PSSA-M embedded field test. The items used as core-to-core linking items in the 2011 PSSA-M mathematics operational administration were selected from items that were developed for the 2009 PSSA-M standalone field test. Therefore the activities that led to the 2011 PSSA-M operational mathematics administration began with the development of the draft test items that appeared in the 2009 PSSA-M standalone field test. The core portions of the 2011 PSSA-M reading and science operational administrations are made up of items that were field tested in the 2010 PSSA-M standalone field test. Therefore the activities that led to the 2011 PSSA-M operational reading and science administrations began with the development of draft test items that appeared in the 2010 PSSA-M standalone field test.

STEPS IN THE DEVELOPMENT PROCESS

A series of major activities took place in the development of the PSSA-M for mathematics. 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 test of items in spring 2009; item review with data; embedded field test of items in spring 2010; item review with data; and final selection of items to compose the 2011 PSSA-M mathematics assessment for grades 4–8 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	2009 FT for 2010 OP	Item modifications implemented in preparation for 2009 standalone field test
January 2009	2009 FT for 2010 OP	Item review and bias, fairness, and sensitivity review for candidate items for the 2009 standalone field test
February–March 2009	2009 FT for 2010 OP	Forms construction for the 2009 standalone field test
May 2009	Cognitive Interviews	Cognitive Interviews conducted in Pennsylvania schools
May 2009	2009 FT for 2010 OP	PSSA-M Mathematics Standalone Field Test
June–July 2009	2010 FT for 2011 OP	Item modifications (revisions and/or enhancements) implemented in preparation for 2010 embedded field test
July–August 2009	2010 FT for 2011 OP	Item review and bias, fairness, and sensitivity review for candidate items of the 2010 embedded field test
August 2009	2009 FT for 2010 OP	Statistical review of the 2009 field tested items

Table 3–1 (continued). PSSA-M Mathematics Development Timeline

Time Frame	Assessment	Activity
September 2009–January 2010	2010 OP & 2010 FT for 2011 OP	Forms construction for the 2010 operational assessment with embedded field test
April 2010	2010 OP & 2010 FT for 2011 OP	2010 operational assessment with embedded field test
August 2010	2010 FT for 2011 OP	Statistical review of the 2010 field tested items
August–September 2010	2010 OP & 2010 FT for 2011 OP	Forms construction for the 2011 operational assessment
March 2011	2011 OP	2011 operational assessment

Note. FT = Field Test
OP = Operational

A series of major activities took place in the development of the PSSA-M for reading. 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; pilot test in spring 2009; field test of items in spring 2010; item review with data; and final selection of items to compose the 2011 PSSA-M reading assessments. These activities are summarized in Table 3–2 below, and they are further described in the paragraphs that follow.

Table 3–2. PSSA-M Reading Development Timeline

Time Frame	Assessment	Activity
Sept 2008–Jan 2009	2009 Pilot Test and 2010 FT	Item modifications implemented in preparation for 2009 Reading Pilot Test and 2010 Reading Standalone Field Tests
Jan 2009	2009 Pilot Test and 2010 FT	Item review and bias, fairness, and sensitivity review for candidate items for the 2009 Reading Pilot Test and 2010 Reading Standalone Field Tests
Jan–March 2009	2009 Pilot Test	Forms construction for the 2009 Reading Pilot Test
May 2009	Cognitive Interviews	Cognitive Interviews conducted in Pennsylvania schools
May 2009	2009 Pilot Test	PSSA-M Reading Pilot Test

Table 3–2 (continued). PSSA-M Reading Development Timeline

Time Frame	Assessment	Activity
May–July 2009	2010 FT for 2011 OP	Item modifications (revisions and/or enhancements) implemented in preparation for 2010 Reading Standalone Field Tests
July–Aug 2009	2010 FT for 2011 OP	Item review and bias, fairness, and sensitivity review for candidate items of the 2010 Reading Standalone Field Tests
Aug–Oct 2009	2010 FT for 2011 OP	Forms construction for the 2010 Reading Standalone Field Tests
April–May 2010	2010 FT for 2011 OP	PSSA-M Reading Standalone Field Tests
Aug 2010	2010 FT for 2011 OP	Statistical review of the 2010 field tested items
Aug–Sept 2010	2010 OP for 2011 OP	Forms construction for the 2010 operational assessments for Reading
March–April 2011	2011 OP	2011 operational assessments for Reading

Note. FT = Field Test
OP = Operational

A series of major activities took place in the development of the PSSA-M for science. 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 test of items in spring 2010; item review with data; and final selection of items to compose the 2011 PSSA-M science assessments. These activities are summarized in Table 3–3 below, and they are further described in the paragraphs that follow.

Table 3–3. PSSA-M Science Development Timeline

Time Frame	Assessment	Activity
Sept 2008–Jan 2009	2010 FT	Item modifications implemented in preparation for 2010 Science Standalone Field Tests
Jan 2009	2010 FT	Item review and bias, fairness, and sensitivity review for candidate items for the 2010 Science Standalone Field Tests
May 2009	Cognitive Interviews	Cognitive Interviews conducted in Pennsylvania schools
May–July 2009	2010 FT for 2011 OP	Item modifications (revisions and/or enhancements) implemented in preparation for 2010 Science Standalone Field Tests
July–Aug 2009	2010 FT for 2011 OP	Item review and bias, fairness, and sensitivity review for candidate items of the 2010 Science Standalone Field Tests

Table 3–3 (continued). PSSA-M Science Development Timeline

Time Frame	Assessment	Activity
Aug– Oct 2009	2010 FT for 2011 OP	Forms construction for the 2010 Science Standalone Field Tests
April–May 2010	2010 FT for 2011 OP	PSSA-M Science Standalone Field Tests
Aug 2010	2010 FT for 2011 OP	Statistical review of the 2010 field tested items
Aug–Sept 2010	2010 OP for 2011 OP	Forms construction for the 2010 operational assessments for Science
March–April 2011	2011 OP	2011 operational assessments for Science

Note. FT = Field Test
OP = Operational

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 2011 PSSA-M operational administration, the initial planning meetings for the item modification process for the 2010 field test occurred throughout 2008. Item modification began in fall 2008, with the item review meetings occurring in 2009. See Tables 3–1, 3–2, and 3–3 for additional details.

Review of the Items

In September 2008, a pool of mathematics items from grades 4–8 and 11 was reviewed. In September 2008 and again in May 2009, pools of PSSA reading items from grades 4–8 and 11 and PSSA science items for grades 8 and 11 were reviewed. The review of the items focused on whether each item might lend itself well to revision and/or enhancement for possible field testing of PSSA-M items in spring 2010. The pool of candidate items was comprised of PSSA reading and science items that had been field tested in earlier administrations.

Training

To begin the process, WestEd and DRC selected and trained mathematics, reading, and science staff to review PSSA items for possible revision and/or enhancement. Qualified 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 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 content 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*. 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 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 of the Principles of Universal Design 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 3–4 shows the number of mathematics multiple-choice (MC) and open-ended (OE) items submitted to PDE for the item review meeting held in August of 2009.

Table 3–4. Mathematics Number of Items (MC and OE) Presented in August 2009 Item Review Meeting

Grade	August 2009 MC	August 2009 OE	Total Items
4	15	4	19
5	19	4	23
6	18	4	22
7	19	3	22
8	19	2	21
11	19	2	21
Total	109	19	128

Tables 3–5 and 3–6 show the number of reading multiple-choice (MC) items, open-ended (OE) items, and passages submitted to PDE for the item review meetings held in January and August of 2009.

Table 3–5. Reading Number of Items (MC and OE) Presented in January 2009 and August 2009 Item Review Meetings

Grade	January 2009 Modified MC	August 2009 Modified MC	August 2009 New MC	Total MC	January 2009 Modified OE	August 2009 Modified OE	August 2009 New OE	Total OE	Total Items
4	63	65	61	189	9	6	0	15	204
5	64	73	54	191	9	6	0	15	206
6	62	72	56	190	10	6	0	16	206
7	71	56	62	189	10	5	0	15	204
8	64	73	51	188	8	8	0	16	204
11	65	76	57	198	9	7	1	17	215
Total	389	415	341	1145	55	38	1	94	1239

Table 3–6. Reading Number of Passages Presented in January 2009 and August 2009 Item Review Meetings

Grade	January 2009 Passages	August 2009 Passages	Total Passages
4	9	6	15
5	9	6	15
6	9	6	15
7	10	5	15
8	8	7	15
11	8	7	15
Total	53	37	90

Table 3–7 shows the number of science multiple-choice (MC) and open-ended (OE) items submitted to PDE for the item review meetings held in January and August of 2009.

Table 3–7. Science Number of Items (MC and OE) Presented in January 2009 and August 2009 Item Review Meetings

Grade	January 2009 Modified MC	August 2009 Modified MC	Total MC	January 2009 Modified OE	August 2009 Modified OE	Total OE	Total Items
8	67	71	138	11	9	20	158
11	69	64	133	11	6	17	150
Total	136	135	271	22	15	37	308

SUMMARY OF 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 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 or spring 2010 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 when appropriate: When passages are segmented, items follow an order that parallels how information generally appears in the passage and/or prompt. (For example, for the reading PSSA-M, when appropriate, students will be provided the same passage/prompt 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 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 multistep mathematics item, as appropriate, subquestions or steps to break up or help students think through multistep problems/item are provided

- Adding additional directions to explain a process or activity
- Adding prereading information to clarify the purpose of a passage or prompt
- 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 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 in 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 (cognitive complexity), estimated difficulty, answer key (MC items), and calculator use (mathematics).

All items undergoing field testing in 2010 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). Sample IDEAS Item Cards are 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 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, the items were reviewed by two separate committees at different stages. The first committee to meet was the Bias, Fairness, and Sensitivity committee, and the second was the Item Content committee. The first Bias, Fairness, and Sensitivity meeting was held in Harrisburg, PA, on January 12 and 13 of 2009, and the first Item Content meeting, also held in Harrisburg, PA, took place January 14 through 16 of 2009. A second set of meetings was held to review additional modified items. The second Bias, Fairness, and Sensitivity meeting was held in Harrisburg, PA, on July 28 and 29 of 2009, and the second Item Content meeting, also held in Harrisburg, PA, took place on August 5 and 6 of 2009. Summaries, guidelines, and procedures for each meeting are presented below.

BIAS, FAIRNESS, AND SENSITIVITY REVIEW

Prior to 2010 field testing, all revised and/or enhanced PSSA-M items were submitted to a Bias, Fairness, and Sensitivity committee for review. The committee members consisted of a cross-representation of ethnic groups across the six members. There was one Latin American, two African Americans, one Native American, and two Caucasians represented in the committee. 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 the acceptability of items with regard to bias, fairness, and sensitivity issues. They also made recommendations for changes to or deletion of items in order to remove the potential for issues of bias, fairness, and/or sensitivity.

The expert, multiethnic 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, 2003–2010). 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 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, region, religion, socioeconomic status, 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 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.

ITEM CONTENT REVIEW

Prior to the 2009 field testing, all revised and/or enhanced items were also submitted to content committees for review. 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 students who may be administered the PSSA-M assessment. The primary responsibility of the content committees was to evaluate the revised and/or enhanced items with regard to the quality of the revision and/or enhancement, the content classification and 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; 2007), items were identified in which the cognitive demand 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, or a flawed graphic in an item revision and/or enhancement, or if the item 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 and 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 an opportunity to ask questions were also included in the PowerPoint training. In addition, the training 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 the 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 8 and 12 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.

READING PILOT

As a result of feedback from both the Technical Advisory Committee (TAC) and the PSSA-M Advisory Task Force, PDE and DRC proposed a small pilot test in reading, Grades 5 and 8 only. The chart that follows shows the major deliverables and deadlines for the 2009 reading pilot test pull lists.

Table 3–8. 2009 PSSA-M Reading Pilot Test Schedule

Date/Timeframe	Task
February 17, 2009	Reading pilot test pull list delivered to DRC from WestEd.
February 27	Pilot DFA template provided to WestEd.
March 5	DRC provided initial draft of reading pilots to WestEd.
March 10	WestEd provided initial reading pilot feedback to DRC.
March 13	DRC received pilot DFA from WestEd.
March 13	Reading pilot test forms provided to PDE/WestEd for review.
March 18	PDE and WestEd provided final approval to print PSSA-M pilot test.
March 23	PDE received pilot DFA for review.
March 23	Reading Pilot test approved for printer's proof production.
March 27	Reading Pilot test DFA approved by PDE.
April 8	Reading Pilot test admin materials approved for printer's proof production.

The specifications required three pilot test forms per grade. There were two passages, ten multiple-choice items, and one open-ended item per form. The following forms layout was used:

Table 3–9. 2009 PSSA-M Reading Pilot Forms Layout

Enhancement	Form 1	Form 2	Form 3
Passage Segmentation	Passage A: No segmentation Passage B: Segmentation (version 1)	Passage A: Segmentation (version 1) Passage B: No segmentation	Passage A: Segmentation (version 2) Passage B: Segmentation (version 2)
Passage Columns	Both passages: One column (usual margins)	Both passages: Two columns	Passage A: One column (wider margins) Passage B: Two columns
OE Items	Three points, scaffolded into parts a, b, c with single answer space	Three points, scaffolded with answer space following each part	Two points; scaffolded with answer space following each part
Glossing	At least one passage: Footnote style glossing	At least one passage: Bolded-word, glossing at bottom of page	At least one passage: Bolded-word, glossing in margin

Table 3–9 (continued). 2009 PSSA-M Reading Pilot Forms Layout

Enhancement	Form 1	Form 2	Form 3
Helpful Hints In Items	Both passages: Helpful hint above the item	Both passages: Helpful hint below the item	Passage A: Helpful hint as a first statement in the stem Passage B: Helpful hint in parentheses in the stem
Item Placement	Both passages: Two items per page	Passage A: All items on facing page with each passage segment; the “whole passage” (with explanation that it is the parts put together) placed before questions that cut across segments Passage B: Two items per page	Both passages: All items on facing page with each passage segment; direction telling students to use “all segments” to answer the questions placed before questions that cut across segments

The reading pilot tests were administered from May 18, 2009, through May 22, 2009. The pilot tests were scored and the results were shared by DRC with PDE in October 2009.

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 assessments, the interviews were conducted prior to the final development and 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 report titled *Cognitive Interviews in Pennsylvania: Report on Data Collection for the Pennsylvania System of School Assessment Alternate Assessment with Modified Achievement Standards (PSSA-M) Study*.¹ This report is available upon request from PDE. 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, and typographic feature enhancements; challenges with terminology and vocabulary; and findings by item enhancement type.

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

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 reading and science field test items. Mathematics 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 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' understandings 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 more deeply into the student's thinking processes about that item.

This multistep 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 2011 PSSA-M ASSESSMENTS

The PSSA-M, like the PSSA, is based on the Pennsylvania Academic Standards. The 2011 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 3–10. 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%

Table 3–11. Reading Blueprint (Percentage of Total Test Points): PSSA and PSSA-M

Reporting Category	Program	Grade					
		4	5	6	7	8	11
Comprehension and Reading Skills	PSSA	60% - 80%	60% - 80%	50%-70%	50% -70%	40%-60%	40% - 60%
	PSSA-M	60% - 80%	60% - 80%	50%-70%	50% -70%	40%-60%	40% - 60%
Interpretation and Analysis of Fictional and Nonfictional Text	PSSA	20% - 40%	20% - 40%	30% - 50%	30% - 50%	40% - 60%	40% - 60%
	PSSA-M	20% - 40%	20% - 40%	30% - 50%	30% - 50%	40% - 60%	40% - 60%

Table 3–12. Science Blueprint (Percentage of Total Test Points): PSSA and PSSA-M

Reporting Category	Program	Grade	
		8	11
Nature of Science	PSSA	~50%	~50%
	PSSA-M	~50%	~50%
Biological Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%
Physical Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%
Earth and Space Science	PSSA	~17%	~17%
	PSSA-M	~17%	~17%

Operational Layout for 2011 PSSA-M

The PSSA-M mathematics assessments for Grades 4–8 and 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).

The PSSA-M reading assessments for Grades 4–8 and 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 3 points (scale of 0–3).

The PSSA-M science assessments for Grades 8 and 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 2 points (scale of 0–2).

For 2011, each test form contained common items taken by all students. The 2011 PSSA-M was comprised of one form per grade per content area. Tables 3–13 and 3–14 display information about the test form layout.

Table 3–13. 2011 PSSA-M 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 Forms per Grade
Mathematics	2011	30	2	1
Reading	2011	30	2	1
Science	2011	30	2	1

*MC = Multiple-Choice

**OE = Open-Ended

Table 3–14. 2011 PSSA-M Operational Test Layout

Content Area	Grades	Item Stage	Section 1	Section 2
Mathematics	4–8, 11	Core	15 MC	15 MC
Mathematics	4–8, 11	Core	1 OE	1 OE
Reading	4, 5, 7, 8, 11	Core	18 MC	12 MC
Reading	6	Core	19 MC	11 MC
Reading	4–8, 11	Core	1 OE	1 OE
Science	8, 11	Core	15 MC	15 MC
Science	8, 11	Core	1 OE	1 OE

An individual student’s score is obtained by combining the points from the core MC and OE portions of the test as follows:

Table 3–15. 2011 PSSA-M Core Points

Student’s Score	Grades	MC Items	OE Items	Total Score
Mathematics	4–8, 11	30	2 items X 4 points=8 points	38
Reading	4–8, 11	30	2 items X 3 points=6 points	36
Science	8, 11	30	2 items X 2 points=4 points	34

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 was 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 were used for linking purposes. Open-ended items were not repeated as linking items across cores. Approximately 20–50% of the multiple-choice items for each grade were repeated as linking items.

Linking for 2011 and 2012 PSSA-M Assessments

Linking provides a statistical bridge between assessment administrations. The 2012 administration will be linked back to the 2011 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. Approximately 20–50% of the multiple-choice items for each grade will be repeated as linking items.

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

Test Sessions and Timing for 2011 PSSA-M Mathematics Assessments

The test window for the 2011 operational assessment, including make-ups, extended from March 14 through April 15, 2011. 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 table outlines 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 3–16. PSSA-M Mathematics—2011 Administration and Testing Times

Test Section	Suggested Times (In Minutes)			Grade Level Number of Items and Item Type					
	Administration (Total)	Administrative (Pre & Post)	Student Testing	4	5	6	7	8	11
1	65 to 80	15 to 20	50 to 60	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE
2	65 to 80	15 to 20	50 to 60	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE	15 MC 1 OE

Note. MC refers to multiple-choice and OE refers to open-ended items.

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.

Test Sessions and Timing for 2011 PSSA-M Reading Assessments

The test window for the 2011 operational assessment, including make-ups, extended from March 14 through April 15, 2011. The reading 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 table outlines 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 3–17. PSSA-M Reading—2011 Administration and Testing Times

Test Section	Suggested Times (In Minutes)			Grade Level Number of Items and Item Type					
	Administration (Total)	Administrative (Pre & Post)	Student Testing	4	5	6	7	8	11
1	75 to 90	15 to 20	60 to 70	18 MC 1 OE	18 MC 1 OE	19 MC 1 OE	18 MC 1 OE	18 MC 1 OE	18 MC 1 OE
2	55 to 70	15 to 20	40 to 50	12 MC 1 OE	12 MC 1 OE	11 MC 1 OE	12 MC 1 OE	12 MC 1 OE	12 MC 1 OE

Note. MC refers to multiple-choice and OE refers to open-ended items.

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.

Test Sessions and Timing for 2011 PSSA-M Science Assessments

The test window for the 2011 operational assessment, including make-ups, extended from April 4 through April 15, 2011. The science 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 table outlines 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 3–18. PSSA-M Science—2011 Administration and Testing Times

Test Section	Suggested Times (In Minutes)			Grade Level Number of Items and Item Type	
	Administration (Total)	Administrative (Pre & Post)	Student Testing	8	11
1	45 to 60	15 to 20	30 to 40	15 MC 1 OE	15 MC 1 OE
2	45 to 60	15 to 20	30 to 40	15 MC 1 OE	15 MC 1 OE

Note. MC refers to multiple-choice and OE refers to open-ended items.

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 2011 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 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 in these five categories and their percentages of the total number of test points are shown in the following table.

Table 3–19. 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.”

Reporting Categories and Points Distributions for 2011 PSSA and PSSA-M Reading Assessments

The Reading assessment results will be reported in two broad categories:

- A. Comprehension and Reading Skills
- B. Interpretation and Analysis of Fictional and Nonfictional Text

Assessment Anchors associated with Comprehension and Reading Skills are coded with an initial letter “A,” and those related to Interpretation and Analysis of Fictional and Nonfictional Text are coded with an initial letter “B.” The distribution of items in these two categories across genres and their percentages of the total number of test points are shown in the following table.

Table 3–20. Reading Reporting Categories and Genre and Point Distributions

Grade	Reporting Categories				
	A: Comprehension and Reading Skills % Range	B: Interpretation and Analysis of Fictional and Nonfictional Text % Range	Total Points	% of Passages (Genre) Fiction	% Passages (Genre) Nonfiction
Grade 4	60%–80% 22–29 points	20%–40% 7–14 points	36	50%–70%	30%–50%
Grade 5	60%–80% 22–29 points	20%–40% 7–14 points	36	50%–70%	30%–50%
Grade 6	50%–70% 18–25 points	30%–50% 11–18 points	36	40%–60%	40%–60%
Grade 7	50%–70% 18–25 points	30%–50% 11–18 points	36	40%–60%	40%–60%
Grade 8	40%–60% 14–22 points	40%–60% 14–22 points	36	40%–60%	40%–60%
Grade 11	40%–60% 14–22 points	40%–60% 14–22 points	36	30%–50%	50%–70%

Like the mathematics reporting categories, reading 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.”

Reporting Categories and Points Distributions for 2011 PSSA and PSSA-M Science Assessments

The science assessment results will be reported in four categories, coded as A through D:

- A. The Nature of Science
- B. Biological Science
- C. Physical Science
- D. Earth and Space Science

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

Table 3–21. Science Reporting Categories and Point Distributions

Grade	Reporting Categories				Total Points
	A: Nature of Science	B: Biological Sciences	C: Physical Sciences	D: Earth & Space Sciences	
Grade 8	~50% ~17 points	~17% ~5–6 points	~17% ~5–6 points	~17% ~5–6 points	34
Grade 11	~50% ~17 points	~17% ~5–6 points	~17% ~5–6 points	~17% ~5–6 points	34

The science 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,” “Descriptors (Sub-Assessment Anchors),” and “Eligible Content.”

Assessment Anchor Content Standards Subsumed within Reporting Categories for 2011 Modified Assessments

For mathematics, there are 16 Assessment Anchor Content Standards (Assessment Anchors) that occur at all grade levels (Grades 4–8 and 11), although they are not all assessed at each grade level. More specifically, the number targeted for assessment by grade level is 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.

For reading, there are five Assessment Anchors that vary to reflect grade-level appropriateness. Within the Comprehension and Reading Skills Reporting Category, two Assessment Anchors pertain to understanding fiction text and understanding nonfiction text. Within the Interpretation and Analysis of Fiction and Nonfiction Text Reporting Category, three Assessment Anchors pertain to Components of Text, Literary Devices and Concepts, and Organization of Nonfiction Text.

For science, there are 12 Assessment Anchors that exist at each grade. Within The Nature of Science Reporting Category, three Assessment Anchors pertain to Reasoning and Analysis; Processes, Procedures, and Tools of Scientific Investigations; and Systems, Models, and Patterns. Within the Biological Sciences Reporting Category, three Assessment Anchors pertain to Structure and Function of Organisms, Continuity of Life, and Ecological Behavior and Systems. Within the Physical Sciences Reporting Category, three Assessment Anchors pertain to Structure, Properties, and Interaction of Matter and Energy; Forms, Sources, Conversion, and Transfer of Energy; and Principles of Motion and Force. Within the Earth and Space Sciences Reporting Category, three Assessment Anchors pertain to Earth Features and Processes that Change Earth and Its Resources; Weather, Climate, and Atmospheric Processes; and Composition and Structure of the Universe.

Mathematics, reading, and science scores are based on the core (common) sections. Also reported are the student’s mathematics and reading 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, 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* was used for developing items. All items were reviewed for fairness by bias, fairness, 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 stereotype, gender, region/geography, ethnic group/culture, socioeconomic status/class, religion, experiential, and 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-M. 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 is clean and well organized.

A more extensive description of the application of Principles of Universal Design is provided 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 raise issues of fairness or sensitivity.

TEST DEVELOPMENT PROCESS

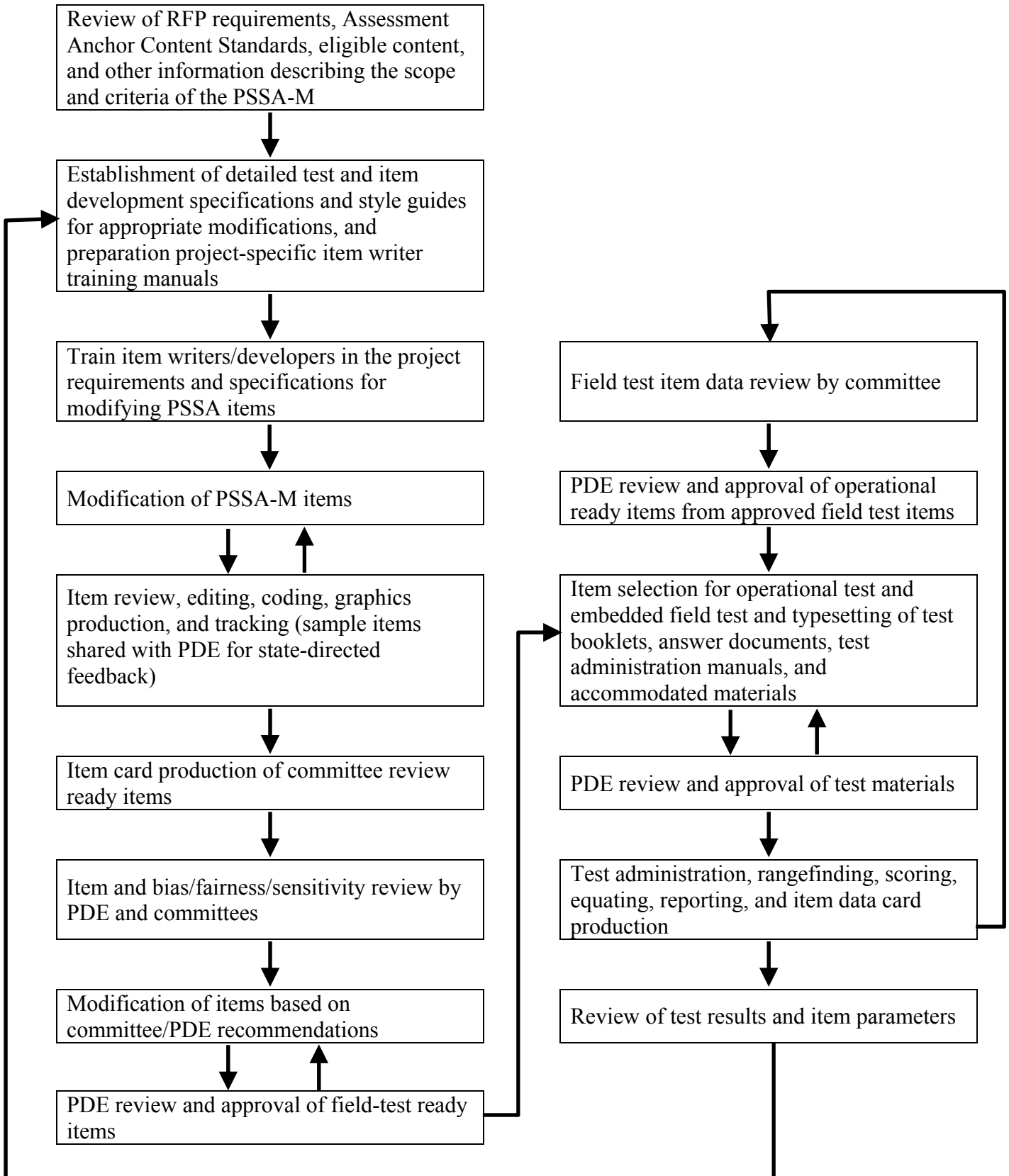
The item development process follows a logical timeline, which is outlined below in Figure 3–1. On the front-end of the schedule, tasks are generally completed with the goal of presenting field test candidate items to committees of Pennsylvania educators. On the back-end of the schedule, all tasks led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

Figure 3–1. Item and Test Development Cycle and Timeline (2010–2011 only)

Steps in Development Cycle	Timeline to/from New Item Review	
Development planning	Fall	↓ -12 to -4 months
Initial item modification	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
Test forms building	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
Operational item selection	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 2010 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)(I)]. 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 apply not only to these groups of students but also 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 was 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 “Assessment Accommodations” later 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 with 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 developed and updated annually (DRC, 2004–2010) was utilized, with PDE approval, which included dimensions of style consistent with universal design.

GUIDELINES FOR UNIVERSALLY DESIGNED ITEMS

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

- 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. Spanish versions for the PSSA-M mathematics and the PSSA-M science were 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 were avoided that may not be necessary (e.g., sidebars, overlays, callout boxes, visual crowding, shading) and that could be potentially distracting to students. 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 work closely with the Pennsylvania Department of Education to help ensure that PSSA-M tests comply with nationally recognized Principles of Universal Design. We support 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 apply to each content area assessment the standards for test accessibility as described in *Test Access: Making Tests Accessible for Students with Visual Impairments: A Guide for Test Publishers, and State Assessment Personnel* (Allman, 2004). To this end, we embrace the following precepts:

- Test directions are carefully worded to allow for alternate responses to open-ended questions.
- During item and bias reviews, test committee members are 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 instructs 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 are 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 formats PSSA-M tests to maximize accessibility for all students by using text that is in a 13-point size and a font style that is easily readable. DRC limits shading, spacing, graphics, charts, and the number of items per page so that there is sufficient white space on each page. Whenever possible, we ensure that graphics, pictures, diagrams, charts, and tables are positioned on the page with the associated test items. We use high contrast for text and background where possible to convey pertinent information. Tests are published on dull-finish paper to avoid the glare encountered on glossy paper. DRC pays 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 ensures consistency across PSSA and PSSA-M assessments by following these Principles of Universal Design.

- High contrast and clarity are 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, Keystone (paper/pencil) Accommodations Guidelines for Students with IEPs and Students with 504 Plans* was developed for use with the 2011 PSSA and PSSA-M. The manual can be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “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 PSSA-M mathematics and PSSA-M science. The Spanish-translation versions are discussed in Chapter Six.

Chapter Five: Field Test Leading to the 2010 Core

STANDALONE FIELD TEST ITEMS

All core items appearing on the 2011 reading and science assessments came from the Spring 2010 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 2010 PSSA-M standalone field test were selected for use as common core items in the 2011 PSSA-M.

Table 5–1. 2010 Spring PSSA-M Reading Field Test

Grade	No. of Passages per FT Form	No. of FT MC per FT Form	No. of FT OE per FT Form	Total No. of Forms	Total No. of Passages	Total No. of FT MC Items per Passage	Total No. of FT OE Items per Passage	Total No. of FT MC Items	Total No. of FT OE Items
4	5	30	3	6	15	12	1	180	15
5	5	30	3	6	15	12	1	180	15
6	5	30	3	6	15	12	1	180	15
7	5	30	3	6	15	12	1	180	15
8	5	30	3	6	15	12	1	180	15
11	5	30	3	6	15	12	1	180	15

Table 5–2. 2010 Spring PSSA-M Science 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
8	24	3	5	120	15	135
11	24	3	5	120	15	135

Core items appearing on the 2011 mathematics assessment came from the spring 2010 mathematics embedded field test and also from the 2009 standalone field test. More information on the 2010 field test designs can be found in specific portions of Chapter Three. Additional information about the 2009 standalone mathematics field test can be found in the *Technical Report for the 2010 Modified Pennsylvania System of School Assessment*. The report can be accessed at www.education.state.pa.us. Type “2010 PSSA Technical Report” in the search box.

STATISTICAL ANALYSIS OF ITEM DATA

All field-tested items were analyzed statistically following conventional item analysis methods. For multiple-choice (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 open-ended (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 in question 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 of items. 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

As stated in the preceding section, “Statistical Analysis of Item Data,” test development content-area specialists used statistics from item and DIF analyses of the 2010 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 for reading and science being in its initial stages, however, it was determined that all PSSA-M reading and science items necessary for the building of two equivalent core forms, both multiple-choice and open-ended, be brought to the data review for approval. Additionally, all mathematics items, both multiple-choice and open-ended, used in the 2010 embedded mathematics field test were brought to the data review, regardless of statistical performance.

The review of the items with data was conducted by over 50 Pennsylvania educators (teachers and PDE staff) broken out into content and grade-level committees. The review took place on August 9–10, 2010. In this review, 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 5–3. 2010 Mathematics Data Review Committee Results

Mathematics	Grade	No. of Items in 2010 Field Test	Field Test Items Examined at 2010 Data Review Committee			Field Test Items Rejected by 2010 Data Review Committee*			Items Classified as “Rejected” from 2010 Field Test (all sources)**		
			No. of		% of FT	No. of		% of FT	No. of		% of FT
			MC	OE		MC	OE		MC	OE	
			MC	OE	FT	MC	OE	FT	MC	OE	FT
	4	27	24	3	100%	0	1	3.7%	0	1	3.7%
	5	27	24	3	100%	1	1	7.4%	1	1	7.4%
	6	27	24	3	100%	2	1	11.1%	2	1	11.1%
	7	27	24	3	100%	3	0	11.1%	3	0	11.1%
	8	27	24	3	100%	1	1	7.4%	1	1	7.4%
	11	27	24	3	100%	2	2	14.8%	2	2	14.8%
	Total	162	144	18	100%	9	6	9.3%	9	6	9.3%

*Rejected as a result of statistics

**Data Review Committee, PDE, and DRC

Table 5–4. 2010 Reading Data Review Committee Results

Reading	Grade	No. of Items in 2010 Field Test	Field Test Items Examined at 2010 Data Review Committee			Field Test Items Rejected by 2010 Data Review Committee*			Items Classified as “Rejected” from 2010 Field Test (all sources)**		
			No. of		% of	No. of		% of	No. of		% of
			MC	OE	FT	MC	OE	FT	MC	OE	FT
	4	195	47	4	26.2%	0	0	0.0%	0	0	0.0%
	5	195	51	4	28.2%	0	0	0.0%	0	0	0.0%
	6	195	50	4	27.7%	1	0	0.5%	1	0	0.5%
	7	195	48	4	26.7%	0	0	0.0%	0	0	0.0%
	8	195	55	4	30.3%	0	0	0.0%	0	0	0.0%
	11	195	48	4	26.7%	0	0	0.0%	0	0	0.0%
Total		1,170	299	24	27.6%	1	0	0.001%	1	0	0.001%

*Rejected as a result of statistics

**Data Review Committee, PDE, and DRC

Table 5–5. 2010 Science Data Review Committee Results

Science	Grade	No. of Items in 2010 Field Test	Field Test Items Examined at 2010 Data Review Committee			Field Test Items Rejected by 2010 Data Review Committee*			Items Classified as “Rejected” from 2010 Field Test (all sources)**		
			No. of		% of	No. of		% of	No. of		% of
			MC	OE	FT	MC	OE	FT	MC	OE	FT
	8	135	47	4	37.8%	0	0	0.0%	0	0	0.0%
	11	135	49	4	39.3%	0	0	0.0%	0	0	0.0%
Total		270	96	8	38.5%	0	0	0.0%	0	0	0.0%

*Rejected as a result of statistics

**Data Review Committee, PDE, and DRC

Chapter Six: Operational Forms Construction for 2010

FINAL SELECTION OF ITEMS AND 2011 PSSA-M FORMS CONSTRUCTION

By the time the final selection of items for the operational 2011 test had begun, the candidate items that emerged from the spring 2010 field test had undergone multiple reviews, including the following:

- 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, multiethnic 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 test item. All items with an item status code of “Acceptable/Active” were candidates to be selected for the 2011 PSSA-M. To have an item status code of “Acceptable/Active” meant that an 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 level. 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 tests.

For the common items, this meant that the combination of multiple-choice (MC) and open-ended (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 functioned 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” (distracters 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 was either in the form of approval or recommendation to replace 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 TEST TO THE 2011 OPERATIONAL TEST

The 2010 operational PSSA-M mathematics test was linked with the 2011 operational PSSA-M mathematics test using core-to-core linking items (items that are repeated from operational form to the next).

In the selection of the core-to-core linking items (part of the overall core pull), content considerations remained 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.

LINKING THE 2011 OPERATIONAL TEST TO THE 2012 OPERATIONAL TEST

The 2011 operational PSSA-M mathematics, reading, and science tests will be linked with the 2012 operational PSSA-M mathematics, reading, and science tests using core-to-core linking items.

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, have 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 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 (MC and OE) 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 form designation.

Spanish Translation of the Mathematics and Science 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-M for mathematics and/or science only. The original translation of the items and the *Directions for Administration Manuals* 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 4–8 and 11 and the science booklets for Grades 8 and 11 were designed with a modified “over/under” format, with the Spanish presented directly above or to the left of the English. The English text was presented in italics and in a smaller font than the Spanish text. Those students using this accommodated version of the mathematics and/or science assessments could write their answers in English, Spanish, or a combination of both English and Spanish, 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 18 students at Grades 4, 6, 7, 8, and 11 in 2011. Spanish translated versions of the 2011 PSSA-M science assessment were used by a total of 12 students at Grades 8 and 11.

Instructions for the appropriate use of these special forms are detailed in the *PSSA, PSSA-M, Keystone (paper/pencil) Accommodations Guidelines for Students with IEPs and Students with 504 Plans*. The manual can be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Testing Accommodations & Security.”

Chapter Seven: Test Administration Procedures

TEST SESSIONS, TEST SECTIONS, AND TEST TIMING

The PSSA-M mathematics, reading, and science tests utilize 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 tests without having to manage two separate booklets.

The PSSA-M tests consist of two untimed sections. Testing-time recommendations are given, but the estimated times are 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	1	15	1	March 14–25	March 28–April 15
5	2	15	1	15	1	March 14–25	March 28–April 15
6	2	15	1	15	1	March 14–25	March 28–April 15
7	2	15	1	15	1	March 14–25	March 28–April 15
8	2	15	1	15	1	March 14–25	March 28–April 15
11	2	15	1	15	1	March 14–25	March 28–April 15

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

Grade	Total No. of MC Items per Form per Administration	Total No. of OE Items per Form per Administration	Total Estimated Administration Time per Form (in Minutes)
4	30	2	130 to 160
5	30	2	130 to 160
6	30	2	130 to 160
7	30	2	130 to 160
8	30	2	130 to 160
11	30	2	130 to 160

Table 7–3. PSSA-M Reading 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	18	1	12	1	March 14–25	March 28–April 15
5	2	18	1	12	1	March 14–25	March 28–April 15
6	2	19	1	11	1	March 14–25	March 28–April 15
7	2	18	1	12	1	March 14–25	March 28–April 15
8	2	18	1	12	1	March 14–25	March 28–April 15
11	2	18	1	12	1	March 14–25	March 28–April 15

Table 7–4. PSSA-M Reading Duration and Testing Load by Grade

Grade	Total No. of MC Items per Form per Administration	Total No. of OE Items per Form per Administration	Total Estimated Administration Time per Form (in Minutes)
4	30	2	130 to 160
5	30	2	130 to 160
6	30	2	130 to 160
7	30	2	130 to 160
8	30	2	130 to 160
11	30	2	130 to 160

Table 7–5. PSSA-M Science 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
8	2	15	1	15	1	April 4–8	April 11–15
11	2	15	1	15	1	April 4–8	April 11–15

Table 7–6. PSSA-M Science Duration and Testing Load by Grade

Grade	Total No. of MC Items per Form per Administration	Total No. of OE Items per Form per Administration	Total Estimated Administration Time per Form (in Minutes)
8	30	2	90 to 120
11	30	2	90 to 120

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

Test administrators are advised to use a testing location that is separate from the administration of the general PSSA assessment. For 2011, students who participated in the PSSA-M mathematics assessment may have participated in the general PSSA reading assessment. These students were allowed to complete both sections of the PSSA-M mathematics assessment before completing the three general PSSA reading sections. Alternating the PSSA-M mathematics 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 in which they were printed in the booklets. Likewise, students who participated in the PSSA-M reading assessment may have participated in the general PSSA mathematics assessment. Test administrators had the option of administering both sections of the PSSA-M reading assessment before administering the three general PSSA mathematics sections or alternating the PSSA-M reading sections with the PSSA mathematics sections, as long as the subject sections were administered in sequence in which they were printed in the booklets.

Since not all students will finish the assessment sections at the same time, test administrators are advised to use the flexibility of the time limits to the students' advantage. For example, test administrators manage the testing time so that students do not feel rushed while they are taking any assessment section, and no student is penalized because he or she works slowly. It is equally stressed to test administrators that a student should not be given an opportunity to waste time. Students are told to close their booklets when they have finished the section of the assessment in which they had been working. Students who finish early are allowed to sit quietly or read for pleasure until all students have 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 work slowly may require extended time. Special assessment situations are arranged for these students. When all students in a testing session have indicated that they have finished an assessment section, test administrators end the section and begin the next section or allow the students to return to regular activities.

Scheduled extended time can be provided by a test administrator, and students may request extended time if they indicate that they have not completed a task. Such requests are granted if the test administrator finds the request to be educationally valid. Test administrators are advised that not permitting ample time for students to complete the assessments may impact the students' and schools' performances.

As a general guideline, however, when all students indicate that they have finished a section, that section is closed. Students requiring time beyond the majority of the student population are allowed to continue immediately following the regularly scheduled session in another setting. When such accommodations are made, school personnel ensure that students are monitored at all times to prevent sharing of information. Students are 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 are asked to print out and distribute a copy of each grade's formula sheets. The formula sheets are posted at www.education.state.pa.us. On the left, select "Programs," "Programs O-R," "Pennsylvania System of School Assessment (PSSA)," and then select "Resource Materials." The formula sheets are listed under "Mathematics Resources."

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

TESTING WINDOW

The testing windows for the 2011 PSSA-M operational assessments were as follows:

Mathematics and Reading:

- Primary testing window – March 14 through 25, 2011
- Make-up testing window – March 28 through April 15, 2011

Science:

- Primary testing window – April 4 through 8, 2011
- Make-up testing window – April 11 through 15, 2011

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

SHIPPING, PACKAGING, AND DELIVERY OF MATERIALS

DRC sent two shipments for the 2011 PSSA-M operational assessment:

- Shipment one contained the *Handbook for Assessment Coordinators* and the *Directions for Administration Manuals* for each grade tested at a school participating in the mathematics, reading, and science assessments. Shipment one was delivered by February 14, 2011.
- 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, reading, and science assessments. Shipment two was delivered by February 28, 2011.

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 103,422 modified assessment booklets and 35,073 modified *Directions for Administration Manuals* for 3,994 testing sites. DRC used United Parcel Service (UPS) and Advanced Shipping Technologies (AST) 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 – March 14 through April 8, 2011
- Make-up return window – April 11 through April 15, 2011

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 materials. Every principal or director was to sign and return the test security affidavit with the return of the testing materials. 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 Manuals* can be found on the PDE website at www.education.state.pa.us. On the left, select "Programs," "Programs O–R," "Pennsylvania System of School Assessment (PSSA)," and the "Test Administration."

TESTING WINDOW ASSESSMENT ACCOMMODATIONS

Three accommodations manuals, *PSSA*, *PSSA-M*, and *Keystone (paper pencil) 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 2011 PSSA-M. Additional information regarding assessment accommodations can be found in Chapter Four of this report. These manuals can be found at www.education.state.pa.us. On the left, select "Programs," "Programs O–R," "Pennsylvania System of School Assessment (PSSA)," and then "Testing Accommodations & Security."

Chapter Eight: Processing and Scoring

RECEIPT OF MATERIALS

Receipt of PSSA-M test materials began on March 21, 2011, and concluded with all make-up tests on April 15, 2011. 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 sorted materials 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 formats. 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 receipt versus expected receipt. 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 returned 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 booklets.

The scannable documents were automatically fed into the image scanners where predefined processing criteria determined which fields were to be captured electronically. Open-ended item 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 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 predefined 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 omissions in predetermined fields. Marks that did not meet the predefined 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 files. 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.

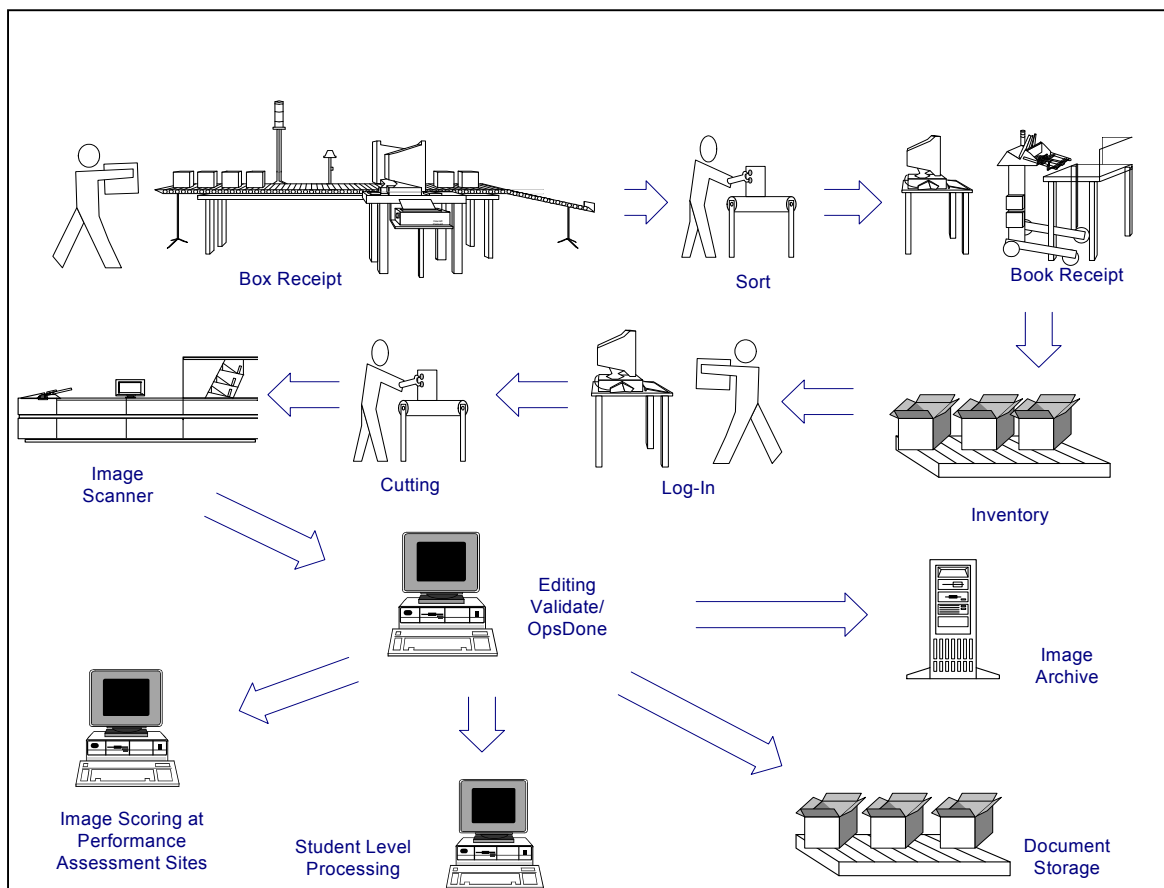
Table 8-1 shows the number of modified booklets received through booklet check-in and the number of modified booklets that contained student responses that were scanned and scored.

Table 8–1. Counts of 2011 PSSA-M Materials Received – Grades 4–8 and 11

Material Description	Booklets Received	Used Booklets Scanned	Total Booklets Shipped
Grade 04 Modified Math	7,199	2,421	7,206
Grade 04 Modified Reading	7,823	3,435	7,827
Grade 05 Modified Math	7,840	3,423	7,840
Grade 05 Modified Reading	8,128	3,997	8,132
Grade 06 Modified Math	7,216	3,668	7,220
Grade 06 Modified Reading	7,469	4,040	7,472
Grade 07 Modified Math	7,165	4,054	7,169
Grade 07 Modified Reading	7,216	4,050	7,216
Grade 08 Modified Math	7,462	4,198	7,463
Grade 08 Modified Reading	7,174	3,712	7,176
Grade 08 Modified Science	6,473	3,329	6,473
Grade 11 Modified Math	7,783	4,394	7,796
Grade 11 Modified Reading	7,436	4,026	7,441
Grade 11 Modified Science	6,990	3,709	6,991

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 scaled scores were calculated.

Student answers were 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

The reading and science items that were part of the 2011 PSSA-M were field tested in 2010. The mathematics items that were part of the 2011 PSSA-M were field tested in either 2009 or 2010. Rangefinding took place after each field test administration. What follows is a description of DRC's rangefinding process for the 2011 PSSA-M items (though those items underwent rangefinding in previous years).

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 math, the 0–3 item-specific scoring guidelines for modified reading, and the 0–2 item-specific scoring guidelines for modified science.

Reading and science responses were pulled from the 2010 modified field tests. Mathematics responses were pulled from the 2009 and 2010 modified field tests. Once examples of responses representing all the score points were selected for each item, sets were assembled for rangefinding. 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 Reading Rangefinding Meeting was held July 7 and 8, 2010, at the Hilton, Harrisburg. Modified Math Rangefinding Meetings were held on July 7, 2010, at the Hilton, Harrisburg, and from June 30 through July 2, 2009, at the Holiday Inn, New Cumberland. The Modified Science Rangefinding Meeting was held July 7 and 8, 2010, at the Hilton, Harrisburg.

Each rangefinding meeting began in a joint session with a review of the history of the assessment and then broke into subject/grade-specific groups. Sets of student responses were presented to the committees one item at a time. Each committee initially reviewed and scored student samples as a group to ensure that everyone was interpreting the scoring guidelines consistently. Committee members then went on to score responses independently. For each student response, committee members' scores were discussed until a consensus was reached. Only those responses for which there was strong agreement among committee members were chosen for inclusion in training materials for DRC raters.

Discussions of student responses included the mandatory use of scoring guideline language. This ensured that committee members remained focused on the specific requirements of each score level. DRC PAS staff took notes addressing how and why the committees arrived at score point decisions, and this information was used by the scoring directors in rater training.

DRC and PDE discussed scoring guideline edits suggested by the rangefinding committees. Changes approved by PDE were then incorporated into the scoring guidelines by DRC Test Development staff. The edited scoring guidelines were used in the preparation of materials and the training of raters.

RATER RECRUITMENT/QUALIFICATIONS

DRC retains a number of raters from year to year. This pool of experienced raters was drawn from to staff the scoring of the 2011 PSSA (including the modified assessment). To complete the rater staffing for this project, DRC placed advertisements in local newspapers and utilized a variety of websites. Open houses were held and applications for rater positions were screened by DRC's recruiting staff. Next, candidates were personally interviewed by DRC staff. In addition, each candidate was required to provide an on-demand writing sample, an on-demand math sample, 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 degrees emphasizing expertise in mathematics, reading, or science. Thus, the rater pool consisted of educators and other professionals with content-specific backgrounds. These individuals were valued for their content-specific knowledge, but they were required to set aside their own biases about student performance and accept the scoring standards outlined in the PSSA and PSSA-M.

LEADERSHIP RECRUITMENT/QUALIFICATIONS

Scoring directors and team leaders were selected by content specialists from a pool of employees who displayed expertise as raters and leaders on previous DRC projects. These individuals had strong backgrounds in mathematics, reading, or science and demonstrated organizational, leadership, and management skills. A majority of scoring directors and team leaders had at least five years of leadership experience working on large-scale assessments, including the PSSA. All scoring directors, team leaders, and raters were required to sign confidentiality agreements before handling secure materials.

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

Team leaders assisted the scoring director with rater training by leading their teams in small group discussions and answering individual questions that raters may not have felt comfortable asking in a large group. Once raters were qualified, team leaders were responsible for maintaining the accuracy and workload of each team member. Ongoing monitoring identified those individuals having difficulty scoring accurately. These raters received one-on-one retraining from the team leader. Any rater who could not be successfully retrained had his/her scores dropped and was released from the project.

TRAINING

As part of preparation for the 2011 modified mathematics, reading, and science assessments, DRC's PAS staff assembled the PDE-approved scoring guidelines and scored student responses approved by rangefinding committees into sets used for training raters. The item-specific scoring guidelines served as the raters' constant reference. Responses that were relevant in terms of the scoring concepts they illustrated were annotated and included in an anchor set. The full range of each score point was clearly represented and annotated in the anchor set, which was used for reference by raters throughout the project.

Training sets and qualifying sets contained student responses reviewed by rangefinding committee members. Raters were instructed on how to apply the scoring guidelines and were required to demonstrate a clear comprehension of each anchor set by performing well on the associated training materials. Responses were selected for training to show raters the range of each score point (e.g., high, mid, and low 2s). Examples of 0s were included in all subjects. This process helped raters recognize the various ways that a student could respond in order to earn each score point outlined and defined in the item-specific scoring guidelines.

The scoring director conducted a team leader training session before training the raters. This session followed the same procedures as rater training, but qualifying standards were more stringent due to the extra responsibilities required of team leaders. During team leader training, all PSSA-M materials were reviewed and discussed. Team leaders were required to annotate all of their training materials with committee justifications from the rangefinding meetings. To facilitate scoring consistency, it was imperative that all team leaders imparted the same rationale for each response. Once the team leaders were qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader per each 8–10 raters ensured a sufficient monitoring rate for each team member.

The 2011 assessment included the opportunity for students to respond in Spanish to modified math and science items. The scoring director responsible for this component of the assessment was a native Spanish speaker who has strong mathematics and science background and who has worked with the PSSA for over ten years. All Spanish raters were bilingual and hired specifically to score the Spanish portion of the assessment. They were required to meet the same training and scoring standards set for the raters of the English version of the assessment.

Rater training began with the scoring director providing an intensive review of the scoring guidelines and anchor papers. Next, raters practiced by independently scoring the responses in the training sets. After each training set, the scoring director or team leaders led a thorough discussion of the responses, either in a large-group or small-group setting.

Once the scoring guidelines, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. Raters who failed to achieve 70 percent exact agreement on the first qualifying set were given additional individual training. Raters 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 individuals were removed from the pool of potential raters in DRC's imaging system and released from the project.

HANDSCORING PROCESS

Student responses were scored independently. All responses were scored once, and ten percent of the responses were scored a second time. The data collected from the ten percent double read portion was used to calculate the exact and adjacent agreement rates in the Scoring Summary Reports. The responses that were used for the ten percent read behind were randomly chosen by the imaging system at the item level. Additional read behinds by the team leaders and scoring directors were done to further ensure reliability.

Raters scored the imaged student responses on PC monitors at DRC Scoring Centers in Sharonville and Columbus, Ohio; Plymouth and Woodbury, Minnesota; Pittsburgh, Pennsylvania; and Austin, Texas. Raters were seated at tables with two imaging stations at each table. Image distribution was controlled, ensuring that student responses were sent only to designated groups of raters qualified to score those items. Imaged student responses were electronically separated for routing to individual raters by item. Raters were only provided with student responses that they were qualified to score. Scores were keyed into DRC's imaging system.

To handle possible alerts (i.e., student responses indicating potential issues related to students' safety and well-being that sometimes require attention at the state or local level), DRC's imaging system allows raters 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 students' responses or any other identifying information about the students. At no time in the alerts process do raters acquire any knowledge concerning a student's personal identity. There were no student response alerts during the scoring of the 2011 modified assessment.

HANDSCORING VALIDITY PROCESS

One of the training tools PAS utilized to ensure rater 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 raters score student responses in a manner consistent with statewide standards both within a single administration of the PSSA-M and across consecutive administrations. In scoring the 2011 PSSA-M, 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 each modified subject selected 40 validity papers for each core open-ended response item. These 40 papers were drawn from a pool of exemplars (responses that are representative of a particular score point and have been verified by the scoring director and the content specialist). The scores on validity papers are considered true scores.

The validity papers were then implemented to test rater accuracy. The responses were scanned into the imaging system and dispersed intermittently to the raters. By the end of the project, raters had scored all 40 validity papers for any items they were qualified to score. Raters were unaware that they were being dealt 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 raters who received validity papers had already successfully completed the training/qualifying process.

Next, the scores that the raters assigned to the validity papers were compared to the true scores in order to determine the validity of the raters' 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 sort of data was also computed for each specific rater. This data was accessed through the Validity Reader Detail Report. Both of these may be run as daily or cumulative reports.

The Validity Reader Detail Report was used to identify particular raters for retraining. If a rater 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. Any time a rater's validity agreement rate fell below 70 percent, the scoring director was cued to examine that rater's scoring. First, the scoring director attempted to ascertain what kind of validity papers the rater was scoring incorrectly. This was done to determine whether there was any sort of a trend (e.g., trending low on the 1–2 line). Once the source of the low agreement was determined, the rater was retrained. If it was determined that the rater 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 raters.

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 3 was given a score of 2 by a significant number of raters within the room, 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 re-training, usually in the form of an annotated handout or a short set of papers without printed scores given to raters as a recalibration test.

Validity was employed on all core modified reading, math, and science CR items. Each 40-paper validity set was formulated to mirror 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 raters were tested on the full spectrum of response types.

The exact rater 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, mid, and low 2s. 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 raters 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 response might teach would apply only to that particular paper. 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.

QUALITY CONTROL

Rater accuracy was monitored throughout the scoring session by means of daily and on-demand reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Inter-rater 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 scoring centers, where they 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 modified open-ended items responses:

The Scoring Summary Report (includes two related reports)

1. The Reader Monitor Report monitored how often raters 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-rater agreement on the ten 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 rater had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each rater 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., not read, complete, awaiting supervisor review, etc.). 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 rater. This report was useful if any responses needed rescoring because of possible rater drift.

The Validity Reports (addressed on previous page) tracked how raters performed by comparing pre-scored responses to raters' scores for the same responses. If a rater's scoring fell below the 70 percent determined agreement rate, remediation occurred. Raters 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 rater reliability. Team leaders read randomly-selected, scored items from each team member. If the team leader disagreed with a rater'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 rater.

Recalibration Sets were used throughout the scoring sessions to ensure accuracy by comparing each rater's scores with the true scores on a pre-selected set of responses. Recalibration sets helped to refocus raters on Pennsylvania scoring standards. This check made sure there was no change in the scoring pattern as the project progressed. Raters failing to achieve 70 percent agreement with the recalibration true scores were given additional training to achieve the highest degree of accuracy possible. Raters 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 raters on the core open-ended responses for the modified mathematics items in the 2011 PSSA. All student responses were read once, and ten percent of responses were read a second time. The data collected from this ten percent double read was used to calculate the exact and adjacent agreement rates.

**Table 8–2. Inter-rater Agreement for 2011 PSSA Modified
Mathematics Grades 4–8 and 11
Open-Ended Response Items and Validity**

Grade	Common Item	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent Agreement	% Exact Validity Agreement
4	1	97	3	100	94
	2	97	3	100	99
5	1	94	6	100	96
	2	96	4	100	98
6	1	97	3	100	97
	2	92	8	100	99
7	1	89	10	99	92
	2	92	8	100	93
8	1	94	6	100	96
	2	99	1	100	97
11	1	92	8	100	94
	2	88	12	100	92

Note. 0–4 possible score points

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.

**Table 8–3. Percentages Awarded for Each Possible Score Point
2011 PSSA Modified Mathematics Grades 4–8 and 11**

Grade	Common Item	%0	%1	%2	%3	%4	%B/NS*
4	1	15	12	23	18	32	0
	2	16	27	19	28	10	0
5	1	40	11	26	17	5	1
	2	7	15	18	24	34	1
6	1	16	50	25	6	4	1
	2	16	43	29	9	2	1
7	1	18	44	21	8	9	1
	2	13	23	24	24	15	1
8	1	58	14	22	2	3	1
	2	14	22	30	27	6	0
11	1	39	29	15	12	0	5
	2	15	40	22	13	5	5

*B=blank and NS=non-scoreable

Table 8–4 shows exact and adjacent agreement rates of raters on the core open-ended responses for the modified reading items in the 2011 PSSA. All student responses were read once and ten percent of responses were read a second time. The data collected from this ten percent double read was used to calculate the exact and adjacent agreement rates.

Table 8–4. Inter-rater Agreement for 2011 PSSA Modified Reading Grades 4–8 and 11 Open-Ended Response Items and Validity

Grade	Common Item	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent Agreement	% Exact Validity Agreement
4	1	85	15	100	81
	2	92	8	100	91
5	1	78	22	100	81
	2	82	18	100	84
6	1	68	32	100	73
	2	73	27	100	80
7	1	78	21	99	89
	2	71	28	99	81
8	1	73	27	100	81
	2	76	24	100	84
11	1	85	15	100	87
	2	79	21	100	77

Note. 0–3 possible score points

Table 8–5 shows the distribution of scores for the modified reading items. All modified reading items are scored with a 0–3 score point range.

**Table 8–5. Percentages Awarded for Each Possible Score Point
2011 PSSA Modified Reading Grades 4–8 and 11**

Grade	Common Item	%0	%1	%2	%3	%B/NS*
4	1	22	26	33	17	1
	2	19	21	19	39	2
5	1	9	33	41	15	1
	2	15	36	30	14	4
6	1	12	42	36	9	1
	2	13	35	34	17	1
7	1	24	29	30	16	1
	2	11	41	39	7	1
8	1	5	24	41	29	1
	2	12	35	37	15	2
11	1	6	34	34	22	3
	2	5	42	31	18	3

*B=blank and NS=non-scoreable

Table 8–6 shows exact and adjacent agreement rates of raters on the core open-ended responses for the modified science items in the 2011 PSSA. All student responses were read once and ten percent of responses were read a second time. The data collected from this ten percent double read was used to calculate the exact and adjacent agreement rates.

Table 8–6. Inter-rater Agreement for 2011 PSSA Modified Science Grades 8 and 11 Open-Ended Response Items and Validity

Grade	Common Item	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent Agreement	% Exact Validity Agreement
8	1	84	15	99	95
	2	94	6	100	98
11	1	92	8	100	95
	2	90	10	100	93

Note. 0–2 possible score points

Table 8–7 shows the distribution of scores for the modified science items. All modified science items are scored with a 0–2 score point range.

Table 8–7. Percentages Awarded for Each Possible Score Point 2011 PSSA Modified Science Grades 8 and 11

Grade	Common Item	%0	%1	%2	%B/NS*
8	1	26	38	34	2
	2	22	41	35	2
11	1	17	61	18	5
	2	47	28	19	6

*B=blank and NS=non-scoreable

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 processes (e.g., impact results); and, 3) analyses for the technical report and item banking. Very detailed information regarding the attributes of students used for Adequate Yearly Progress (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 students to complete a minimum of four items (multiple-choice or open-ended items) 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 August 24, 2011.

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 representative of the entire state for these quality checks. Available student data typically suffices as long as there is reasonable variability in the total test scores of students.

For 2011 this data included all public school students who 1) had their MC items scanned and scored by April 9 (mathematics/reading) or April 20 (science) 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 18. The state reporting criteria were preliminary, meaning that attributions and final PIMS³ 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). This data file was used to provide impact results to the Technical Advisory Committee (TAC) during the linking review process.

FINAL DATA

This file included all students who met state reporting criteria and post-AYP appeals (including attempt criteria) by August 18⁴ for all subject areas. The final data was post-appeals data, meaning that schools had an opportunity to correct certain fields within the data during the AYP appeals process (e.g., student ethnicity). All other files contained pre-appeals data. The data banked for this administration and the majority of the results included in this technical report were derived using the final data file.

³ Pennsylvania Information Management System

⁴ The AYP reporting file was delivered to PDE on August 24, 2011. 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	Final
Mathematics	4	1635	2374	2375
	5	2242	3364	3366
	6	2942	3599	3600
	7	2295	3969	3972
	8	2778	4109	4114
	11	2327	4266	4269
Reading	4	2979	3387	3388
	5	3329	3943	3947
	6	3344	3982	3983
	7	3258	3971	3974
	8	2946	3642	3647
	11	3080	3917	3919
Sci.	8	2555	3249	3252
	11	2605	3539	3540

Chapter Ten: Summary Demographic, Program, and Accommodation Data for the 2011 PSSA Modified

ASSESSED STUDENTS

As stated in earlier chapters, the target population for the PSSA-M consists of 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; however, they 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) has a grade-level standards-aligned IEP that clearly documents that the student requires significant instructional accommodations to successfully access grade-level content, and 3) demonstrates persistent academic difficulties with 4) a lack of academic progress. More detailed information on the PSSA-M eligibility criteria may be accessed at www.education.state.pa.us. On the left, select “Programs,” “Programs S–Z,” and then “Special Education.” From the “Special Education” page select “Assessment” to access the relevant documents regarding eligibility.

Results for this chapter are presented in sets of tables for the three PSSA-M subject areas (mathematics, reading, and science). Accompanying each numbered table is a letter (M, R, or S) to designate the subject area. Table set 10–1M, 10–1R, and 10–1S provides a summary of the assessed students for each subject. Presented on the first line is the total number of non-blank answer documents processed by grade level for the 2011 PSSA-M. This number pertains to the total number of records on 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 percentage of students with a PSSA-M score in the subject area, followed by the number and percentage 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 10–2 (M, R, and S). (See the section of this chapter entitled “Composition of Sample Used in Subsequent Tables” for an additional explanation.)

Table 10–1M. Students Assessed on the 2011 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	2,419	3,418	3,665	4,053	4,197	4,390
Students with mathematics scores	2,414 99.8	3,407 99.7	3,652 99.6	4,027 99.4	4,179 99.6	4,322 98.5
Number processed but not assessed (without a total score)	5 0.2	11 0.3	13 0.4	26 0.6	18 0.4	68 1.5
Students with mathematics scores used in state summaries	2,375	3,366	3,600	3,972	4,114	4,269

Table 10–1R. Students Assessed on the 2011 PSSA-M: Reading

	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	3,434	3,989	4,036	4,046	3,711	4,021
Students with reading scores	3,427 99.8	3,979 99.7	4,021 99.6	4,023 99.4	3,692 99.5	3,960 98.5
Number processed but not assessed (without a total score)	7 0.2	10 0.3	15 0.4	23 0.6	19 0.5	61 1.5
Students with reading scores used in state summaries	3,388	3,947	3,983	3,974	3,647	3,919

Table 10–1S. Students Assessed on the 2011 PSSA-M: Science

	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Number of non-blank answer documents processed	3,324		3,707	
Students with a science score	3,283	98.8	3,574	96.4
Number processed but not assessed (without a total score)	41	1.2	133	3.6
Students with a science score used in state summaries	3,252		3,540	

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

- Extended absence from school that continued beyond the assessment window
- Being absent without make-up for at least one section of a subject area test
- Failure of a student to meet the attempt criteria on one or more sections of a subject area test with no exclusion code marked by school personnel. For mathematics, reading, and science 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 Tables 10–2M, 10–2R, and 10–2S.

Table 10–2M. Counts of Students without Scores on the 2011 PSSA-M: Mathematics

Reason for Non-Assessment	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 20.0	0 0.0	1 7.7	4 15.4	2 11.1	15 22.1
Absent Without Make-up	0 0.0	1 9.1	1 7.7	2 7.7	1 5.6	4 5.9
Non-Attempt	1 20.0	5 45.5	5 38.5	8 30.8	6 33.3	23 33.8
Medical Emergency	2 40.0	3 27.3	1 7.7	4 15.4	3 16.7	6 8.8
Other Reasons	1 20.0	2 18.2	5 38.5	8 30.8	6 33.3	20 29.4
Total Not Assessed	5	11	13	26	18	68

Table 10–2R. Counts of Students without Scores on the 2011 PSSA-M: Reading

Reason for Non-Assessment	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	0 0.0	0 0.0	1 6.7	2 8.7	2 10.5	15 24.6
Absent Without Make-up	0 0.0	1 10.0	2 13.3	0 0.0	0 0.0	5 8.2
Non-Attempt	3 42.9	5 50.0	5 33.3	11 47.8	11 57.9	20 32.8
Medical Emergency	2 28.6	2 20.0	2 6.7	4 17.4	3 15.8	5 8.2
Other Reasons	2 28.6	2 20.0	5 33.3	6 25.8	3 15.8	16 26.2
Total Not Assessed	7	10	15	23	19	61

Table 10–2S. Counts of Students without Scores on the 2011 PSSA-M: Science

Reason for Non-Assessment	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Extended Absence from School	11	26.8	37	27.8
Absent Without Make-up	6	14.6	29	21.8
Non-Attempt	11	26.8	39	29.3
Medical Emergency	7	17.1	12	9.0
Other Reasons	6	14.6	16	12.0
Total Not Assessed	41		133	

COMPOSITION OF SAMPLE USED IN SUBSEQUENT TABLES

Students included in the following demographic analyses were those who contributed to state summary statistics, using the post-appeals Adequate Yearly Progress (AYP) individual student data file provided to the Pennsylvania Department of Education on August 29, 2011. Students not included in the present state summary data were those who were 1) enrolled in a Pennsylvania school after October 1, 2010, 2) coded as ELL and enrolled after May 7, 2010, 3) a foreign exchange student, 4) home schooled, 5) enrolled in a non-public school, or 6) do not have a subject area test score.

Demographic data for students taking the PSSA-M is presented separately for each subject area in Tables 10–3M, 10–3R, and 10–3S. Results for accommodations received were collected separately by subject area and are presented in separate tables as well. For example, tables involving accommodations for reading (Tables 10–4R, 10–5R, 10–6R, and 10–7R) were calculated for those students having a reading score.

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 Tables 10–3M, 10–3R, and 10–3S. Percentages are based on students with scores in a subject area, which are shown at the bottom of the appropriate table. Included are students receiving education in a non-traditional setting, such as court-agency placement.

In 2011, PSSA-M is based on post-appeals files rather than on the pre-appeals file as in 2010, making it more comparable with the PSSA. A comparison between selected demographic characteristics of PSSA-M and PSSA data was thereby improved. Such comparisons, involving gender and ethnicity, revealed more male and minority (Black/African American, Latino/Hispanic) students received the PSSA-M than those receiving the standard PSSA, regardless of subject area. A detailed account of these comparisons may be found in Appendix O of this report.

Table 10–3M. Demographic Characteristics of Students Taking the 2011 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	956 40.3	1,440 42.8	1,482 41.2	1,580 39.8	1,675 40.7	1,690 39.6
Male	1,415 59.6	1,923 57.1	2,116 58.8	2,384 60.0	2,431 59.1	2,573 60.3
Race/Ethnicity						
American Indian or Alaskan Native	6 0.3	8 0.2	4 0.1	11 0.3	2 0.0	7 0.2
Asian or Pacific Islander	19 0.8	38 1.1	29 0.8	30 0.8	40 1.0	32 0.7
Black/African American non-Hispanic	564 23.7	775 23.0	745 20.7	861 21.7	863 21.0	862 20.2
Latino/Hispanic	263 11.1	352 10.5	367 10.2	339 8.5	394 9.6	326 7.6
White non-Hispanic	1,478 62.2	2,158 64.1	2,401 66.7	2,656 66.9	2,760 67.1	2,993 70.1
Multi-Racial/Ethnic	42 1.8	31 0.9	50 1.4	65 1.6	47 1.1	40 0.9
Educational Category and Other Demographic Groups						
IEP (not gifted)	2,375 100.0	3,366 100.0	3,600 100.0	3,972 100.0	4,114 100.0	4,269 100.0
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	903 38.0	1,139 33.8	970 26.9	794 20.0	866 21.1	682 16.0
Title III Served	85 3.6	103 3.1	112 3.1	80 2.0	90 2.2	41 1.0
Title III Not Served	37 1.6	41 1.2	35 1.0	32 0.8	41 1.0	17 0.4
Migrant Student	3 0.1	2 0.1	4 0.1	0 0.0	3 0.1	0 0.0
ELL (enrolled after 5-7-10)	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
ELL (enrolled before 5-7-10)	122 5.1	145 4.3	148 4.1	114 2.9	132 3.2	58 1.4
Exited ESL/bilingual program and in first year of monitoring	3 0.1	8 0.2	7 0.2	8 0.2	8 0.2	3 0.1
Exited ESL/bilingual program and in second year of monitoring	3 0.1	4 0.1	9 0.3	9 0.2	10 0.2	8 0.2
Former ELL no longer monitored	4 0.2	20 0.6	32 0.9	37 0.9	53 1.3	54 1.3
Economically Disadvantaged	1,510 63.6	2,132 63.3	2,198 61.1	2,331 58.7	2,368 57.6	2,129 49.9

**Table 10–3M (continued). Demographic Characteristics of
Students Taking the 2011 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
Enrollment						
Current Enrollment in school of residence after 10-1-10	61 2.6	77 2.3	97 2.7	107 2.7	85 2.1	130 3.0
Current Enrollment in district of residence after 10-1-10	25 1.1	49 1.5	55 1.5	55 1.4	61 1.5	83 1.9
Current Enrollment as PA resident after 10-1-10	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Enrolled in school of residence after 10-1-09 but on/before 10-1-10	431 18.1	587 17.4	847 23.5	776 19.5	492 12.0	539 12.6
Enrolled in district of residence after 10-1-09 but on/before 10-1-10	239 10.1	321 9.5	323 9.0	374 9.4	348 8.5	374 8.8
Education in Non-Traditional Settings						
Court/agency placed	6 0.3	3 0.1	6 0.2	11 0.3	22 0.5	50 1.2
Students with mathematics scores used in state summaries	2,375	3,366	3,600	3,972	4,114	4,269

**Table 10–3R. Demographic Characteristics of
Students Taking the 2011 PSSA-M: Reading**

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	1,191 35.2	1,518 38.5	1,480 37.2	1,489 37.5	1,372 37.6	1,420 36.2
Male	2,195 64.8	2,423 61.4	2,501 62.8	2,480 62.4	2,273 62.3	2,484 63.4
Race/Ethnicity						
American Indian or Alaskan Native	7 0.2	10 0.3	5 0.1	11 0.3	3 0.1	7 0.2
Asian or Pacific Islander	29 0.9	47 1.2	29 0.7	45 1.1	42 1.2	32 0.8
Black/African American non-Hispanic	674 19.9	827 21.0	757 19.0	769 19.4	746 20.5	823 21.0
Latino/Hispanic	334 9.9	375 9.5	393 9.9	373 9.4	360 9.9	315 8.0
White non-Hispanic	2,295 67.7	2,636 66.8	2,742 68.8	2,710 68.2	2,457 67.4	2,693 68.7
Multi-Racial/Ethnic	47 1.4	46 1.2	53 1.3	60 1.5	37 1.0	33 0.8

**Table 10–3R (continued). Demographic Characteristics of
Students Taking the 2011 PSSA-M: Reading**

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)	3,388 100.0	3,947 100.0	3,983 100.0	3,974 100.0	3,647 100.0	3,919 100.0
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	1,233 36.4	1,296 32.8	1,049 26.3	822 20.7	803 22.0	632 16.1
Title III Served	118 3.5	108 2.7	118 3.0	95 2.4	85 2.3	41 1.0
Title III Not Served	45 1.3	50 1.3	47 1.2	38 1.0	19 0.6	15 0.4
Migrant Student	2 0.1	3 0.1	3 0.1	0 0.1	0 0.0	0 0.0
ELL (enrolled after 5-7-10)	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
ELL (enrolled before 5-7-10)	163 4.8	158 4.0	165 4.1	135 3.4	121 3.3	56 1.4
Exited ESL/bilingual program and in first year of monitoring	4 0.1	12 0.3	10 0.3	9 0.2	6 0.2	6 0.2
Exited ESL/bilingual program and in second year of monitoring	6 0.2	4 0.1	10 0.3	14 0.4	9 0.2	11 0.3
Former ELL no longer monitored	6 0.2	26 0.7	23 0.6	44 1.1	51 1.4	44 1.1
Economically Disadvantaged	2,050 60.5	2,390 60.6	2,374 59.6	2,314 58.2	2,153 59.0	2,016 51.4
Enrollment						
Current Enrollment in school of residence after 10-1-10	73 2.2	71 1.8	93 2.3	83 2.1	74 2.0	116 3.0
Current Enrollment in district of residence after 10-1-10	36 1.1	46 1.2	54 1.4	43 1.1	53 1.5	72 1.8
Current Enrollment as PA resident after 10-1-10	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
Enrolled in school of residence after 10-1-09 but on/before 10-1-10	538 15.9	650 16.5	949 23.8	779 19.6	467 12.8	496 12.7
Enrolled in district of residence after 10-1-09 but on/before 10-1-10	295 8.7	331 8.4	331 8.3	356 9.0	328 9.0	333 8.5
Education in Non-Traditional Settings						
Court/agency placed	4 0.1	3 0.1	8 0.2	10 0.3	22 0.6	48 1.2
Students with reading scores used in state summaries	3,388	3,947	3,983	3,974	3,647	3,919

**Table 10–3S. Demographic Characteristics of
Students Taking the 2011 PSSA-M: Science**

Demographic or Educational Characteristic	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Gender				
Female	1,242	38.2	1,300	36.7
Male	2,005	61.7	2,233	63.1
Race/Ethnicity				
American Indian or Alaskan Native	4	0.1	10	0.3
Asian or Pacific Islander	41	1.3	28	0.8
Black/African American non-Hispanic	620	19.1	696	19.7
Latino/Hispanic	339	10.4	293	8.3
White non-Hispanic	2,209	67.9	2,466	69.7
Multi-Racial/Ethnic	34	1.0	36	1.0
Educational Category and Other Demographic Groups				
IEP (not gifted)	3,252	100.0	3,540	100.0
Student exited IEP in last 2 years	0	0.0	0	0.0
Title I	646	19.1	560	15.8
Title III - Served	85	2.6	40	1.1
Title III - Not Served	30	0.9	17	0.5
Migrant Student	3	0.1	0	0.0
ELL (enrolled after 5-7-10)	5	0.2	3	0.1
ELL (enrolled before 5-7-10)	110	3.4	53	1.5
Exited ESL/bilingual program and in first year of monitoring	8	0.2	4	0.1
Exited ESL/bilingual program and in second year of monitoring	10	0.3	11	0.3
Former ELL no longer monitored	35	1.1	48	1.4
Economically Disadvantaged	1,857	57.1	1,802	50.9
Enrollment				
Current Enrollment in school of residence after 10-1-10	62	1.9	91	2.6
Current Enrollment in district of residence after 10-1-10	49	1.5	59	1.7
Current Enrollment as PA resident after 10-1-10	0	0.0	0	0.0
Enrolled in school of residence after 10-1-09 but on/before 10-1-10	395	12.1	437	12.3

**Table 10–3S (continued). Demographic Characteristics of
Students Taking the 2011 PSSA-M: Science**

Demographic or Educational Characteristic	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Enrolled in district of residence after 10-1-09 but on/before 10-1-10	282	8.7	303	8.6
Education in Non-Traditional Settings				
Court/agency placed	18	0.6	30	0.8
Students with science scores used in state summaries	3,252		3,540	

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 entitled *PSSA, PSSA-M, and Keystone (paper/pencil) Accommodations Guidelines for Students with IEPs and Students with 504 Plans* was updated for use with the 2011 PSSA and PSSA-M. The manual can be accessed at www.education.state.pa.us. On the left side, select “Programs,” “Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Testing Accommodations & Security.”

The frequency with which these accommodations were utilized is summarized separately by subject area for each accommodation category in Tables 10–4M, 10–4R, and 10–7S. 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.

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 in the 2011 PSSA-M varied by subject as follows: mathematics and science, 13; and reading, 9. As depicted in Tables 10–4M, 10–4R, and 10–4S, the actual frequencies were quite low, generally representing less than five-tenths of one percent of assessed students statewide. The most notable exceptions were test directions read aloud (each subject), and test items/questions read aloud (mathematics and science).

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 in the 2011 PSSA-M varied by subject as follows: mathematics and science, 12, and reading, 9. Tables 10–5M, 10–5R, and 10–5S summarize the frequency with which these accommodations were utilized, most of which are quite low, typically 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 2011. As depicted in Tables 10–6M, 10–6R, and 10–6S, small group testing and testing in a separate setting were the most commonly used accommodations for each subject.

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 2011. As depicted in Tables 10–7M, 10–7R, and 10–7S, the most common accommodation was scheduled extended time.

**Table 10–4M. Incidence of Presentation
Accommodations Received on the 2011 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	0 0.0	0 0.0	1 0.0	3 0.1
Large Print Format	16 0.7	12 0.4	14 0.4	10 0.3	15 0.4	29 0.7
Electronic Screen Reader	0 0.0	1 0.0	0 0.0	0 0.0	0 0.0	7 0.2
Test directions read aloud (provided by live reader)	1,012 42.6	1,436 42.7	1,094 30.4	955 24.0	898 21.8	486 11.4
Test directions signed, interpreted for ELL student, or recorded	27 1.1	21 0.6	16 0.4	19 0.5	22 0.5	19 0.4
Test items/questions read aloud (provided by live reader) or signed	1,508 63.5	2,008 59.7	1,488 41.3	1,246 31.4	1,015 24.7	363 8.5
Test items / questions interpreted for ELL student	16 0.7	13 0.4	14 0.4	11 0.3	16 0.4	8 0.2
Amplification device	5 0.2	6 0.2	7 0.2	1 0.0	1 0.0	1 0.0
Magnification device	0 0.0	1 0.0	1 0.0	2 0.1	0 0.0	7 0.2
Reading windows, reading guides	40 1.7	57 1.7	25 0.7	12 0.3	12 0.3	3 0.1
Other (per <i>Accommodations Guidelines</i>)	48 2.0	67 2.0	76 2.1	87 2.2	79 1.9	72 1.7
Spanish version for mathematics	1 0.0	0 0.0	1 0.0	5 0.1	7 0.2	4 0.1

**Table 10–4R. Incidence of Presentation
Accommodations Received on the 2011 PSSA-M: Reading**

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	2 0.1	0 0.0	1 0.0	0 0.0	1 0.0	1 0.0
Large Print Format	19 0.6	21 0.5	17 0.4	8 0.2	16 0.4	25 0.6
Electronic Screen Reader	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	5 0.1
Test directions read aloud (provided by live reader)	1,421 41.9	1,479 37.5	1,171 29.4	924 23.3	763 20.9	414 10.6
Test directions signed, interpreted for ELL student, or recorded	29 0.9	31 0.8	25 0.6	18 0.5	21 0.6	18 0.5
Amplification device	15 0.4	17 0.4	7 0.2	3 0.1	4 0.1	1 0.0
Magnification device	2 0.1	0 0.0	0 0.0	5 0.1	0 0.0	6 0.2
Reading windows, reading guides	211 6.2	216 5.5	72 1.8	7 0.2	18 0.5	7 0.2
Other (per <i>Accommodations Guidelines</i>)	72 2.1	103 2.6	100 2.5	82 2.1	50 1.4	40 1.0

**Table 10–4S. Incidence of Presentation
Accommodations Received on the 2011 PSSA-M: Science**

Type of Presentation Accommodation	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Braille Format	1	0.0	1	0.0
Large Print Format	19	0.6	10	0.3
Electronic Screen Reader	0	0.2	7	0.2
Test directions read aloud (provided by live reader)	497	15.3	349	9.9
Test directions signed, interpreted for ELL student, or recorded	6	0.2	15	0.4
Test items/questions read aloud (provided by live reader) or signed	655	20.1	258	7.3
Test items/questions interpreted for ELL student	16	0.5	13	0.4
Amplification device	2	0.1	1	0.0

**Table 10–4S (continued). Incidence of Presentation
Accommodations Received on the 2011 PSSA-M: Science**

Type of Presentation Accommodation	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Magnification device	0	0.0	6	0.2
Reading windows, reading guides	0	0.0	1	0.0
Other (per <i>Accommodations Guidelines</i>)	40	1.2	32	0.9
Spanish version for science	8	0.2	4	0.1

**Table 10–5M. Incidence of Response Accommodations
Received on the 2011 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	50 2.1	53 1.6	31 0.9	18 0.5	19 0.5	10 0.2
Test administrator scribed open- ended responses at student's direction	107 4.5	105 3.1	70 1.9	25 0.6	26 0.6	11 0.3
Test administrator transcribed student responses	39 1.6	40 1.2	38 1.1	24 0.6	25 0.6	13 0.3
Qualified interpreter for ELL	1 0.0	0 0.0	0 0.0	0 0.0	1 0.0	0 0.0
Typewriter, word processor or computer	1 0.0	3 0.1	3 0.1	8 0.2	9 0.2	8 0.2
Braille/Notetaker	1 0.0	0 0.0	0 0.0	0 0.0	1 0.0	3 0.1
Augmentative communication device	0 0.0	2 0.1	3 0.1	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	1 0.0	0 0.0	0 0.0	0 0.0	6 0.1
Manipulative	40 1.7	20 0.6	22 0.6	3 0.1	7 0.2	5 0.1
Translation dictionary for ELL students	0 0.0	0 0.0	0 0.0	6 0.2	1 0.0	2 0.0
Other (approved by PDE)	8 0.3	21 0.6	26 0.7	51 1.3	39 0.9	27 0.6

**Table 10–5R. Incidence of Response Accommodations
Received on the 2011 PSSA-M: Reading**

Type of Response 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
Test administrator marked multiple-choice responses at student’s direction	62 1.8	45 1.1	32 0.8	11 0.3	22 0.6	7 0.2
Test administrator scribed open-ended responses at student’s direction	269 7.9	218 5.5	121 3.0	49 1.2	38 1.0	12 0.3
Test administrator transcribed student responses	91 2.7	68 1.7	54 1.4	28 0.7	42 1.2	18 0.5
Typewriter, word processor, or computer	5 0.1	8 0.2	14 0.4	14 0.4	20 0.5	15 0.4
Braille/Notetaker	1 0.0	0 0.0	0 0.0	0 0.0	1 0.0	2 0.1
Augmentative communication device	0 0.0	0 0.0	2 0.1	0 0.0	0 0.0	0 0.0
Audio recording of student responses	0 0.0	1 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	0 0.0	0 0.0	7 0.2
Other (approved by PDE)	20 0.6	22 0.6	25 0.6	35 0.9	34 0.9	21 0.5

**Table 10–5S. Incidence of Response Accommodations
Received on the 2011 PSSA-M: Science**

Type of Response Accommodation	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Test administrator marked multiple-choice responses	21	0.6	5	0.1
Test administrator scribed open-ended responses at student’s direction	29	0.9	11	0.3
Test administrator transcribed student responses	29	0.9	15	0.4
Qualified interpreter for ELL student	0	0.0	0	0.0
Typewriter, word processor, or computer	9	0.3	11	0.3
Braille/Notetaker	1	0.0	0	0.0
Augmentative communication device	0	0.0	0	0.0
Audio recording of student responses	0	0.0	0	0.0
Electronic Screen Reader	0	0.0	6	0.2
Manipulative	1	0.0	0	0.0
Translation dictionary for ELL students	1	0.1	0	0.0
Other (approved by PDE)	34	1.0	22	0.0

**Table 10–6M. Incidence of Setting Accommodations
Received on the 2011 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	3 0.1	5 0.1	6 0.2	7 0.2	9 0.2	6 0.1
Separate Setting	986 41.5	1,322 39.3	1,104 30.7	1,026 25.8	1,109 27.0	936 21.9
Small Group Testing	1,898 79.9	2,631 78.2	2,567 71.3	2,687 67.6	2,581 62.7	2,220 52.0
Other (PDE approved)	16 0.7	45 1.3	47 1.3	10 0.3	17 0.4	10 0.2

**Table 10–6R. Incidence of Setting Accommodations
Received on the 2011 PSSA-M: Reading**

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 0.1	6 0.2	3 0.1	4 0.1	10 0.3	5 0.1
Separate Setting	1,511 44.6	1,586 40.2	1,251 31.4	1,157 26.6	952 26.19	854 21.8
Small Group Testing	2,696 79.6	3,013 76.3	2,852 71.6	2,656 66.8	2,298 63.0	2,023 51.6
Other (PDE approved)	14 0.4	39 1.0	50 1.3	9 0.2	13 0.4	13 0.3

**Table 10–6S. Incidence of Setting Accommodations
Received on the 2011 PSSA-M: Science**

Type of Setting Accommodation	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Hospital/Home Testing	10	0.3	3	0.1
Separate Setting	821	25.2	709	20.0
Small Group Testing	1,859	57.2	1,705	48.2
Other (PDE Approved)	3	0.1	5	0.1

**Table 10–7M. Incidence of Timing Accommodations
Received on the 2011 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	533 22.4	778 23.1	642 17.8	557 14.0	579 14.1	643 15.1
Requested Extended Time	50 2.1	59 1.8	103 2.9	96 2.4	96 2.3	84 2.0
Multiple Test Sessions	118 5.0	170 5.1	143 4.0	159 4.0	171 4.2	165 3.9
Changed Test Schedule	46 1.9	58 1.7	66 1.8	66 1.7	65 1.6	29 0.7

**Table 10–7R. Incidence of Timing Accommodations
Received on the 2011 PSSA-M: Reading**

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	758 22.4	837 21.2	687 17.2	545 143.7	527 14.5	596 15.2
Requested Extended Time	84 2.5	96 2.4	120 3.0	112 2.8	95 2.6	92 2.3
Multiple Test Sessions	154 4.5	204 5.2	141 3.5	160 4.0	144 3.9	157 4.0
Changed Test Schedule	47 1.4	68 1.7	84 2.1	74 1.9	60 1.6	23 0.6

**Table 10–7S. Incidence of Timing Accommodations
Received on the 2011 PSSA-M: Science**

Type of Timing Accommodation	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Scheduled Extended Time	311	9.6	445	12.6
Requested Extended Time	42	1.3	23	.6
Multiple Test Sessions	115	3.5	88	2.5
Changed Test Schedule	10	0.3	4	0.1

ACCOMMODATION RATE

The incidence of students receiving one or more of the available accommodations for each subject area are provided in Tables 10–8M, 10–8R, and 10–8S. The category of Non-Accommodated indicates students who did not receive any accommodations during the testing.

The general pattern of findings for mathematics and reading reveals a consistently high percentage of students receiving an accommodation, which diminished across grade levels. Science also displayed a high incidence of students receiving an accommodation, although this was slightly lower than for the other two subject areas.

Table 10–8M. Accommodation Rate on the 2011 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	329 13.9	500 14.7	751 20.9	1,008 25.4	1,147 27.9	1,754 41.1
Accommodated	2,046 86.1	2,866 85.1	2,849 79.1	2,964 74.6	2,967 72.1	2,515 58.9
	2,375	3,366	3,600	3,972	4,114	4,269

Table 10–8R. Accommodation Rate on the 2011 PSSA-M: Reading

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	486 14.3	694 17.6	849 21.3	1,048 26.4	1,055 28.9	1,629 41.6
Accommodated	2,902 85.7	3,253 82.4	3,134 78.7	2,926 73.6	2,592 71.1	2,290 58.4
	3,388	3,947	3,983	3,974	3,647	3,919

Table 10–8S. Accommodation Rate on the 2011 PSSA-M: Science

Student Subgroup	Gr. 8		Gr. 11	
	N	Pct	N	Pct
Non-Accommodated	1,132	34.8	1,647	46.5
Accommodated	2,120	65.2	1,893	53.5
	3,252		3,540	

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, 10–4R, and 10–7S, the most frequently occurring accommodations for assessed students were as follows:

- 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 Tables 10–9M, 10–9R, and 10–9S. 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 May 7, 2010. All other assessed students, including those who exited an ESL/bilingual program and in the first or second year of monitoring were regarded as non-ELL. Students coded as ELL and enrolled in a U.S. school after May 7, 2010, are excluded from state summary statistics as stated earlier in this chapter.

Because the combination of accommodations listed in the tables and grades assessed differs somewhat by subject area, it is useful to reference the number of instances of accommodations for which the results in the tables apply. For example, mathematics with five accommodations displayed and six assessed grade levels results in 30 possible instances. There are 24 instances for reading and 10 for science. Of the 64 possible comparisons overall, the non-ELL students received a larger percentage of accommodations in 33 instances, ELL students in 22 instances, and in nine remaining instances the difference was less than one percent.

Several accommodations did show a tendency to occur more frequently for a particular group of students. The most consistent findings are described below:

- Students tested in a separate setting occurred more frequently for the non-ELL group than for the ELL group in 13 of the 14 instances (at six grades for mathematics, five grades for reading, and both grades for science).
- Directions read aloud occurred more frequently for the ELL group in 10 of 14 instances (at three grades for mathematics, all six grades for reading, and one grade for science).
- Test items/questions read aloud or signed, an accommodation available only for mathematics and science, occurred more frequently for the ELL group in six of the eight instances (at five grades for mathematics and one for science).

Tables 10–9M, 10–9R, and 10–9S present the number and percentage of non-ELL and ELL students who received the selected accommodations for each of the assessed grade levels.

**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	956	42.4	56	45.9
Test items/ questions read aloud or signed	1,420	63.0	88	72.1
Tested in separate setting	941	41.8	45	36.9
Small group testing	1,797	79.8	101	82.8
Scheduled extended time	505	22.4	28	23.0
Column N for Gr. 4	2,253		122	
Gr. 5	N	Pct	N	Pct
Test directions read aloud	1,369	42.5	67	46.2
Test items/ questions read aloud or signed	1,920	59.6	88	60.7
Tested in separate setting	1,267	39.3	55	37.9
Small group testing	2,522	78.3	109	75.2
Scheduled extended time	756	23.5	22	15.2
Column N for Gr. 5	3,221		145	
Gr. 6	N	Pct	N	Pct
Test directions read aloud	1,058	30.6	36	24.3
Test items/ questions read aloud or signed	1,417	41.0	71	48.0
Tested in separate setting	1,070	31.0	34	23.0
Small group testing	2,462	71.3	105	70.9
Scheduled extended time	618	17.9	24	16.2
Column N for Gr. 6	3,452		148	
Gr. 7	N	Pct	N	Pct
Test directions read aloud	926	24.0	29	25.4
Test items/ questions read aloud or signed	1,208	31.3	38	33.3
Tested in separate setting	1,002	26.0	24	21.1
Small group testing	2,614	67.8	73	64.0
Scheduled extended time	537	13.9	20	17.5
Column N for Gr. 7	3,858		114	
Gr. 8	N	Pct	N	Pct
Test directions read aloud	875	22.0	23	17.4
Test items/ questions read aloud or signed	975	24.5	40	30.3
Tested in separate setting	1,076	27.0	33	25.0
Small group testing	2,506	62.9	75	56.8
Scheduled extended time	562	14.1	17	12.9
Column N for Gr. 8	3,982		132	

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

Accommodation Received	Non-ELL Students		ELL Students	
	N	Pct	N	Pct
Gr. 11				
Test directions read aloud	479	11.4	7	12.1
Test items/ questions read aloud or signed	360	8.5	3	5.2
Tested in separate setting	925	22.0	11	19.0
Small group testing	2,199	52.2	21	36.2
Scheduled extended time	640	15.2	3	5.2
Column N for Gr. 11	4,211		58	

**Table 10–9R. Incidence of Non-ELL and ELL Students
Receiving Selected Accommodations: Reading**

Accommodation Received	Non-ELL Students		ELL Students	
	N	Pct	N	Pct
Gr. 4				
Test directions read aloud	1,347	41.8	74	45.4
Tested in separate setting	1,437	44.6	74	45.4
Small group testing	2,554	79.2	142	87.1
Scheduled extended time	715	22.2	43	26.4
Column N for Gr. 4	3,225		163	
Gr. 5				
Test directions read aloud	1,413	37.3	66	41.8
Tested in separate setting	1,526	40.3	60	38.0
Small group testing	2,897	76.5	116	73.4
Scheduled extended time	808	21.3	29	18.4
Column N for Gr. 5	3,789		158	
Gr. 6				
Test directions read aloud	1,120	29.3	51	30.9
Tested in separate setting	1,209	31.7	42	25.5
Small group testing	2,733	71.6	119	72.1
Scheduled extended time	656	17.2	31	18.8
Column N for Gr. 6	3,818		165	
Gr. 7				
Test directions read aloud	883	23.0	41	30.4
Tested in separate setting	1,028	26.8	29	21.5
Small group testing	2,566	66.8	90	66.7
Scheduled extended time	527	13.7	18	13.3
Column N for Gr. 7	3,839		135	

**Table 10–9R (continued). Incidence of Non-ELL and ELL Students
Receiving Selected Accommodations: Reading**

Accommodation Received	Non-ELL Students		ELL Students	
	N	Pct	N	Pct
Gr. 8				
Test directions read aloud	736	20.9	27	22.3
Tested in separate setting	924	26.2	28	23.1
Small group testing	2,226	63.1	72	59.5
Scheduled extended time	509	14.4	18	14.9
Column N for Gr. 8	3,526		121	
Gr. 11				
Test directions read aloud	407	10.5	7	12.5
Tested in separate setting	843	21.8	11	19.6
Small group testing	2,004	51.9	19	33.9
Scheduled extended time	594	15.4	2	3.6
Column N for Gr. 11	3,863		56	

**Table 10–9S. Incidence of Non-ELL and ELL Students
Receiving Selected Accommodations: Science**

Accommodation Received	Non-ELL Students		ELL Students	
	N	Pct	N	Pct
Gr. 8				
Test directions read aloud	487	15.5	10	8.7
Test items/ questions read aloud or signed	636	20.3	19	16.5
Tested in separate setting	801	25.5	20	17.4
Small group testing	1,810	57.7	49	42.6
Scheduled extended time	300	9.6	11	9.6
Column N for Gr. 8	3,137		115	
Gr. 11				
Test directions read aloud	342	9.8	7	12.5
Test items/ questions read aloud or signed	253	7.3	5	8.9
Tested in separate setting	700	20.1	9	16.1
Small group testing	1,677	48.1	28	50.0
Scheduled extended time	443	12.7	2	3.6
Column N for Gr. 11	3,484		56	

GLOSSARY OF ACCOMMODATIONS TERMS

Table 10–10 provides brief descriptions of accommodations terms as they are 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 of each accommodation abstracted from the *PSSA, PSSA-M, and Keystone (paper/pencil) Accommodations Guidelines for Students with IEPs and Students with 504 Plans*. This manual can be found at www.education.state.pa.us. On the left, select “Programs,” “Programs O–R,” “Pennsylvania System of School Assessment (PSSA),” and then “Testing Accommodations & Security.”

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 in writing and all items in mathematics and science.
Qualified interpreter for ELL student	An interpreter may translate directions or clarify instructions for the assessments. The interpreter may translate, not define, specific words or test questions on the mathematics and science tests. On the reading, the interpreter may only translate directions and may not translate or define words in the passages or test questions.
Test directions read aloud, signed, or recorded (provided by live reader)	Directions for all PSSA tests may be read aloud, signed, or presented by audio recording.

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

Type of Testing Accommodation	Explanation
Test items/questions read aloud or signed (provided by live reader)	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.
Amplification device	In addition to hearing aids, students may require an amplification device to enhance clarity.
Audio CD Format	An audio CD version of mathematics and science test items/questions may be taken by students with severe hearing disabilities as documented by their IEP.
Other (PDE approval required)	Other presentation accommodations indicated in the <i>Accommodation Guidelines</i> 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	
Braille/Note taker (per <i>Accommodations Guidelines</i>)	Students using this device as part of their regular instructional program may use it on the PSSA; however, it may only be used without a thesaurus, spelling or grammar checker, etc.
Test administrator scribed open-ended responses at student’s direction	A test administrator may record word-for-word what a student dictates 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 and the test administrator will mark the response in the answer booklet).
Test administrator transcribed (copied) student responses (per <i>Accommodations Guidelines</i>)	For writing prompts the test administrator may transcribe handwriting that is extremely difficult to read. On reading, mathematics, or science assessments, 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.

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

Type of Testing Accommodation	Explanation
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 <i>Accommodations Guidelines</i>)	An allowable accommodation as a typing function only for students with the 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.
Audio recording of student responses (per <i>Accommodations Guidelines</i>)	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). It 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 <i>Accommodations Guidelines</i> 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.

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

Type of Testing Accommodation	Explanation
Other (per <i>Accommodations Guidelines</i> 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 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 whether 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 multiple-choice (MC) items and item means for open-ended (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 of students 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 0s and 1s (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. Therefore, 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 the item was relatively easy and/or 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. (As a result of this, some assert that this index would be more accurately 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 (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.⁵ 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., the items 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 of 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 including the following:

- The minimum and maximum values are listed.
- Mean scores are indicated by the red dot.

⁵ 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.

- P_{25} , P_{50} , and P_{75} are indicated by the red lines.
- Marginal “rugs” indicate the density of the individual data points.

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 these scatterplots.

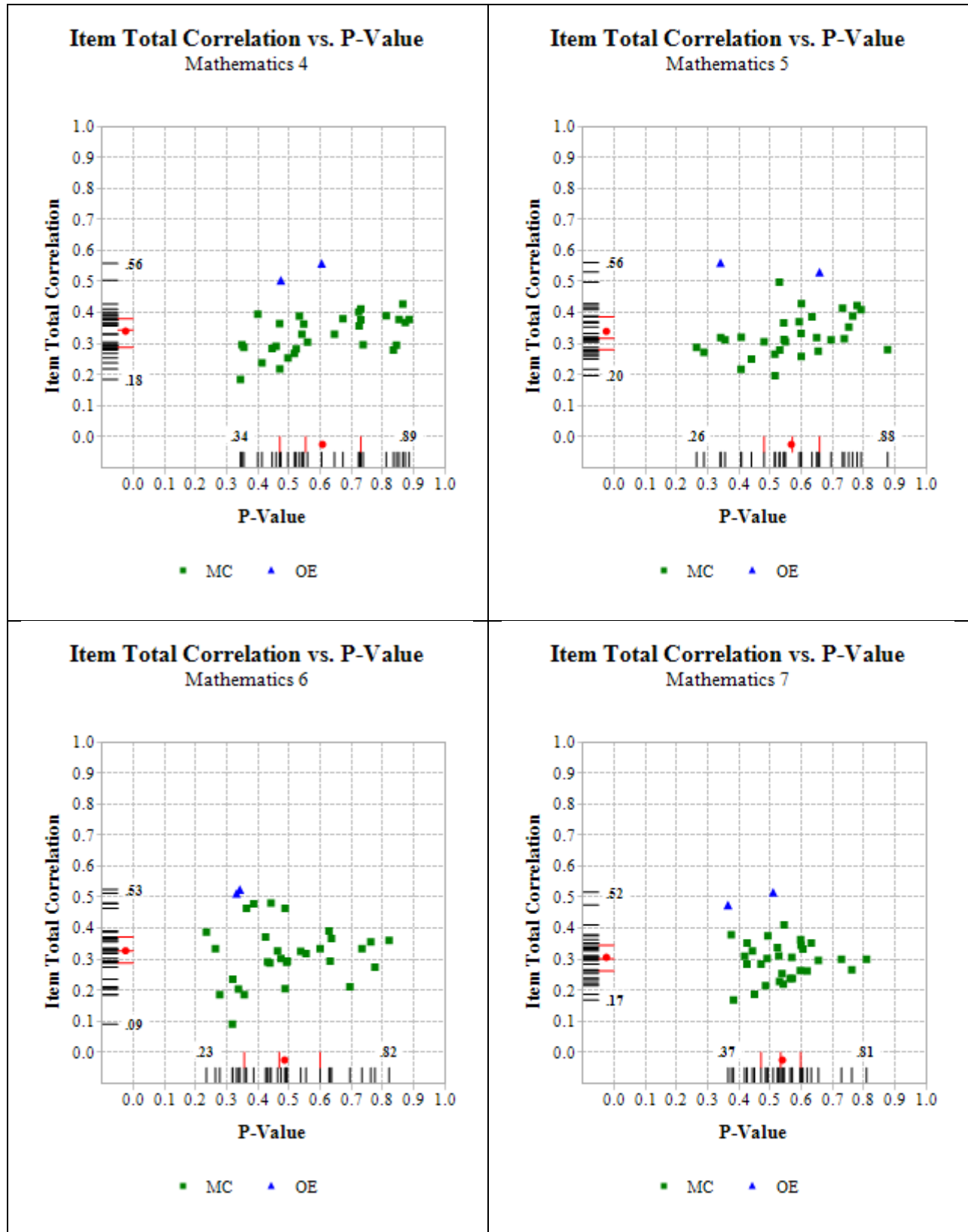
OBSERVATIONS AND INTERPRETATIONS

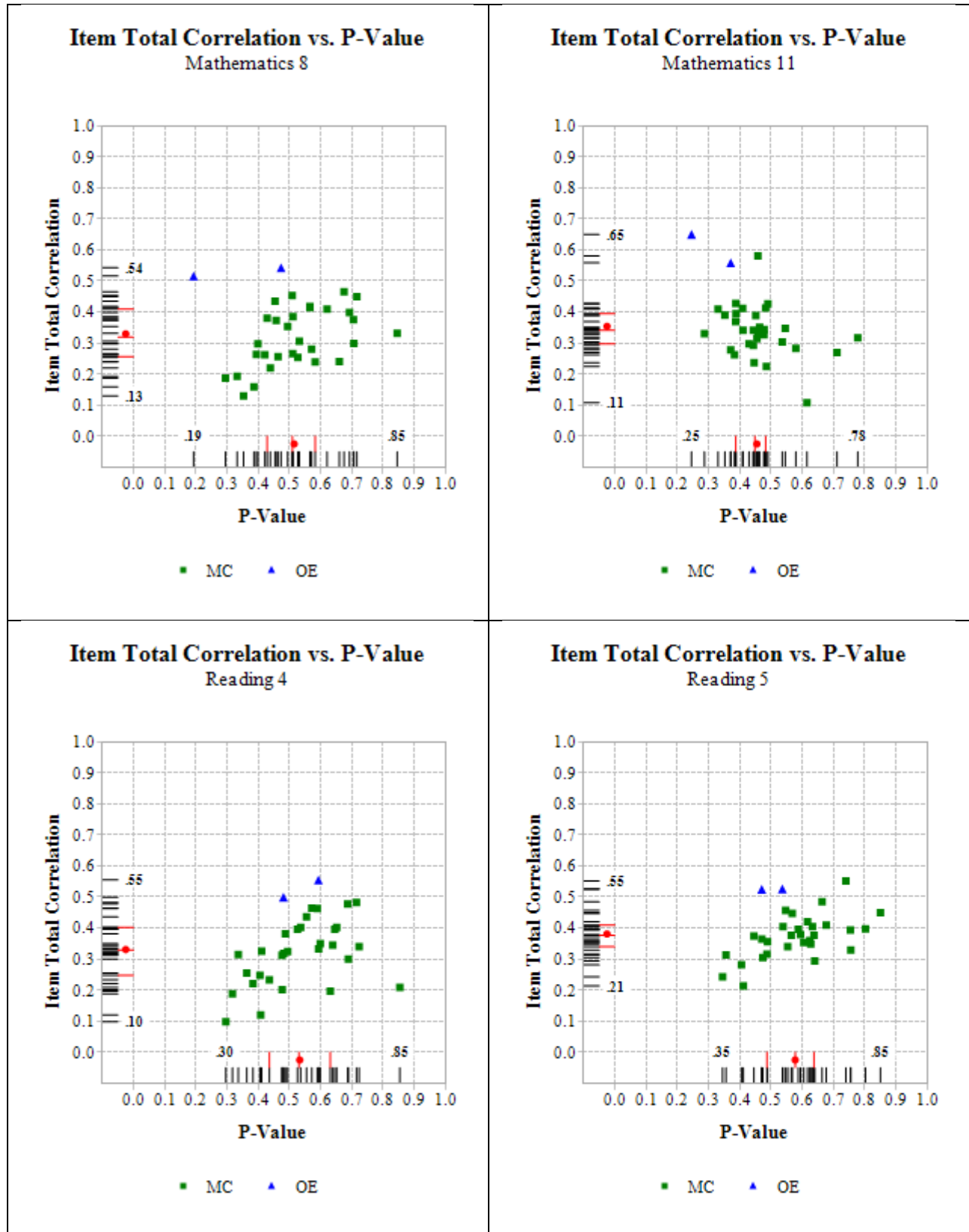
From the difficulty distributions illustrated in the scatterplots, a wide range of item difficulties appeared on each PSSA-M test, which was a desired goal. To support the visuals, Table 11–1 provides break-out results for the MC and OE items. Additional summary statistics (for the MC items only) are provided in Table 11–2. The mean p -values for the MC items ranged from about 0.47–0.65, while the mean proportion-correct values for the OE items ranged from about 0.31–0.58. 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).

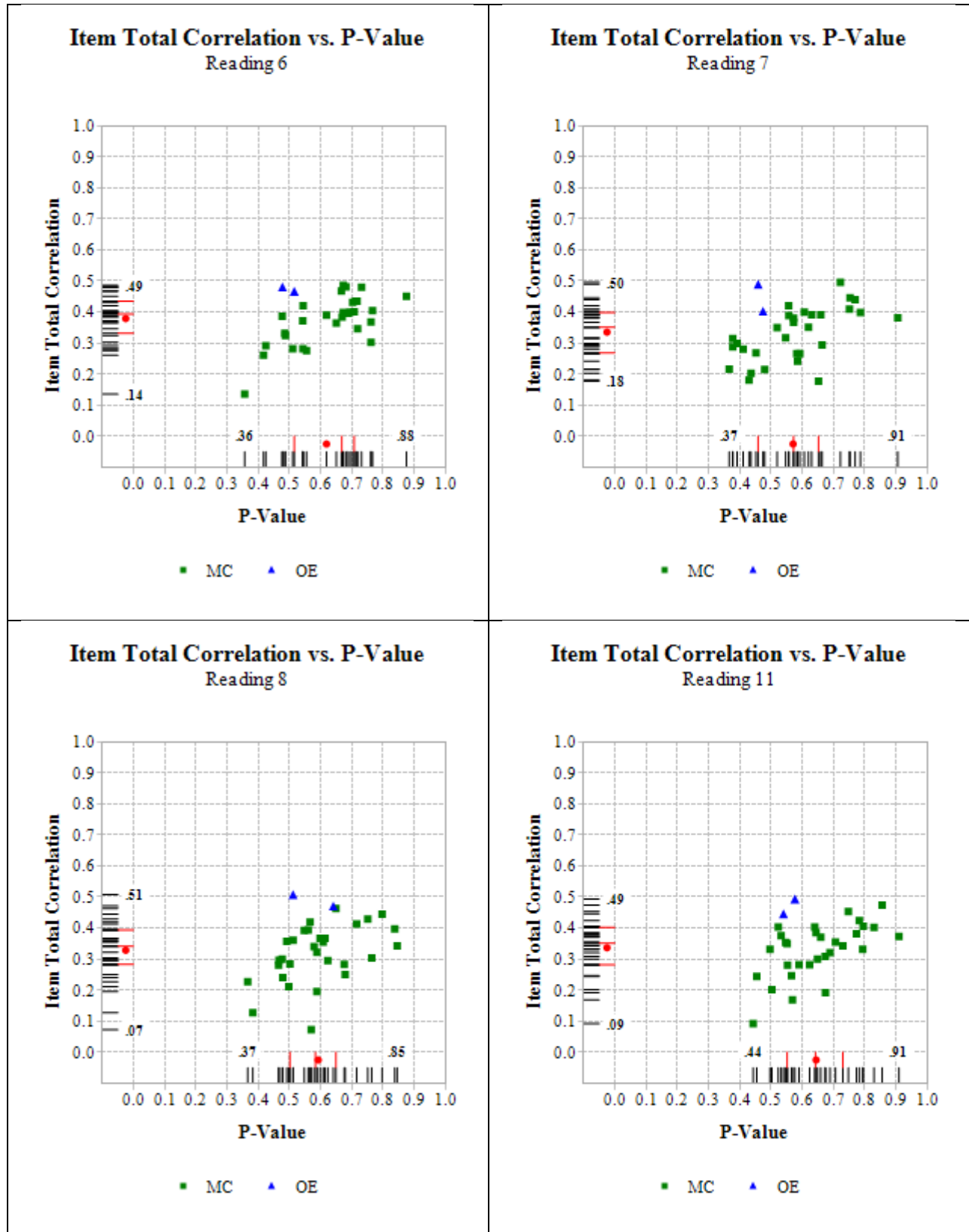
A small number of items had lower item discriminations (e.g., below 0.20). Some of these were observed on items that were very easy or very hard. The mean point-biserial correlations ranged from 0.29–0.37 and 0.36–0.60 for the MC and OE items, respectively. While these values are somewhat lower than 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







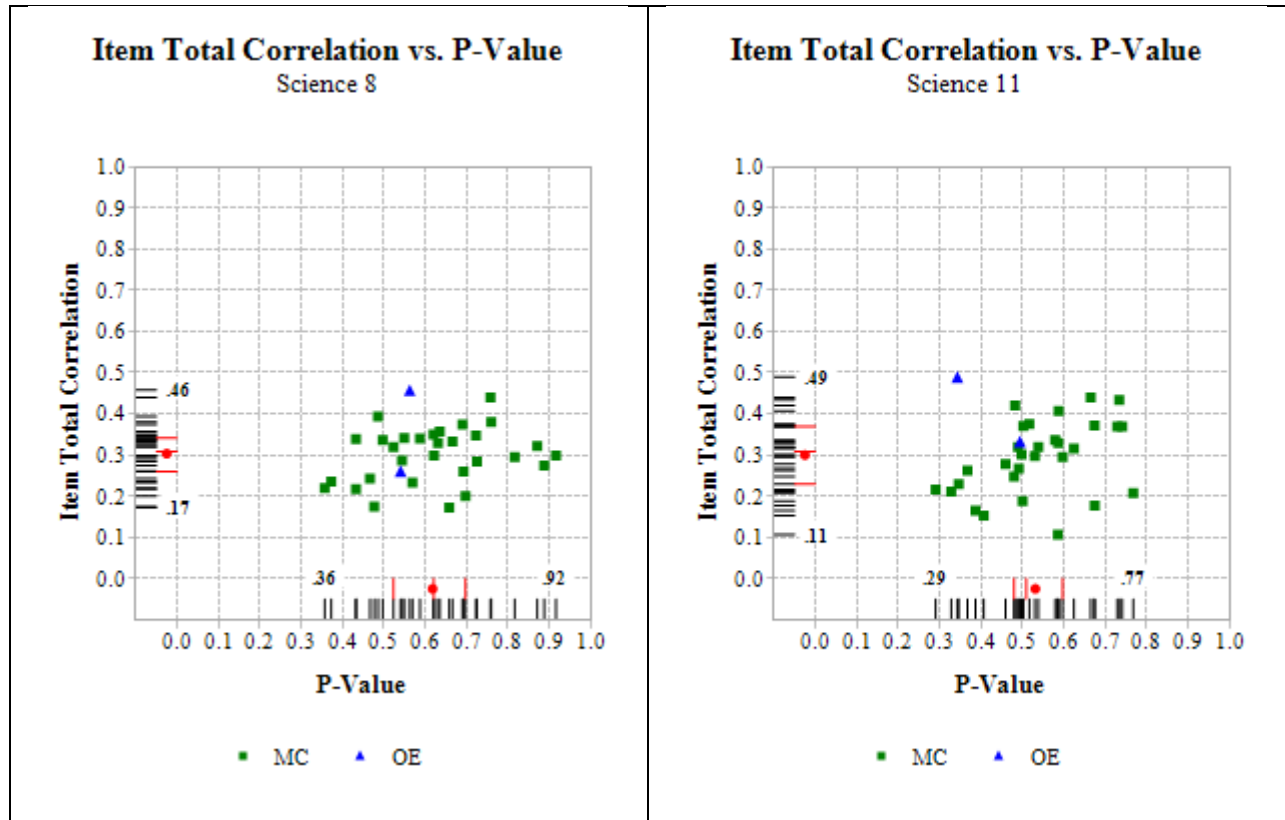


Table 11–1. Sum and Mean Statistics for MC and OE Items

		Multiple-Choice Items				Open-Ended Items			
		Points	Sum	Mean (%/100)	Mean I-T Corr.	Points	Sum	Mean (%/100)	Mean I-T Corr.
Mathematics	4	30	18.361	0.612	0.327	8	4.315	0.539	0.531
	5	30	17.178	0.573	0.326	8	3.998	0.500	0.545
	6	30	14.871	0.496	0.315	8	2.692	0.337	0.519
	7	30	16.379	0.546	0.293	8	3.500	0.437	0.495
	8	30	15.823	0.527	0.316	8	2.664	0.333	0.529
	11	30	13.945	0.465	0.337	8	2.467	0.308	0.604
Reading	4	30	15.987	0.533	0.317	6	3.224	0.537	0.527
	5	30	17.479	0.583	0.371	6	3.023	0.504	0.526
	6	30	18.823	0.627	0.373	6	2.981	0.497	0.474
	7	30	17.338	0.578	0.328	6	2.801	0.467	0.446
	8	30	17.809	0.594	0.317	6	3.459	0.577	0.489
	11	30	19.522	0.651	0.328	6	3.351	0.558	0.470
Sci.	8	30	18.679	0.623	0.300	4	2.208	0.552	0.359
	11	30	16.169	0.539	0.292	4	1.675	0.419	0.411

Note. I-T Corr. is the item-test score correlation. The means for the I-T correlations were not computed using Fisher’s Z transformation. However, this is not expected to affect any conclusions based on these result.

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

		P-Value				Point Biserial			
		Min	Max	Mean	Med	Min	Max	Mean	Med
Mathematics	4	0.34	0.89	0.61	0.55	0.18	0.43	0.33	0.33
	5	0.26	0.88	0.57	0.57	0.20	0.50	0.33	0.31
	6	0.23	0.82	0.50	0.48	0.09	0.48	0.31	0.32
	7	0.38	0.81	0.55	0.54	0.17	0.41	0.29	0.30
	8	0.30	0.85	0.53	0.51	0.13	0.46	0.32	0.30
	11	0.29	0.78	0.46	0.45	0.11	0.58	0.34	0.34
Reading	4	0.30	0.85	0.53	0.53	0.10	0.48	0.32	0.32
	5	0.35	0.85	0.58	0.59	0.21	0.55	0.37	0.37
	6	0.36	0.88	0.63	0.67	0.14	0.49	0.37	0.39
	7	0.37	0.91	0.58	0.58	0.18	0.50	0.33	0.33
	8	0.37	0.85	0.59	0.58	0.07	0.46	0.32	0.33
	11	0.44	0.91	0.65	0.65	0.09	0.47	0.33	0.35
Sci.	8	0.36	0.92	0.62	0.63	0.17	0.44	0.30	0.31
	11	0.29	0.77	0.54	0.52	0.11	0.44	0.29	0.30

ITEM OMIT RATES

Omit rates are analyzed here at the test level (Table 11–3). Individual item omit rates can be found in Appendix I. 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 this analysis. All MC item omit rates were less than 2% and all OE item omit rates were less than 3%. 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. Grade 11 for all subjects had the lowest percent of zero item omits. The PSSA-M would not be considered speeded based on the Swineford (1956) criteria.⁶

Table 11–3. Omit Rates for Tests

		% Attempting 75% of Items	% Attempting All Items
Mathematics	4	100.00	98.27
	5	100.00	98.40
	6	100.00	97.50
	7	99.97	97.18
	8	99.93	96.57
	11	99.84	91.40
Reading	4	99.82	96.25
	5	99.95	96.35
	6	99.87	94.28
	7	99.80	94.61
	8	99.89	94.98
	11	99.85	93.57
Sci.	8	100.00	96.00
	11	99.89	93.79

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

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 all items 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 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 ($x = 0, 1, 2, \dots, m_i$) 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, the Rasch model provides estimates of a person's ability 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

Since 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 were 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' performances. 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., 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 standardized residual setting was used for the PCA because standardized residuals are better for decomposing the unexplained variance into contrasts (Linacre, 2009).

Table 12–1 presents the PCA results for the PSSA-M 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 percentage 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 be seen in Table 12–1, the primary dimension in the Rasch model explained about 29–48 percent of the total variance across Grades 4–8 and 11. If the data fit the model in such a way that only random noise was present, about 29–48 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. According to Reckase (1979), the variance explained by the primary dimension should be greater than 20 percent to indicate unidimensionality. The variance explained for all subjects and grades exceeded this threshold, demonstrating a unidimensional trait of the data.

Another 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 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 1.4 to 1.9. This indicates that the second dimension accounted for only 1.4 to 1.9 units out of 32 units of item residual variance. Smith and Miao (1994) used simulation studies to show that eigenvalues less than 1.4 are at the random level while Raïche (2005) suggested that, on occasion, eigenvalues as high as 2.0 are at the random level. The fact that all eigenvalues of the first contrast are less than 2.0 along with the amount of variance explained by the primary dimension provides reasonable evidence of unidimensionality for all PSSA-M tests.

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

		Eigenvalue	Empirical	Modeled	
Mathematics	4	Total raw variance in observations	61.3	100.0%	100.0%
		Raw variance explained by measures	29.3	47.8%	46.9%
		Raw unexplained variance (total)	32.0	52.2%	53.1%
		Unexplained variance in 1st contrast	1.6	2.6%	
	5	Total raw variance in observations	57.4	100.0%	100.0%
		Raw variance explained by measures	25.4	44.3%	43.8%
		Raw unexplained variance (total)	32.0	55.7%	56.2%
		Unexplained variance in 1st contrast	1.7	2.9%	
	6	Total raw variance in observations	52.5	100.0%	100.0%
		Raw variance explained by measures	20.5	39.0%	38.8%
		Raw unexplained variance (total)	32.0	61.0%	61.2%
Unexplained variance in 1st contrast		1.8	3.4%		
7	Total raw variance in observations	45.3	100.0%	100.0%	
	Raw variance explained by measures	13.3	29.4%	29.0%	
	Raw unexplained variance (total)	32.0	70.6%	71.0%	
	Unexplained variance in 1st contrast	1.9	4.2%		
8	Total raw variance in observations	55.6	100.0%	100.0%	
	Raw variance explained by measures	23.6	42.5%	41.6%	
	Raw unexplained variance (total)	32.0	57.5%	58.4%	
	Unexplained variance in 1st contrast	1.8	3.2%		
11	Total raw variance in observations	49.0	100.0%	100.0%	
	Raw variance explained by measures	17.0	34.7%	33.9%	
	Raw unexplained variance (total)	32.0	65.3%	66.1%	
	Unexplained variance in 1st contrast	1.7	3.5%		

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

		Eigenvalue	Empirical	Modeled		
Reading		Total raw variance in observations	48.2	100.0%	100.0%	
	4		Raw variance explained by measures	16.2	33.6%	33.9%
			Raw unexplained variance (total)	32.0	66.4%	66.1%
			Unexplained variance in 1st contrast	1.6	3.3%	
			Total raw variance in observations	52.2	100.0%	100.0%
	5		Raw variance explained by measures	20.2	38.7%	38.2%
			Raw unexplained variance (total)	32.0	61.3%	61.8%
			Unexplained variance in 1st contrast	1.5	2.8%	
			Total raw variance in observations	52.2	100.0%	100.0%
	6		Raw variance explained by measures	20.2	38.7%	37.9%
			Raw unexplained variance (total)	32.0	61.3%	62.1%
			Unexplained variance in 1st contrast	1.6	3.0%	
			Total raw variance in observations	50.5	100.0%	100.0%
	7		Raw variance explained by measures	18.5	36.7%	35.7%
			Raw unexplained variance (total)	32.0	63.3%	64.3%
			Unexplained variance in 1st contrast	1.5	3.1%	
		Total raw variance in observations	46.8	100.0%	100.0%	
8		Raw variance explained by measures	14.8	31.7%	31.0%	
		Raw unexplained variance (total)	32.0	68.3%	69.0%	
		Unexplained variance in 1st contrast	1.5	3.2%		
		Total raw variance in observations	49.2	100.0%	100.0%	
11		Raw variance explained by measures	17.2	34.9%	33.7%	
		Raw unexplained variance (total)	32.0	65.1%	66.3%	
		Unexplained variance in 1st contrast	1.4	2.9%		

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

		Eigenvalue	Empirical	Modeled		
Science		Total raw variance in observations	51.2	100.0%	100.0%	
	8		Raw variance explained by measures	19.2	37.5%	36.6%
			Raw unexplained variance (total)	32.0	62.5%	63.4%
			Unexplained variance in 1st contrast	1.4	2.8%	
			Total raw variance in observations	45.0	100.0%	100.0%
	11		Raw variance explained by measures	13.0	28.9%	28.9%
			Raw unexplained variance (total)	32.0	71.1%	71.1%
			Unexplained variance in 1st contrast	1.6	3.6%	

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, \dots, X_n is locally independent with respect to the latent variable θ if, for all $x = (x_1, x_2, \dots, x_n)$ and θ ,

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

This formula essentially states that the probability of any pattern of responses across all items (\mathbf{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 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 shown below. (This is a weaker form because higher-order dependencies among items are allowed.) Based on the WLI, the following expression can be derived:

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

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 determine student 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 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 (P₁₀, P₂₅, P₅₀, P₇₅, P₉₀) — 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 tests.

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

Statistic	Mathematics					
	4	5	6	7	8	11
N	496	496	496	496	496	496
Mean	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
SD	0.04	0.04	0.04	0.04	0.04	0.04
Minimum	-0.12	-0.15	-0.11	-0.15	-0.11	-0.16
P ₁₀	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07
P ₂₅	-0.05	-0.05	-0.05	-0.05	-0.06	-0.05
P ₅₀	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04
P ₇₅	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
P ₉₀	0.01	0.01	0.01	0.01	0.02	0.01
Maximum	0.19	0.40	0.20	0.20	0.19	0.15
>0.20	0	1	0	0	0	0

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

Statistic	Reading					
	4	5	6	7	8	11
N	496	496	496	496	496	496
Mean	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
SD	0.03	0.03	0.04	0.03	0.03	0.03
Minimum	-0.12	-0.11	-0.11	-0.11	-0.12	-0.10
P ₁₀	-0.07	-0.07	-0.07	-0.07	-0.07	-0.06
P ₂₅	-0.05	-0.05	-0.05	-0.05	-0.05	-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	-0.01
P ₉₀	0.01	0.01	0.01	0.01	0.01	0.00
Maximum	0.10	0.09	0.23	0.16	0.09	0.10
>0.20	0	0	1	0	0	0

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

Statistic	Science	
	8	11
N	496	496
Mean	-0.03	-0.03
SD	0.03	0.03
Minimum	-0.11	-0.12
P ₁₀	-0.06	-0.07
P ₂₅	-0.05	-0.05
P ₅₀	-0.03	-0.03
P ₇₅	-0.01	-0.01
P ₉₀	0.00	0.00
Maximum	0.12	0.15
>0.20	0	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 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 be 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

		Infit Mean Square					Outfit Mean Square				
		Mean	SD	Min	Max	[0.7,1.3]	Mean	SD	Min	Max	[0.7,1.3]
Mathematics	4	0.99	0.07	0.83	1.12	32/32	0.98	0.14	0.70	1.22	32/32
	5	0.99	0.07	0.86	1.13	32/32	0.99	0.11	0.77	1.18	32/32
	6	0.99	0.08	0.86	1.19	32/32	1.00	0.14	0.76	1.39	31/32
	7	1.00	0.05	0.91	1.10	32/32	0.99	0.06	0.88	1.14	32/32
	8	0.99	0.09	0.85	1.17	32/32	0.99	0.13	0.76	1.24	32/32
	11	1.00	0.08	0.79	1.19	32/32	1.00	0.11	0.74	1.28	32/32
Reading	4	0.99	0.09	0.84	1.18	32/32	1.00	0.13	0.74	1.33	31/32
	5	0.99	0.08	0.79	1.16	32/32	0.99	0.13	0.63	1.25	30/32
	6	0.99	0.08	0.82	1.20	32/32	0.98	0.15	0.57	1.38	30/32
	7	0.99	0.09	0.82	1.14	32/32	0.98	0.14	0.56	1.20	31/32
	8	0.99	0.09	0.85	1.23	32/32	0.99	0.14	0.71	1.34	31/32
	11	0.99	0.09	0.81	1.21	32/32	0.97	0.15	0.63	1.30	30/32
Sci.	8	0.99	0.07	0.86	1.19	32/32	0.98	0.12	0.71	1.21	32/32
	11	0.99	0.08	0.85	1.16	32/32	1.00	0.11	0.77	1.24	32/32

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 minimum and maximum values and standard deviations suggest that the PSSA-M items covered a relatively wide range of difficulties.

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

		N	Mean	SD	Min	Max
Mathematics	4	32	0.32	0.95	-1.42	1.67
	5	32	0.35	0.81	-1.52	1.88
	6	32	0.25	0.80	-1.53	1.54
	7	32	0.16	0.51	-1.24	0.93
	8	32	0.19	0.69	-1.63	1.61
	11	32	0.33	0.59	-1.32	2.08
Reading	4	32	0.00	0.66	-1.81	1.16
	5	32	0.02	0.66	-1.61	1.20
	6	32	0.04	0.67	-1.62	1.38
	7	32	0.04	0.72	-2.14	1.04
	8	32	0.00	0.63	-1.45	1.10
	11	32	0.02	0.68	-1.82	1.05
Sci.	8	32	0.03	0.80	-2.01	1.32
	11	32	0.03	0.61	-1.15	1.19

Note. The mean logit values for mathematics are not necessarily 0.0 because the items have been placed on a scale that was developed in prior years.

VISUALIZING THE *P*-VALUE-LOGIT RELATIONSHIP

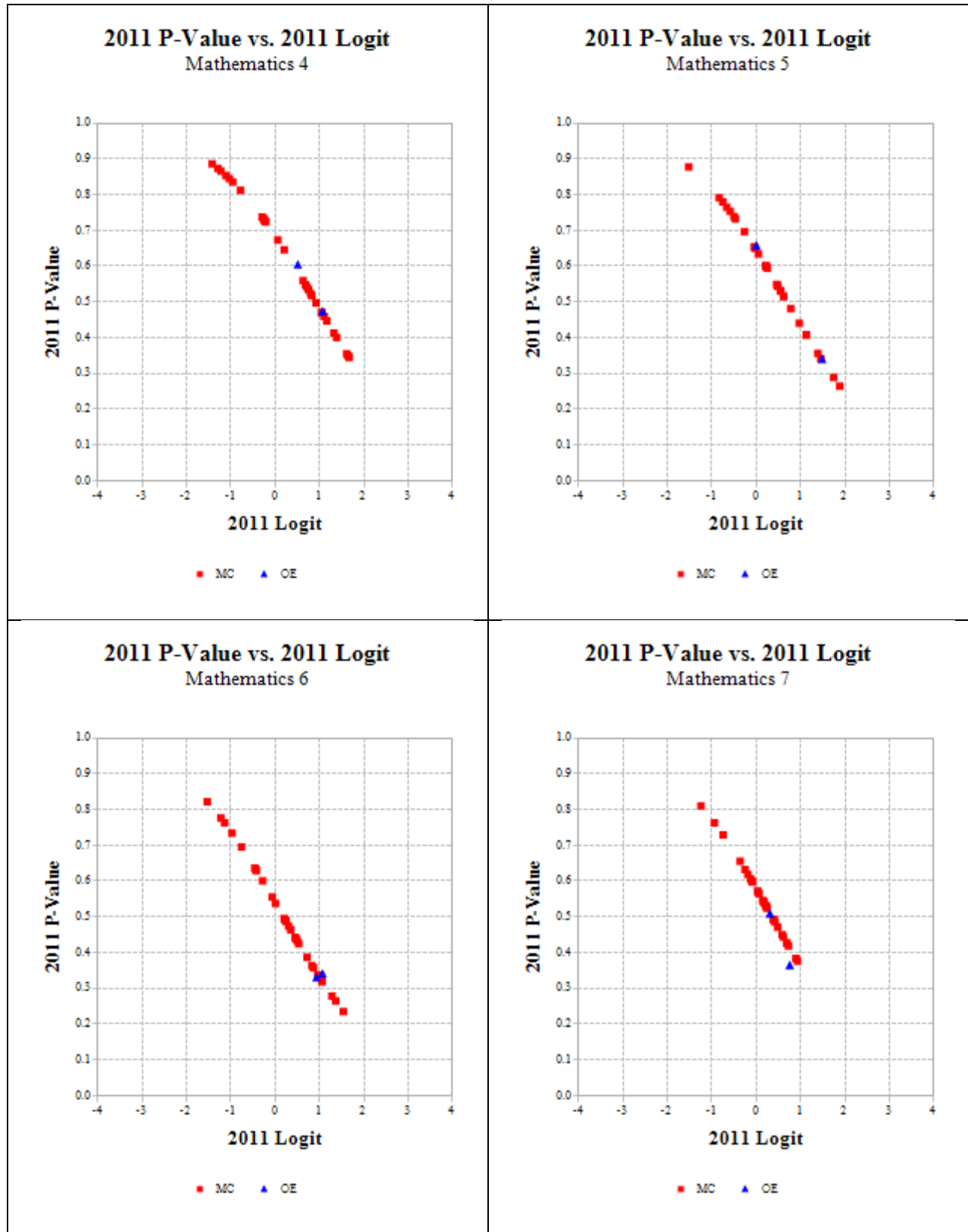
During the PSSA-M administration, test forms were spiraled within classrooms. In effect, students were administered the same set of common items but different nonoperational items (e.g., field test item sets). Cross checks can be made to ensure the calibration and linking processes are reasonable across 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. As a result 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.

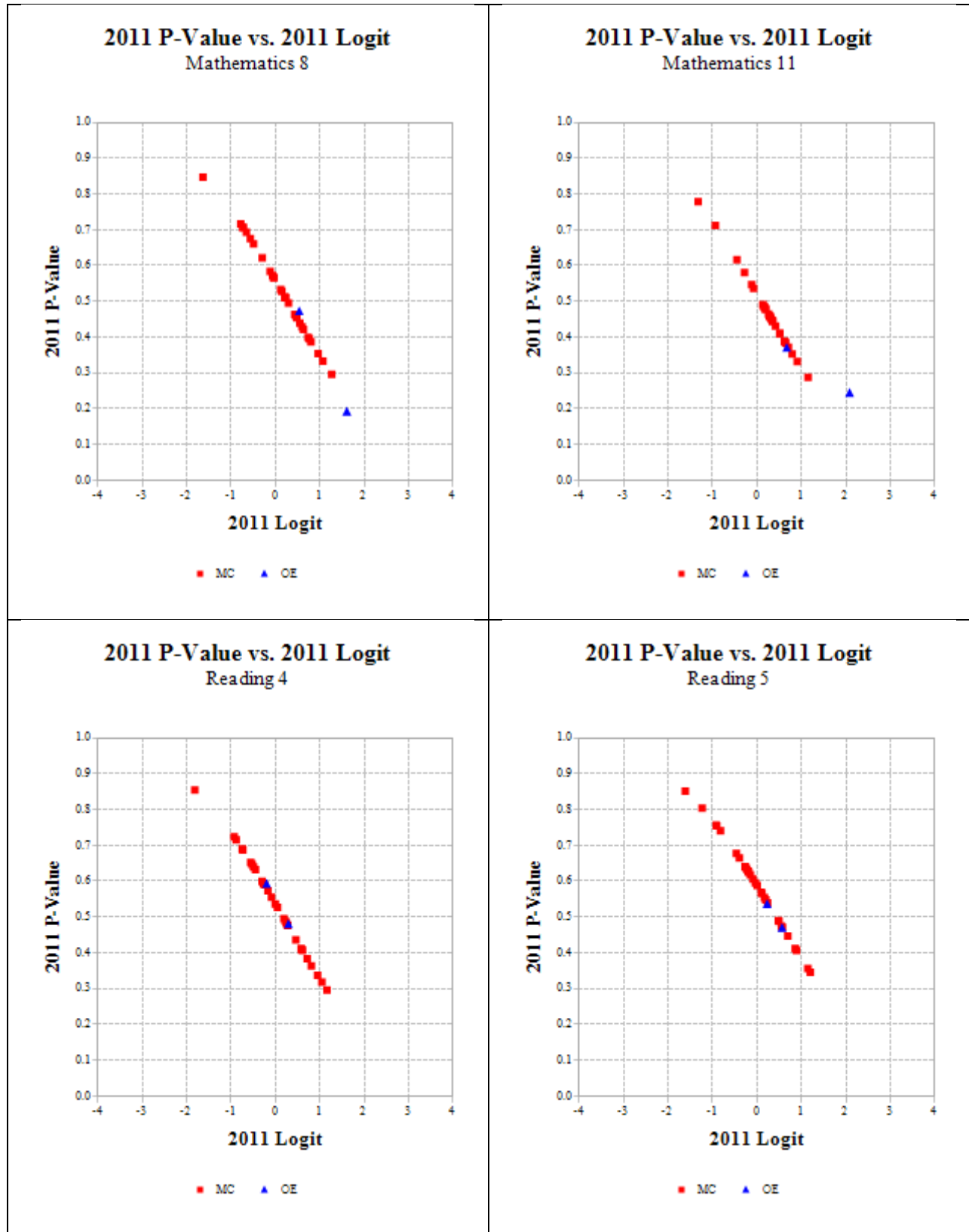
Figure 12–1 shows plots of the *p*-value-logit relationship for the operational items. The curves are nearly linear in the center but curve towards asymptotes of one and zero, respectively, on the left and right. The graphs show that items with lower *p*-values (indicating a more difficult item that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

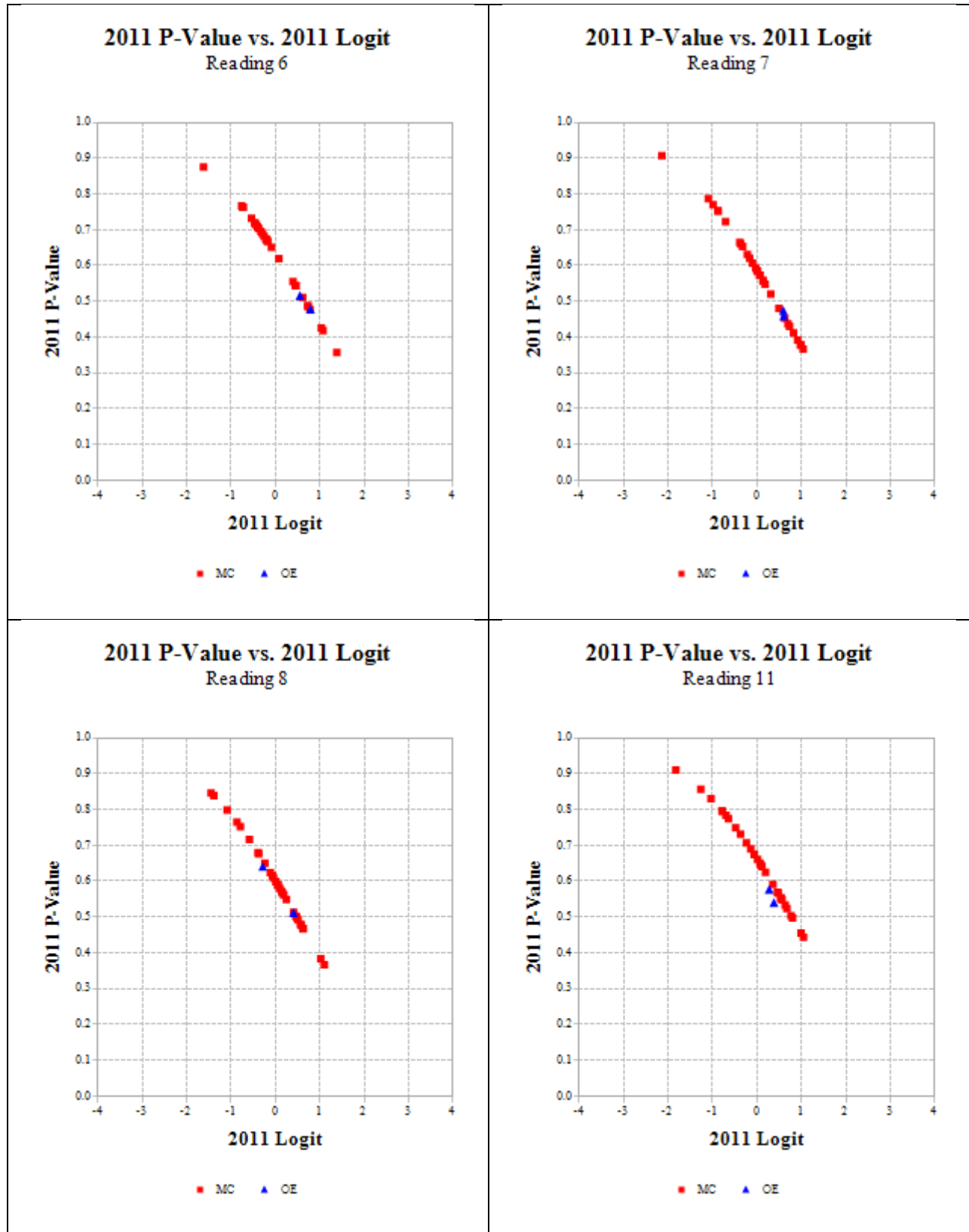
The spread of the graph points is indicative of the dispersion of item difficulties in the operational items. The dispersion and coordinates of items are roughly similar across grades for mathematics.

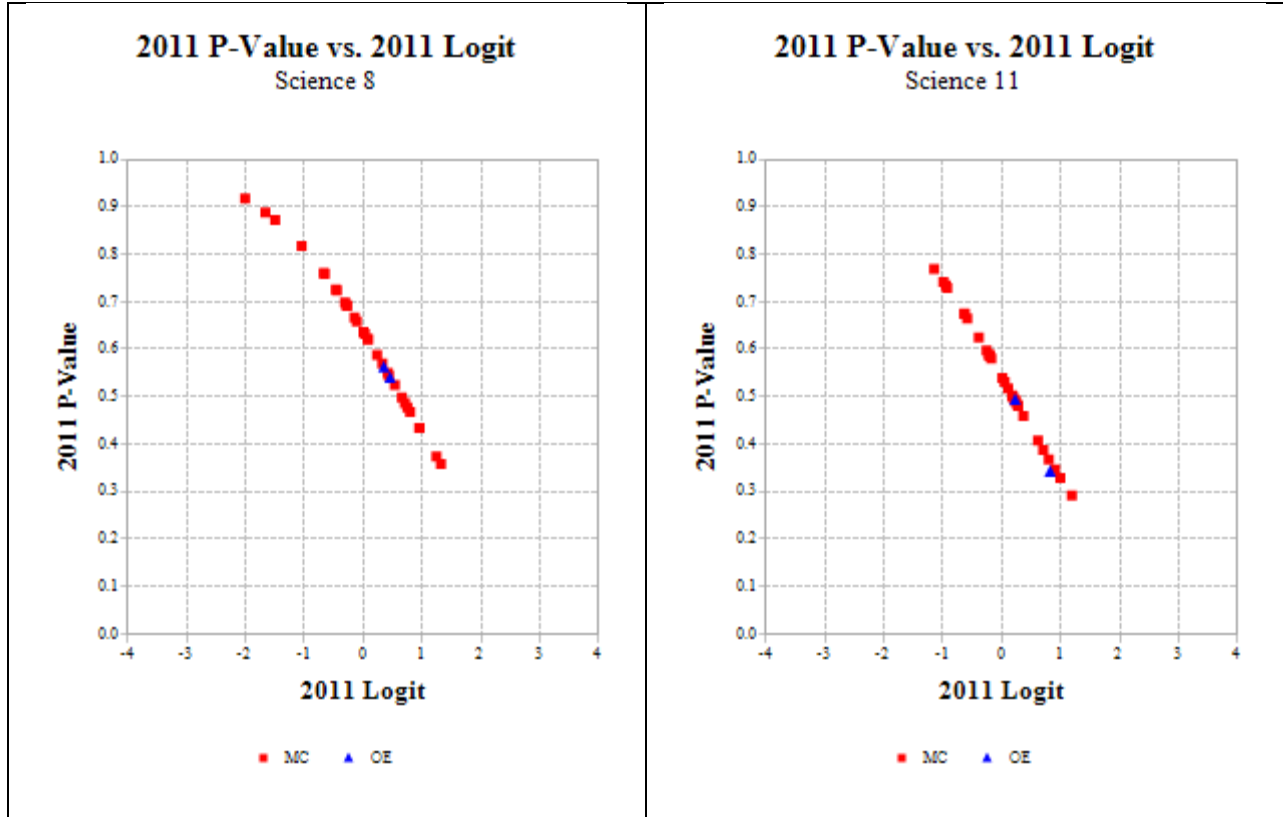
Common OE items are also graphed in Figure 12–1. These items appear with triangular markers. The OE items generally fall on the same curve as the MC items, but subtle differences can occur. The OE items were placed on the MC item difficulty (*p*-value) scale, which ranges from 0.00 to 1.00, by dividing 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. 2011 P-Values on 2011 Logit Values







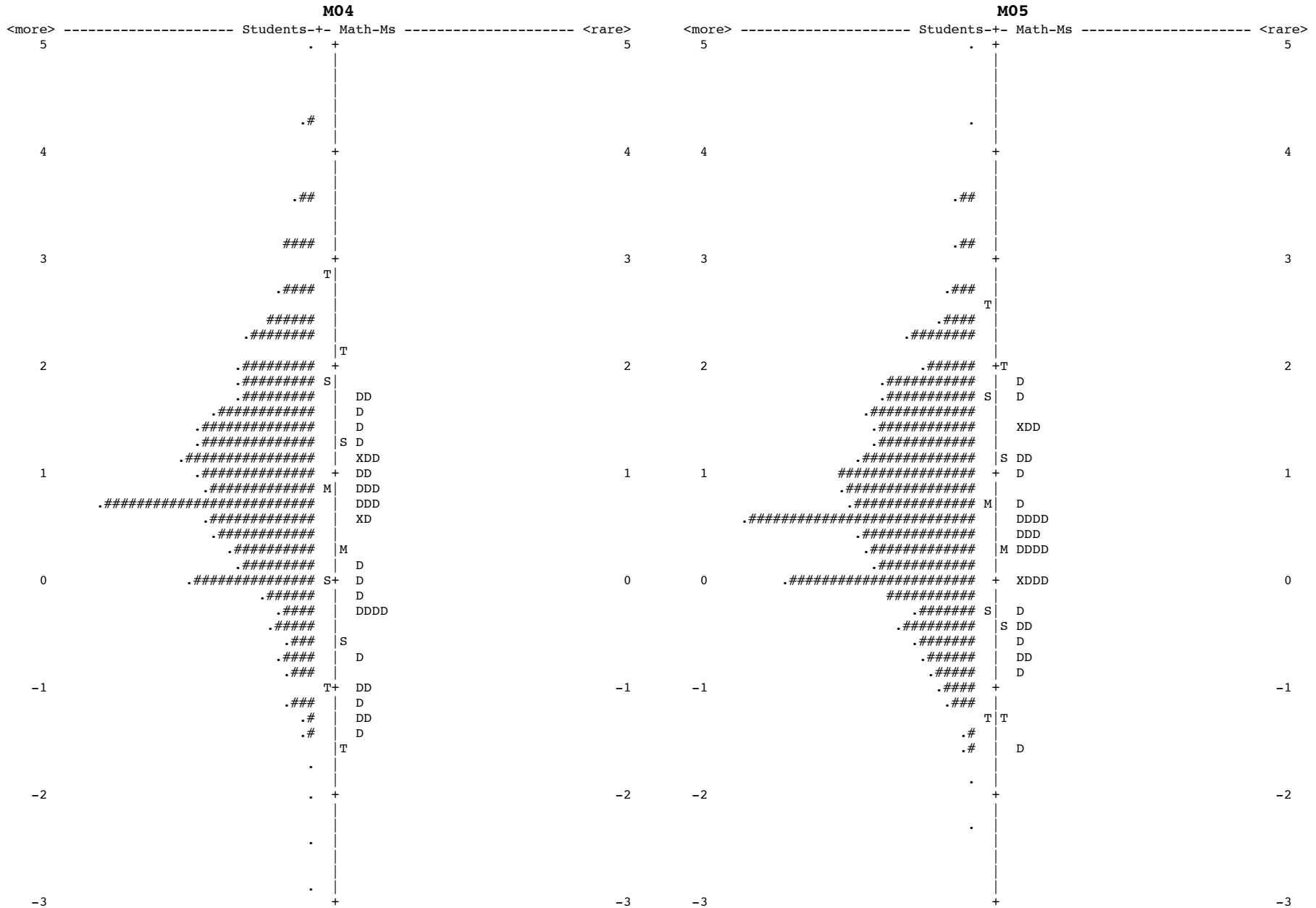


Item Difficulty-Student Ability Maps

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 (Xs) were relatively more difficult than the MC items (Ds). However, the OEs provide more information for higher ability students.

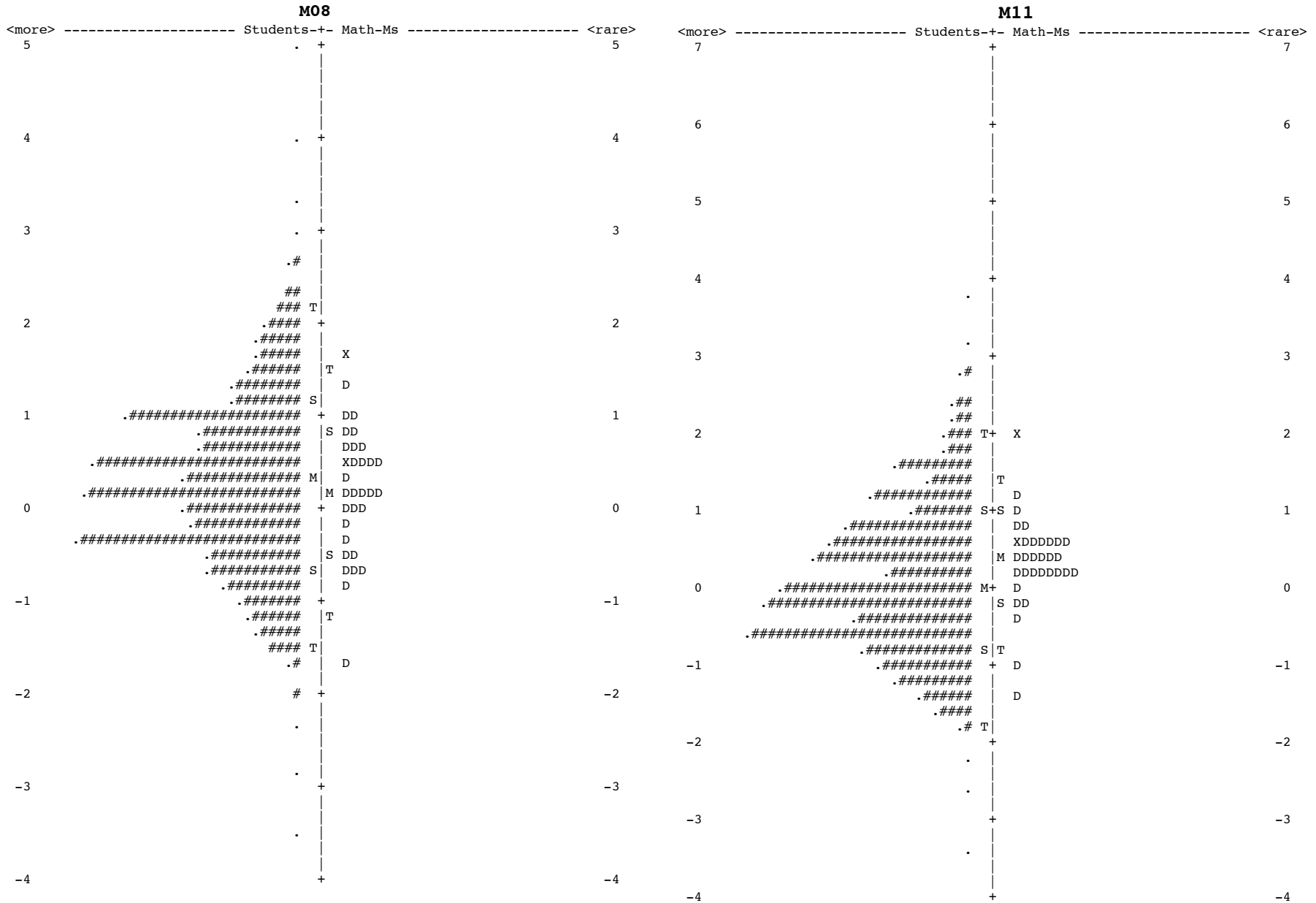
Figure 12–2. Item-Student Maps



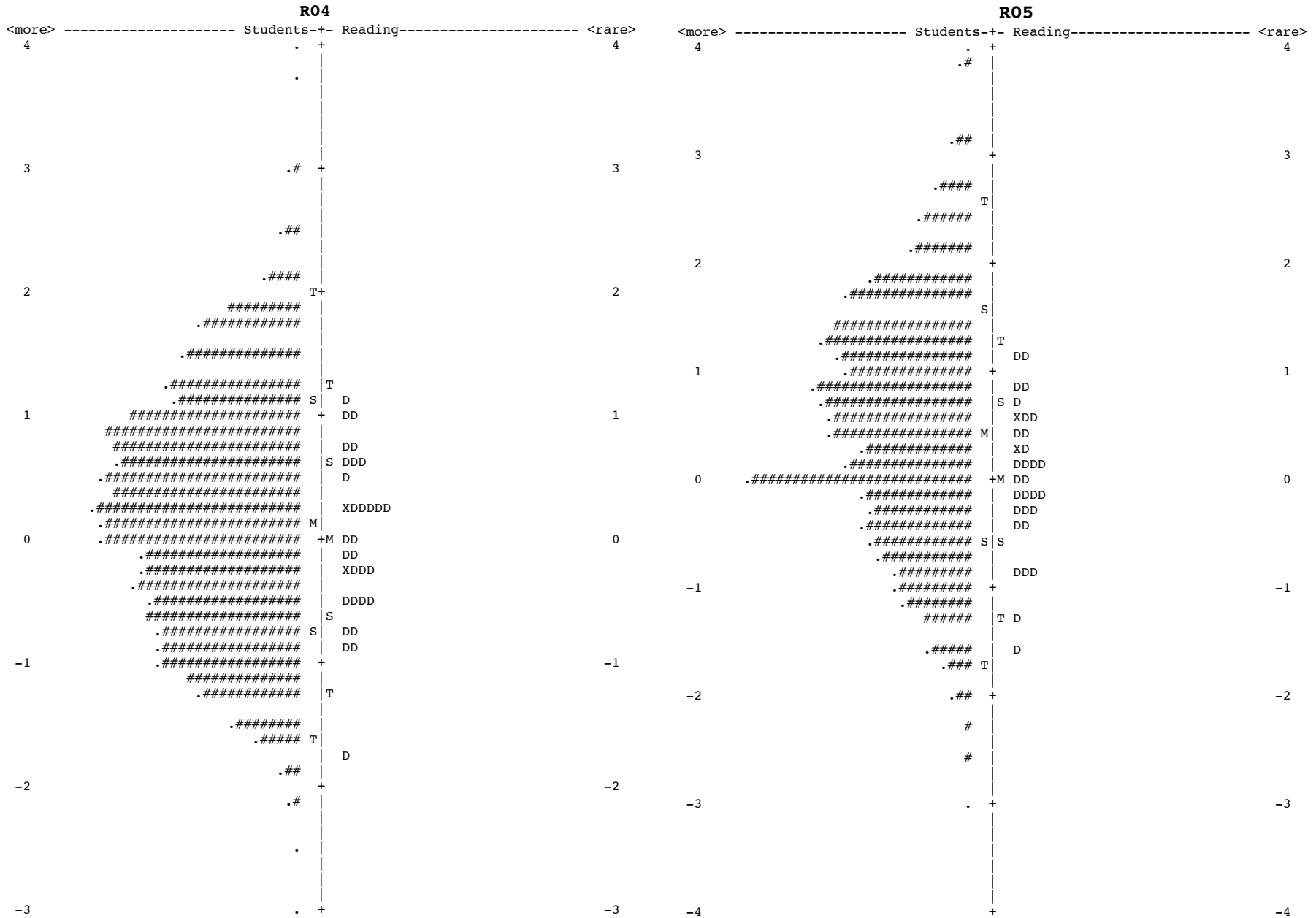
Chapter Twelve: Rasch Item Calibration

M06				M07			
<more>	Students	Math-Ms	<rare>	<more>	Students	Math-Ms	<rare>
5	.		5	4	.		4
4	.		4	3	.#		3
3	#		3	2	.##		2
2	.##		2	1	.###		1
1	.###		1	0	.####		0
0	#####		0	-1	#####		-1
-1	#####		-1	-2	#####		-2
-2	#####		-2	-3	#####		-3

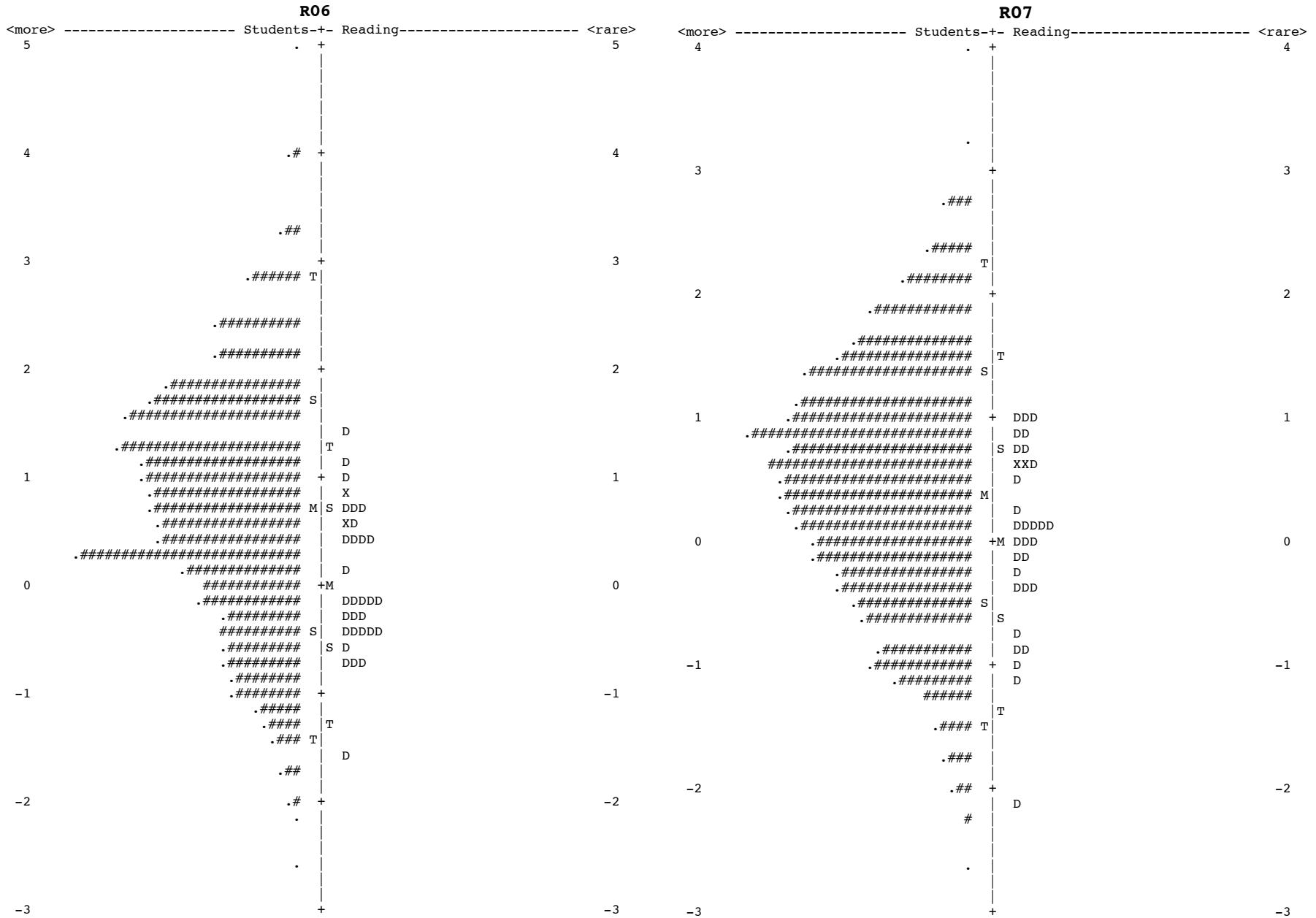
Chapter Twelve: Rasch Item Calibration



Chapter Twelve: Rasch Item Calibration



Chapter Twelve: Rasch Item Calibration



Chapter Twelve: Rasch Item Calibration

R08				R11			
<more>	Students	Reading	<rare>	<more>	Students	Reading	<rare>
4	.	+	4	4	.	+	4
	.				.		
3	.#	+	3	3	.#	+	3
	##				##		
	###	T			#####	T	
2	#####	+	2	2	#####	+	2
	#####				#####		
	#####	S			#####	S	
	#####	T			#####	T	
1	#####	+	1	1	#####	+	1
	#####				#####		
	#####	D			#####	D	
	#####	D			#####	D	
	#####	S			#####	S	
	#####	M			#####	M	
	#####	D			#####	D	
	#####	DDDD			#####	DDDD	
0	#####	+M	0	0	#####	+M	0
	#####				#####		
	#####	D			#####	D	
	#####	XD			#####	XD	
	#####	DD			#####	DD	
	#####	S			#####	S	
	#####	S			#####	S	
	#####	DD			#####	DD	
-1	#####	+	-1	-1	#####	+	-1
	#####				#####		
	#####	D			#####	D	
	#####	T			#####	T	
	##	T			#####	T	
	##				#####		
	.				#####		
-2	.	+	-2	-2	#####	+	-2
	.				#####		
	.				#####		
	.				#####		
-3	.	+	-3	-2	.	+	-2
	.				.		
	.				.		
-4	.	+	-4	-3	.	+	-3
	.				.		
	.				.		

Chapter Twelve: Rasch Item Calibration

S08				S11			
<more>	Students	Science	<rare>	<more>	Students	Science	<rare>
4	.	+	4	4	.	+	4
	##				.		
3	.###	+	3	3	.	+	3
	.#####				##		
	.#####	T			.###		
2	#####	+	2	2	.#####	T+	2
	.#####				.#####		
	.#####	S T			.#####		
	.#####				#####		
	#####	D			#####	T D	
	.#####	D			.#####		
1	.#####	+	1	1	.#####	+	1
	.#####	S			.#####	XD	
	#####	DDD			.#####	DD	
	.#####	D			.#####	S D	
	.#####	M			.#####		
	.#####	XDD			.#####	D	
	.#####	XDD			.#####	M XDDDD	
	.#####	D			.#####	DDD	
	.#####	DD			.#####	DD	
0	.#####	+M	0	0	.#####	+M	0
	.#####	DD			.#####	D	
	.#####	DDD			.#####	DDDD	
	.#####				.#####	D	
	#####	DD			.#####	S S DDD	
	.#####	DD			.#####		
	.#####	S			.#####	D	
	.#####				.#####	+ DD	-1
-1	####	+	-1	-1	.#####	T	
	##	T			.#####	D	
	.##				.#####	T	
	.##	T D			.#####		
	.#	D			####		
-2	.	+	-2	-2	.##	+	-2
	.	D			.#		
-3		+	-3	-3	.	+	-3

Chapter Thirteen: Performance Level Setting

The performance level setting for the PSSA-M reading and science tests was conducted by Data Recognition Corporation (DRC) using the Ordered-Item Booklet (OIB) Angoff (Yes/No) method during a workshop held in Harrisburg, Pennsylvania, on May 17–20, 2011. 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 2011 Pennsylvania System of School Assessment-Modified Reading and Science Assessments (Data Recognition, 2011)*⁷

A history (dates and methodology) of prior performance level setting events are provided in Table 13–1. For additional details about any given event, refer to the technical report for the year that the event occurred (Data Recognition Corporation, 2010, 2011).

Table 13–1. Performance Level Setting/Validation Event Dates and Methodology

Subject	Grade	Methodology	Event Date
Mathematics-M	4–8, 11	OIB Angoff	Spring 2010
Reading-M	4–8 11	OIB Angoff	Spring 2011
Science-M	8, 11	OIB Angoff	Spring 2011

SUMMARY

Figure 13–1 presents general information about the performance levels setting event. The purpose of the event was to establish three cut scores for the modified reading assessment at Grades 4–8 and 11 and for the modified science assessment at Grades 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 panelists from across the state and targeted educators who had experience in reading or science at a specific grade level and were knowledgeable about modified content standards. A total of 33 educators (22 general educators and 11 special educators) attended the reading performance level setting, and a total of 21 educators (14 general educators and 7 special educators) attended the Science performance level setting. For reading, the 33 panelists were assigned to one or two grade-level panels based on their experience. For science, the 21 panelists were assigned to either the Grade 8 or Grade 11 panel. The numbers of panelists in each grade panel ranged from 8 to 13. The ratio of special educators to general educators ranged from 1:3 to 5:7 across the panels.

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

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

Official Title	Performance Levels Setting for PSSA-M Reading and Science		
Event Dates	May 17–20, 2011		
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 Reading & Modified Science	Grades	4, 5, 6, 7, 8, and 11 for Reading, 8 and 11 for Science
Panelists	Reading: 33 Total Science: 21 Total at least 8 per Group	Tables	1 Table per Group
Rounds	Three (3) plus Articulation	Impact Groups	Total Group

The entire PSSA-M performance level event included the following components:

1. A 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 six reading grades (Grades 4–8 and 11).

Table 13–2 summarizes the final recommended raw, theta, and scale score cuts and their associated conditional standard errors of measurement (CSEMs) for each grade. On May 26, 2011 the Pennsylvania State Board of Education approved the panelists’ recommendations at all grades.

Table 13–2. Final Cut Scores and Conditional Standard Errors of Measurement (CSEM) for Scale Score Cuts

	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	
Reading	4	8	-1.2893	1150	43	17	-0.1101	1275	36	24	0.7217	1363	38
	5	10	-1.0125	1150	41	19	0.1898	1275	37	27	1.3018	1391	42
	6	12	-0.6956	1150	36	22	0.6103	1275	35	29	1.7201	1381	43
	7	12	-0.6788	1150	40	21	0.4809	1275	39	28	1.5026	1385	45
	8	14	-0.4859	1150	40	23	0.6508	1275	40	30	1.7754	1399	51
	11	17	-0.0445	1150	42	25	1.0155	1275	45	32	2.3512	1433	63
Sci.	8	14	-0.3288	1150	43	22	0.7473	1275	44	29	1.9833	1419	58
	11	9	-1.0555	1150	37	19	0.331	1275	33	28	1.7325	1401	42

PSSA-M CUT SCORES

Appendix L provides the Rasch ability and scaled score cuts for each PSSA-M test. For reader convenience, these are documented in the next table in a different format. Table 13–3 documents the Rasch ability (Theta) cut scores for each grade and subject area test. Table 13–4 documents the same but provides the cut scores on the scaled-score metric. PSSA-M scaling procedures are discussed further in Chapter Fourteen.

Table 13–3. PSSA-M Theta (θ) Metric Cut Scores by Grade and Subject Area

		θ Cuts		
		BB/B	B/P	P/A
Mathematics	4	-0.5891	0.8935	1.8540
	5	-0.5352	0.8640	1.9734
	6	-0.9606	0.3441	1.4543
	7	-0.8442	0.5878	1.6086
	8	-0.7858	0.5407	1.8139
	11	-0.5492	0.5317	1.6389
Reading	4	-1.2893	-0.1101	0.7217
	5	-1.0125	0.1898	1.3018
	6	-0.6956	0.6103	1.7201
	7	-0.6788	0.4809	1.5026
	8	-0.4859	0.6508	1.7754
	11	-0.0445	1.0155	2.3512
Sci.	8	-0.3288	0.7473	1.9833
	11	-1.0555	0.3310	1.7325

Note. BB = Below Basic; B = Basic; P = Proficient; and A = Advanced.

Table 13–4. PSSA-M Scaled-Score Metric Cut Scores by Grade and Subject Area

		Scaled-Score Cuts		
		BB/B	B/P	P/A
Mathematics	4	1150	1275	1356
	5	1150	1275	1374
	6	1150	1275	1381
	7	1150	1275	1364
	8	1150	1275	1395
	11	1150	1275	1403
Reading	4	1150	1275	1363
	5	1150	1275	1391
	6	1150	1275	1381
	7	1150	1275	1385
	8	1150	1275	1399
	11	1150	1275	1433
Sci.	8	1150	1275	1419
	11	1150	1275	1401

Note. BB = Below Basic; B = Basic; P = Proficient; and A = Advanced.

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 scaled 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 produce 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 scaled score (LOSS) of 1075. For PSSA-M reading, the LOSS values have been set to 1075 at Grades 4-8 and 1050 and 1000 for Grades 8 and 11, respectively. For PSSA-M science, the LOSS values have been set to 1050 at Grade 8 and 1000 at Grade 11. LOSS values are documented in Table 14–2. See tables in Appendix N for LOSS n -counts.

Highest Obtainable Scaled Scores

A highest obtainable scaled 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 N for HOSS n -counts.

RAW-SCORE TO SCALED-SCORE TABLES

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

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

		θ Cuts			Intercept	Slope
		BB/B	B/P	P/A		
Mathematics	4	-0.5891	0.8935	1.854	1199.67	84.31
	5	-0.5352	0.864	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
Reading	4	-1.2893	-0.1101	0.7217	1286.67	106.00
	5	-1.0125	0.1898	1.3018	1255.27	103.97
	6	-0.6956	0.6103	1.7201	1216.58	95.72
	7	-0.6788	0.4809	1.5026	1223.17	107.79
	8	-0.4859	0.6508	1.7754	1203.43	109.97
	11	-0.0445	1.0155	2.3512	1155.25	117.92
Sci.	8	-0.3288	0.7473	1.9833	1188.19	116.16
	11	-1.0555	0.3310	1.7325	1245.16	90.16

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

		Scaled Score Cuts ¹				Max ²
		Min	BB/B	B/P	P/A	
Mathematics	4	1075	1150	1275	1356	1666
	5	1075	1150	1275	1374	1691
	6	1075	1150	1275	1381	1770
	7	1075	1150	1275	1364	1662
	8	1075	1150	1275	1395	1722
	11	1075	1150	1275	1403	2041
Reading	4	1075	1150	1275	1363	1808
	5	1075	1150	1275	1391	1791
	6	1075	1150	1275	1381	1724
	7	1075	1150	1275	1385	1788
	8	1050	1150	1275	1399	1756
	11	1000	1150	1275	1433	1753
Sci.	8	1050	1150	1275	1419	1769
	11	1000	1150	1275	1401	1694

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.

STRAND (REPORTING CATEGORY) SCORE STRENGTH PROFILE

The following process was followed to derive the strand (reporting category) score strength profile:

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

The strand scaled 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 strand scaled score was converted to H in cases where no strand scaled score equaled or exceeded the Advanced-M scaled score cut. See Chapter Sixteen for more information on strand scores and how they are used in score reports.

Chapter Fifteen: Linking

FOREWARD

This was the second operational administration of the PSSA-M mathematics test and the first operational administration of the PSSA-M reading and science tests. As such, linking/equating was required only for mathematics.

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. 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 was utilized for the PSSA-M operational scores in mathematics. Scores from the new test form were linked to the scale of the old test form. 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 multiple-choice (MC) and open-ended (OE) items, which is considered as within-year linking in this chapter.

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

2. Calibrate selected OE items in an anchored run by putting them on the MC item scale from Step 1:
 - Include all operational (OP) OE items.
 - Fix all MC items from Step 1.
3. Evaluate the stability of the linking items using Robust Z:
 - Include all operational (OP) linking (LK) items.
 - Calculate Robust Z for each item in the linking.

Once the above calculations are 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. The scatterplots with final LK item sets are shown in Figure 15–1.

4. Calculate the mean shift over MC linking items using item difficulties:
 - Include all operational (OP) linking (LK) items.
5. Apply the mean shift to the item parameters calibrated in Steps 1 and 2:
 - Include all operational (OP) MC and OE items.
6. Scale the operational test by fixing all operational 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 was based on the core linking (LK). These items had been used in previous administrations. The LK items were 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. This chapter focuses more on the linking between years.

The linking design employed for the 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., they contribute to student test scores).

All LK items were 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 are summarized in 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

		Core		Core Links	
		MC	OE	MC	OE
Mathematics	4	30	2	14	0
	5	30	2	14	0
	6	30	2	14	0
	7	30	2	12	0
	8	30	2	15	0
	11	30	2	15	0

LINKING METHOD FOR PSSA-M MATHEMATICS

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

Between-year linking utilized the current 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 current and banked parameters was then determined. The mean of the differences was then used to statistically adjust the new parameters to the bank scale. The final (linking) item parameters were then used to estimate student abilities, which were, in turn, transformed to scaled scores. (Transformation formulas are provided in Chapter Fourteen.)

RESULTS SUMMARY

Table 15–3 shows the initial and ending number of linking items and associated shift parameters and the correlation of item difficulties across years for each subject and grade. No LK items were dropped in any grade. At first glance, some of the mean shift values may appear large. However, the shift constants are being applied to parameter estimates from Step 1 in the equating process (where the mean of the unanchored MC items is fixed at zero). The adjustment needed to place the Step 1 estimates on the current scale can be large in magnitude as it must take into account multiple factors (e.g., changes in student ability since the base-year administration, and differences in difficulty).

Table 15–3. Summary Data for Linking Items

		Initial Counts		Final Counts		Initial Shift	Final Shift	Final Corr.
		MC	OE	MC	OE			
Mathematics	4	14	0	14	0	0.2869	0.2869	0.99
	5	14	0	14	0	0.3202	0.3202	0.99
	6	14	0	14	0	0.1985	0.1985	0.98
	7	12	0	12	0	0.1362	0.1362	0.97
	8	15	0	15	0	0.1299	0.1299	0.99
	11	15	0	15	0	0.2571	0.2571	0.97

Appendix J provides the statistics for the linking items used. The previous and current values for item sequence, p -values, and logits are also provided. Appendix M provides the mean raw and scaled score points across years. Together, these appendices provide a summary of how the items and test changed 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

This technical report uses figures to help one visualize the across-year differences in linking items for each subject 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 a 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.

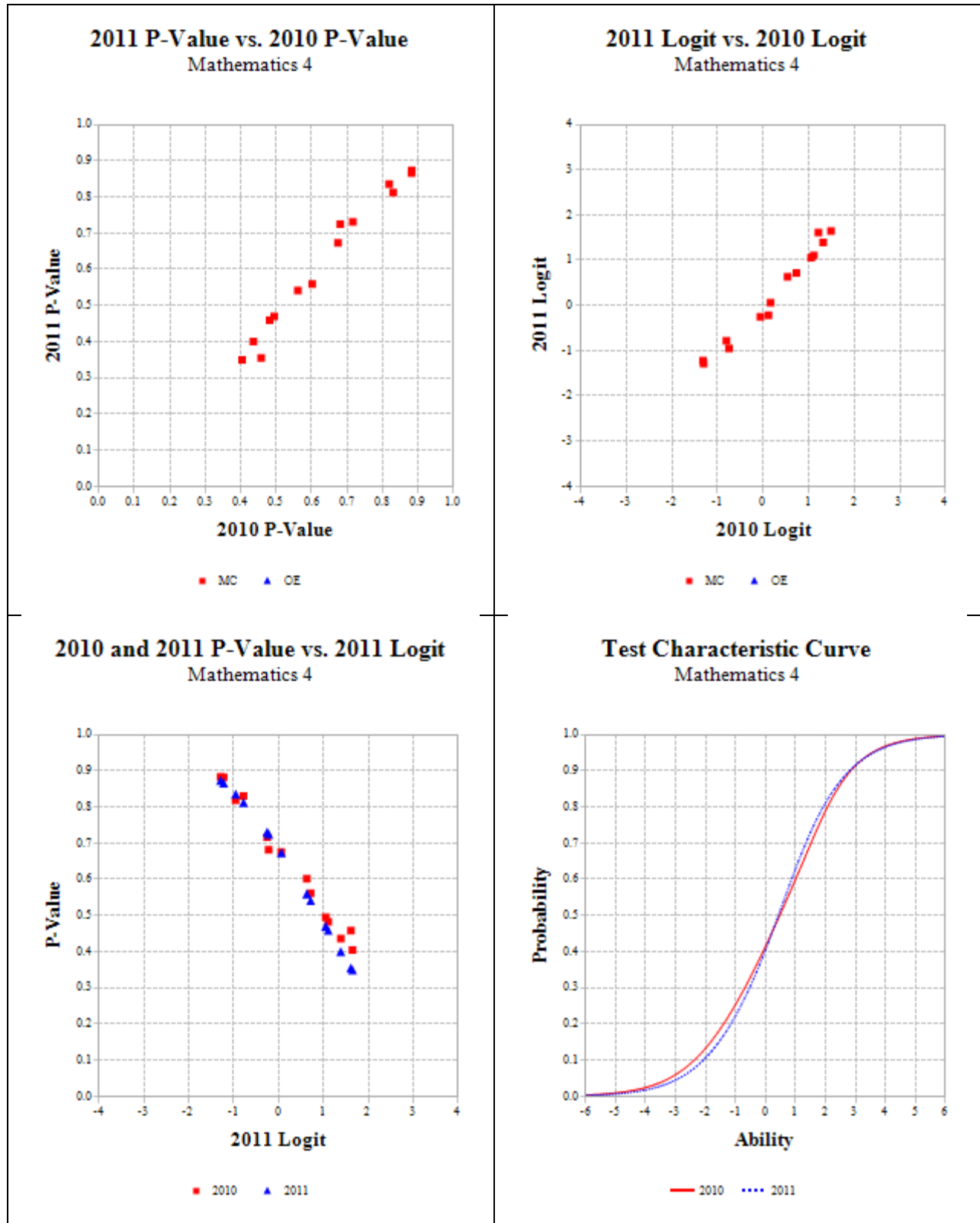
In Figure 15–1, the two lines sloping downward toward the right relate item *p*-values for the two years to the new-year logit difficulties. Again, these graphs have some similarity with the set of graphs that were part of Chapter Twelve. Both show the *p*-value-on-logit relationship, with the Chapter 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.

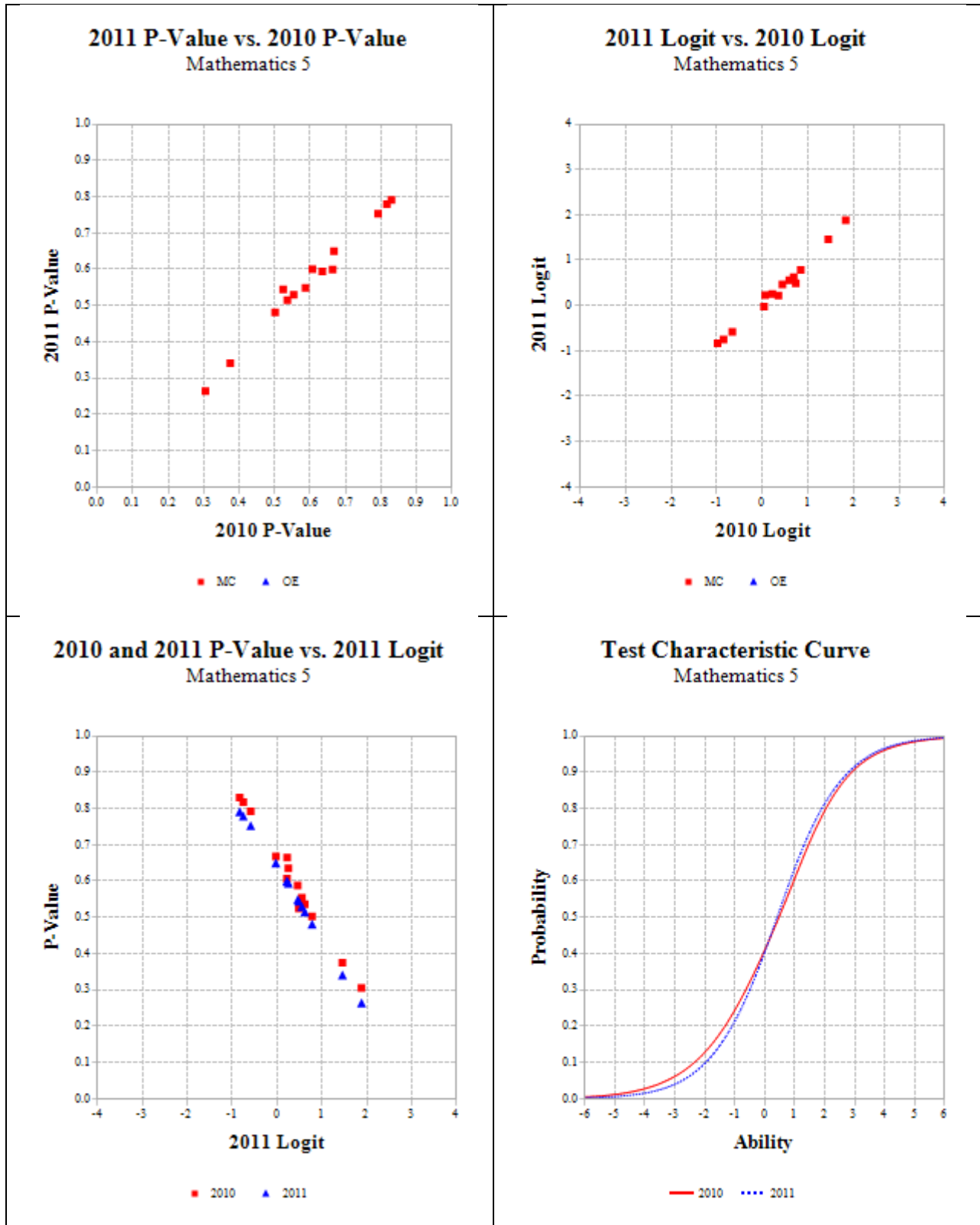
TEST CHARACTERISTIC CURVES

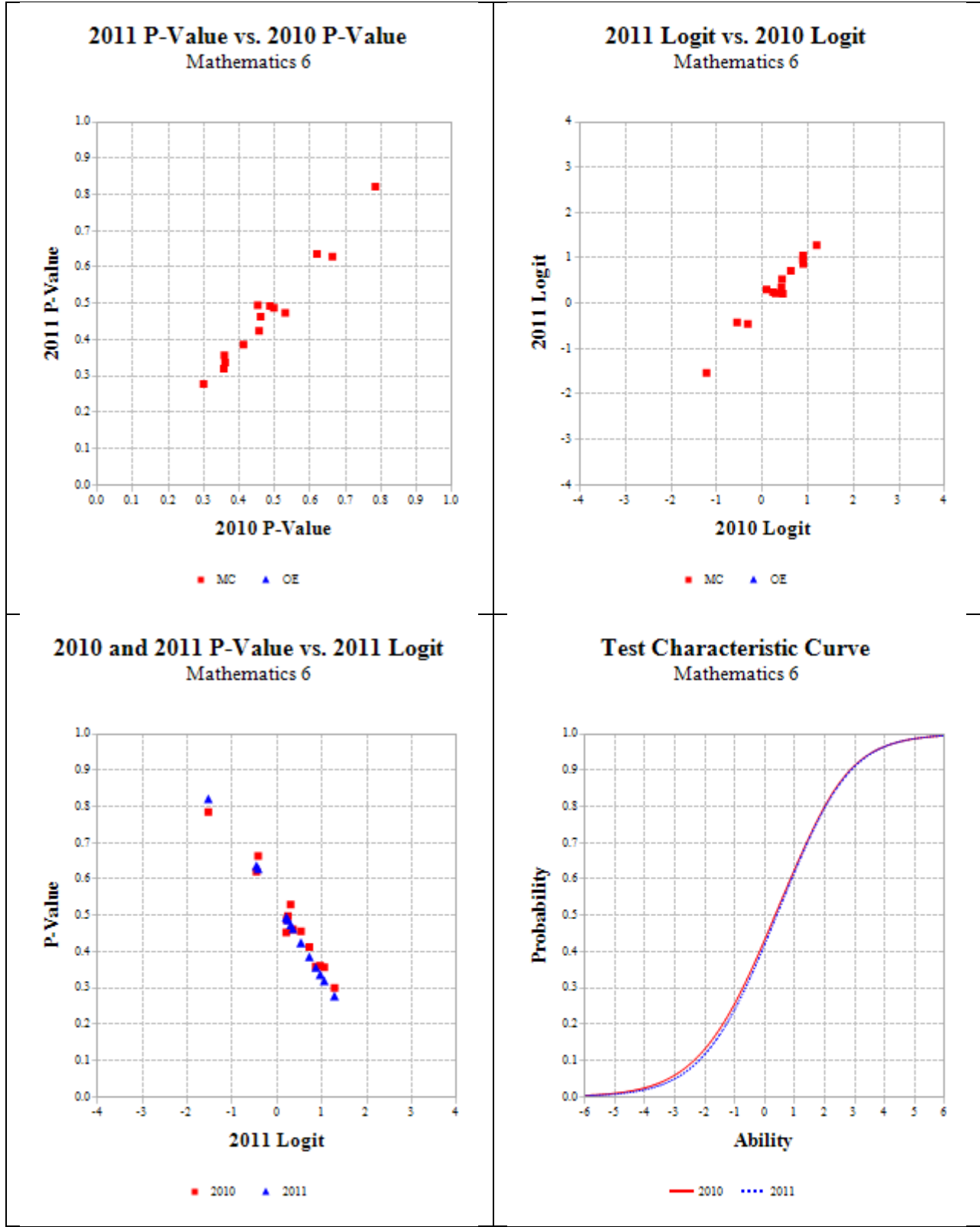
The old and new-year Test Characteristic Curves (TCCs) by grade and subject are shown in the figures⁸. TCCs show the similarity between the new- and old-year tests in terms of difficulty in the logit metric (new-year results are for the final, linked values). Assuming equal numbers of items for the two 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.

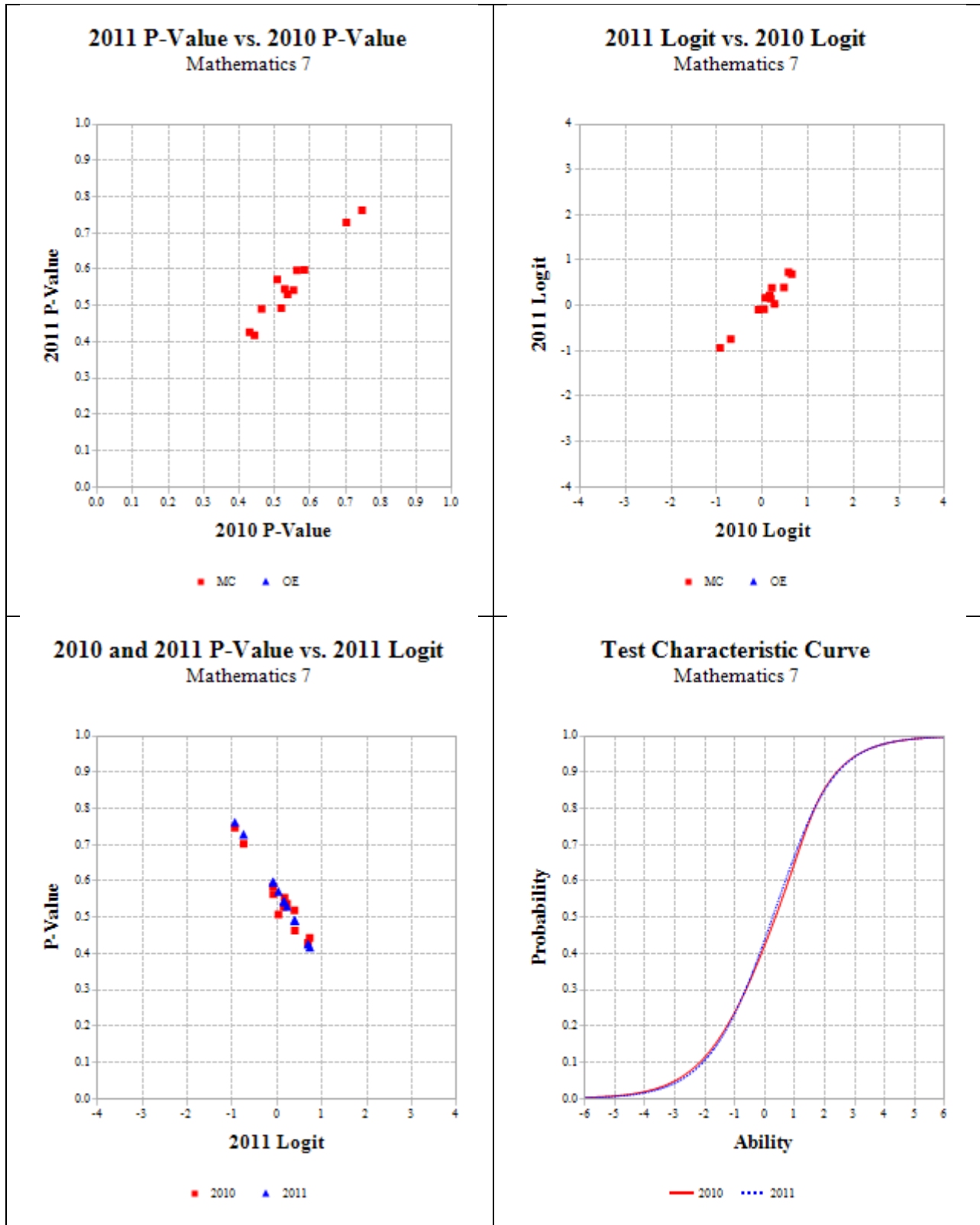
⁸ In the TCC figures, the Y-Axis Probability represents total test raw score expressed on a proportion-correct metric.

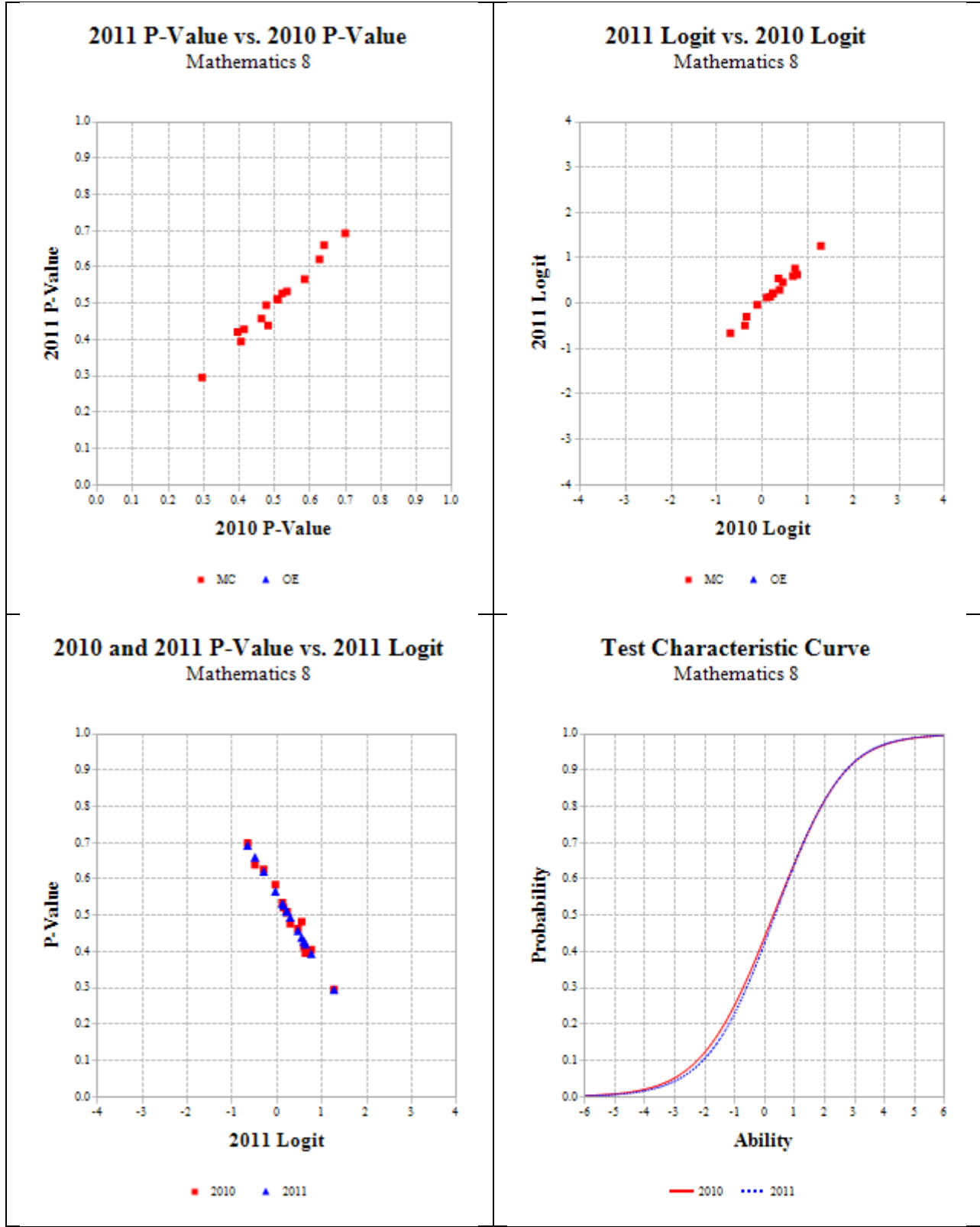
Figure 15–1. Item Stability Plots and Test Characteristic Curves

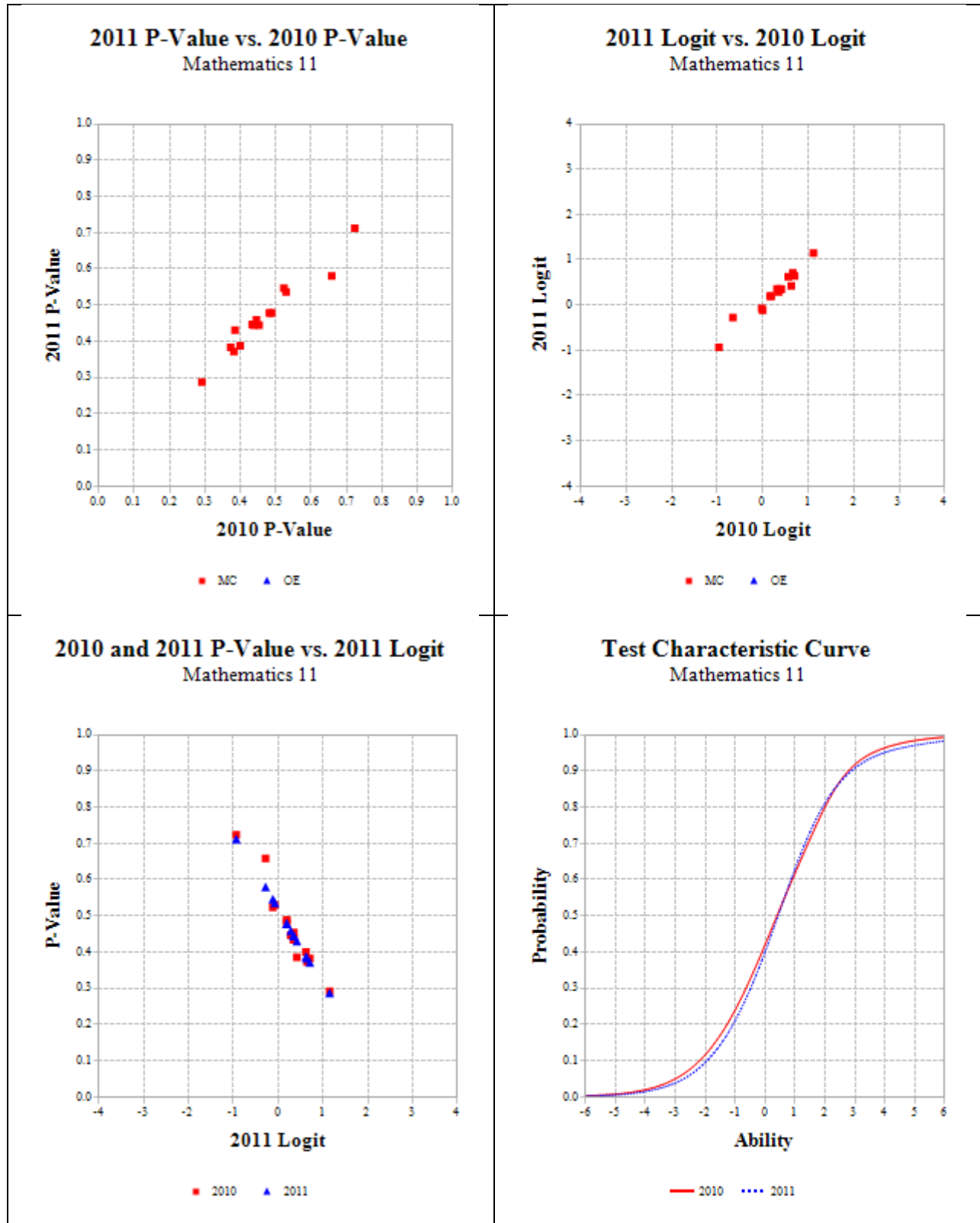












Chapter Sixteen: Scores and Score Reports

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 an 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

Item Type	Subject		
	Mathematics	Reading	Science
Multiple-Choice (1 point)	■	■	■
Open-Ended (2 point)			■
Open-Ended (3 point)		■	
Open-Ended (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 a student’s 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 is 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 of 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**

		Scaled Score Cuts ¹				Max ²
		Min	BB/B	B/P	P/A	
Mathematics	4	1075	1150	1275	1356	1666
	5	1075	1150	1275	1374	1691
	6	1075	1150	1275	1381	1770
	7	1075	1150	1275	1364	1662
	8	1075	1150	1275	1395	1722
	11	1075	1150	1275	1403	2041
Reading	4	1075	1150	1275	1363	1808
	5	1075	1150	1275	1391	1791
	6	1075	1150	1275	1381	1724
	7	1075	1150	1275	1385	1788
	8	1050	1150	1275	1399	1756
	11	1000	1150	1275	1433	1753
Sci.	8	1050	1150	1275	1419	1769
	11	1000	1150	1275	1401	1694

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 STRAND (REPORTING CATEGORY) SCORES

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

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

Strand (Reporting Category) Scores

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

Strand scores cannot be compared across years because they are not statistically linked. Also, it is not advisable to compare strand 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, strand 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 strand scores of a school against the scores of another reference group (e.g., the state average). Hence, strand 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 strand 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 Proficient-M on the PSSA-M. A High score on the strength profile indicates performance that is comparable to Advanced-M.

APPROPRIATE SCORE USES

Individual Students

Scaled scores on the PSSA-M indicate a student's achievement of 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 Item Response Theory (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 percentage of students beyond the Proficient-M standard.

Strand 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 achieved this score, it could not 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 do 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 are deliberately used instead 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 in 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 only on the variances of the strand scores, but also on the covariances among the strand scores. (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 to which 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 considered.

REPORTS

The 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 June 10, 2011. This score report provided parents and students with their first glimpse of performance on the spring 2011 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 Parents:

I am pleased to provide you with information about your child's performance on the 2011 Pennsylvania System of School Assessment (PSSA) exam. The annual PSSA is a standards-based assessment used to measure a student's attainment of the academic standards while also determining the degree to which school programs enable students to attain academic proficiency.

For additional information about the PSSA, visit the Pennsylvania Department of Education's website at *Education.state.pa.us*, or contact your child's school.

Sincerely,
Ronald J. Tomalis
Secretary of Education



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

MATHEMATICS		
How did perform OVERALL?		
Performance Level: Advanced	Score: 1526	
Below Basic	Basic	Proficient
700	1167	1304
		1509
		1594
		2425
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: 1462–1594.		
How did perform by REPORTING CATEGORY?		
Reporting Categories	Student's Points	Total Points Possible
Numbers and Operations	9	11
Measurement	11	11
Geometry	10	13
Algebraic Concepts	22	27
Data Analysis and Probability	7	10

WRITING		
How did perform OVERALL?		
Performance Level: Advanced	Score: 1918	
Below Basic	Basic	Proficient
700	952	1236
		1806
		2364
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: 1875–1961.		
How did perform by REPORTING CATEGORY?		
Reporting Categories	Student's Points	Total Points Possible
Composition	70	80
Informational	40	40
Persuasive	30	40
Revising and Editing	18	20
Informational	4	4
Persuasive	3	4
Multiple Choice	11	12

READING		
How did perform OVERALL?		
Performance Level: Proficient	Score: 1429	
Below Basic	Basic	Proficient
700	1112	1257
		1492
		2511
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: 1343–1515.		
How did perform by REPORTING CATEGORY?		
Reporting Categories	Student's Points	Total Points Possible
Comprehension and Reading Skills	17	22
Interpretation and Analysis of Fictional and Nonfictional Text	22	30

SCIENCE		
How did perform OVERALL?		
Performance Level: Proficient	Score: 1330	
Below Basic	Basic	Proficient
1050	1150	1275
		1347
		1822
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: 1302–1358.		
How did perform by REPORTING CATEGORY?		
Reporting Categories	Student's Points	Total Points Possible
The Nature of Science	27	38
Biological Sciences	7	12
Physical Sciences	10	14
Earth and Space Sciences	9	10

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.

Individual Student Report

An individual student report is provided for all students who took the PSSA-M. This report was delivered to Pennsylvania school districts on September 7, 2011. Districts are responsible for sending the reports 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

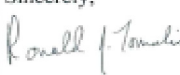
Student Report

Dear Parents:

The following report is designed to provide you with specific information about your child's strengths and needs as measured by the 2011 Grade 11 Pennsylvania System of School Assessment (PSSA). The PSSA is an annual exam designed to measure a student's attainment of academic standards. I encourage you to use this information to talk with your child's teacher(s) to develop methods to enhance your student's education.

For additional information about the PSSA, visit the Pennsylvania Department of Education's website at www.education.state.pa.us, or contact your child's school.

Sincerely,



Ronald J. Tomalis
Secretary of Education

Student Name: [Redacted]

PA Student ID: [Redacted]

School: [Redacted]

District: [Redacted]

Test Date: Spring 2011

Grade: 11

Student's PSSA Results by Subject

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

Table of Contents


Page 1..... General Overview

Page 2..... Math, Reading, and Science Detailed Results

Page 3..... Writing Detailed Results

Page 4..... Making the Most of Your Senior Year!

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

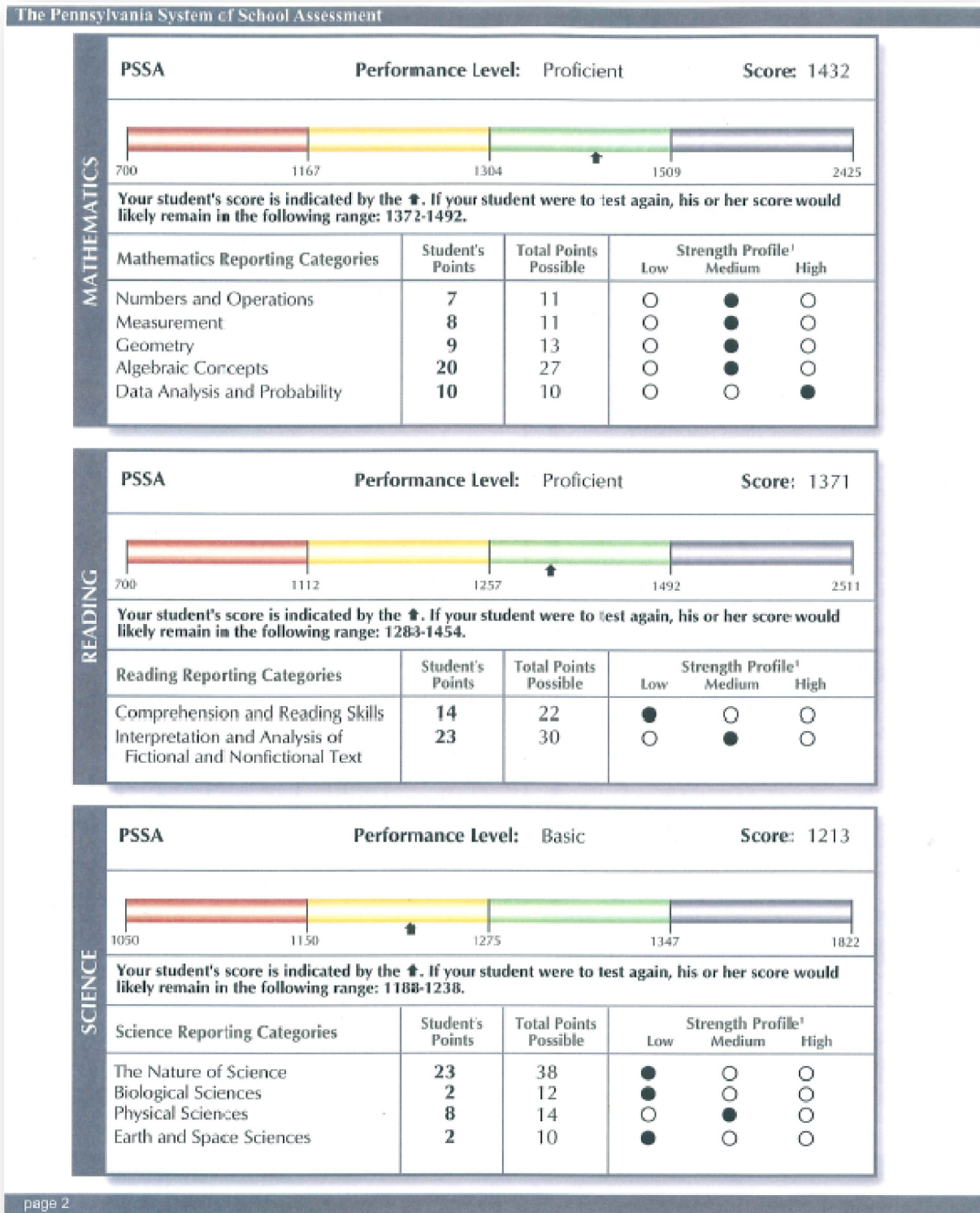


pennsylvania
DEPARTMENT OF EDUCATION

The Pennsylvania System of School Assessment page 1

www.education.state.pa.us

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



page 2

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

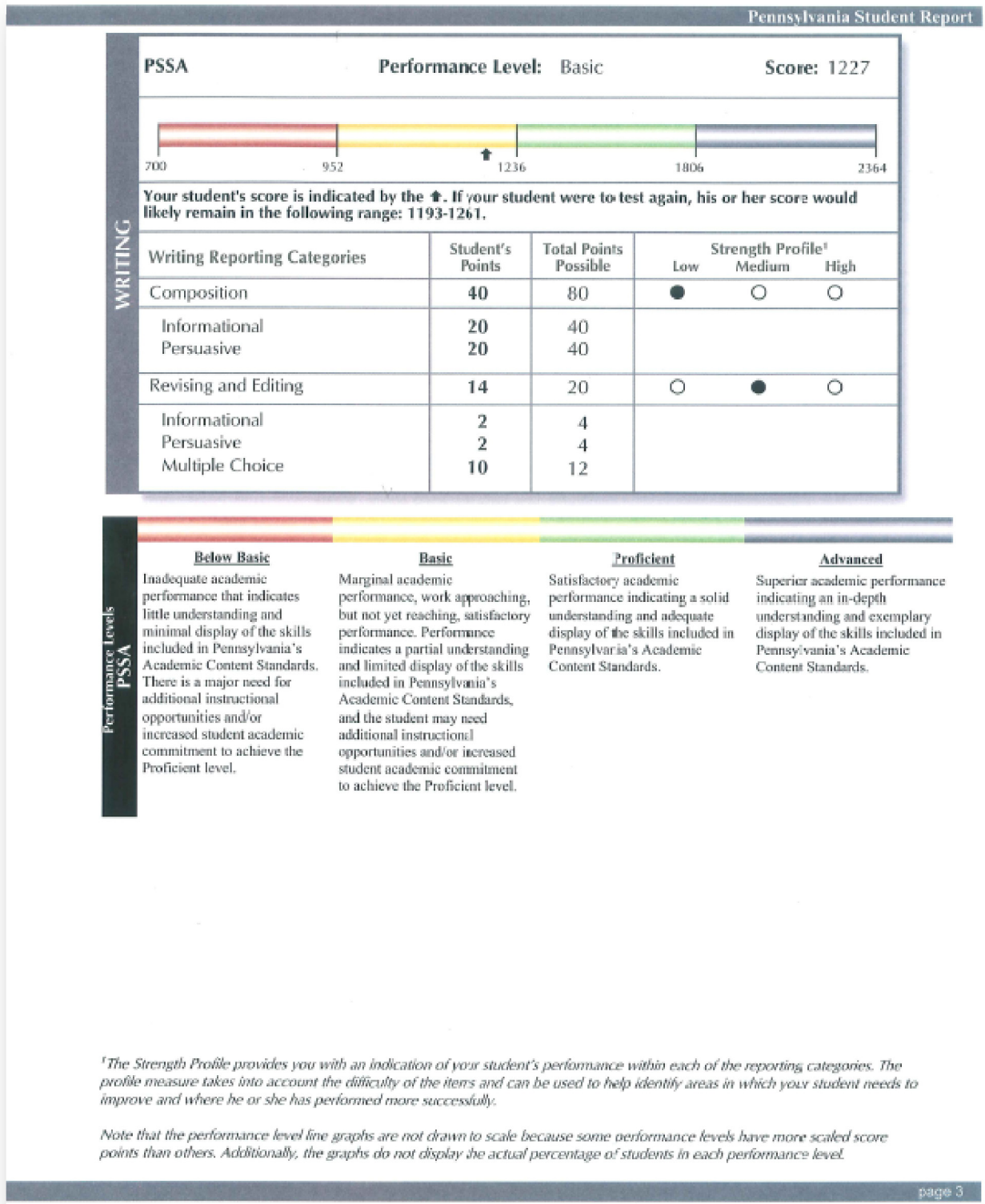


Figure 16–5. Page 4 of the Individual Student Report

Making the Most of Your Senior Year!

Education pays...
Education pays in higher earnings and lower unemployment rates.

Unemployment rate in 2010 (%)

Less than a high school diploma	14.9
High school diploma	10.3
Some college, no degree	9.2
Associate degree	7.0
Bachelor's degree	5.4
Master's degree	4.0
Professional degree	2.4
Doctoral degree	1.9

Average: 8.2%
Source: Bureau of Labor Statistics, Current Population Survey

Median weekly earnings in 2010 (\$)

Less than a high school diploma	444
High school diploma	626
Some college, no degree	712
Associate degree	767
Bachelor's degree	1,038
Master's degree	1,272
Professional degree	1,610
Doctoral degree	1,550


Average: \$782

Checklist for your future:
Everyone needs to find his/her career path in life. Your senior year should serve as your springboard to your future – make it count!

- Talk to your parents/guardians about your future plans – these could include college, career schools, apprenticeships, the military or other options.
- Make an appointment for you and your parents/guardians to discuss these plans with your guidance counselor.
- For any postsecondary interests you may have, seek out events related to those opportunities through college or career fairs.
- Take advantage of dual enrollment, which your school may already offer. Dual enrollment provides college courses that can give you college credit(s) during your senior year.
- Ask specific questions about the admission requirements, such as what courses you may need to take during your senior year and what college exams and other types of scores may be needed for entrance into any kind of school, training or service.
- KEEP GOING! High school isn't over yet!

Career planning

- Contact your school's guidance or career counselor for information on determining your career interests.
- Free printed materials and career exploration CDs from Pennsylvania Higher Education Assistance Agency (PHEAA) are available by visiting: www.educationplanner.org.
- Take interest surveys and explore valuable career information at www.pacareerzone.com.
- Search the latest job postings at <https://paworkstats.geosoline.com/>



page 4

The Pennsylvania System of School Assessment
www.education.state.pa.us

Data Reporting Tool, 2011 PSSA-1209-64321

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 presented in Chapter Eleven (classical item statistics) and Chapter Twelve (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. Appendix M provides performance level percentages for prior years.

Table 17–1. Performance Level Percentages for the 2011 PSSA-M

Subject	Grade	Percentage in Each Performance Level			
		Below Basic	Basic	Proficient	Advanced
Mathematics	4	5.8	40.5	36.2	17.5
Reading		3.8	32.3	33.7	30.2
Mathematics	5	10.4	44.4	35.7	9.5
Reading		7.9	30.7	37.1	24.3
Mathematics	6	11.4	45.8	33.2	9.7
Reading		10.1	33.6	37.8	18.4
Mathematics	7	5.5	53.5	32.5	8.5
Reading		11.7	37.7	36.9	13.7
Mathematics	8	13.0	48.7	31.7	6.7
Reading		13.9	40.0	36.3	9.8
Science		12.0	40.1	38.2	9.7
Mathematics	11	24.4	43.1	25.4	7.0
Reading		17.4	37.1	40.2	5.3
Science		5.8	47.7	40.7	5.8

SCALED SCORES

Summary Statistics

Table 17–2 provides the scaled score means and standard deviations. (See the section Every Test Has a Unique Scale in Chapter Sixteen for caveats regarding interpretation of scaled scores.)

Table 17–2. Means and Standard Deviations for the 2011 PSSA-M Scaled Scores

Grade	Mathematics		Reading		Science	
	Mean	SD	Mean	SD	Mean	SD
4	1277.8	83.9	1305.6	98.2		
5	1261.2	85.5	1301.8	110.8		
6	1260.6	89.5	1281.9	101.5		
7	1256.0	74.5	1268.3	99.3		
8	1250.8	87.0	1257.4	101.5	1266.6	107.4
11	1227.8	107.2	1249.1	110.4	1263.5	79.2

Scaled Score Distributions

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

RAW SCORES

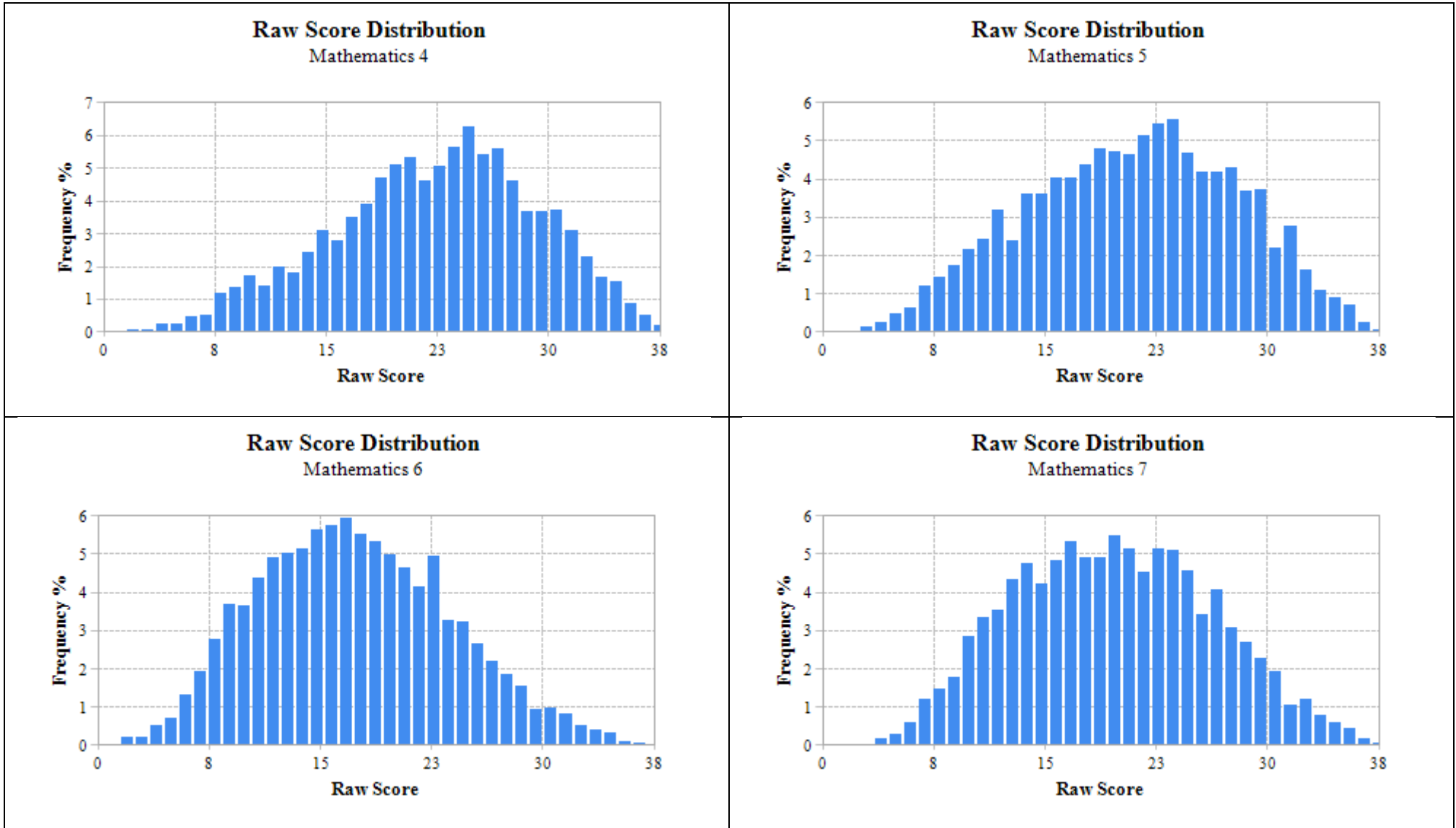
Summary Statistics

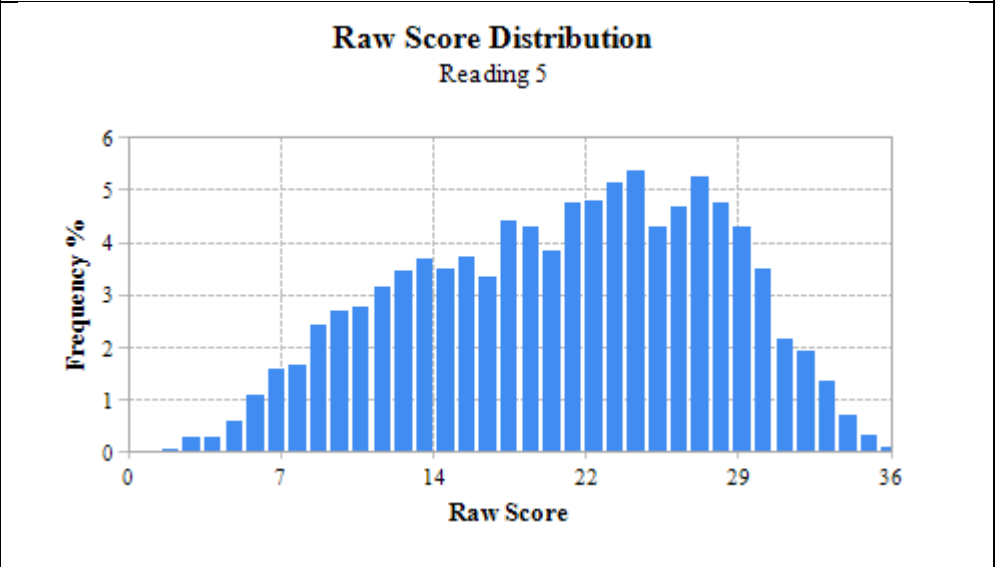
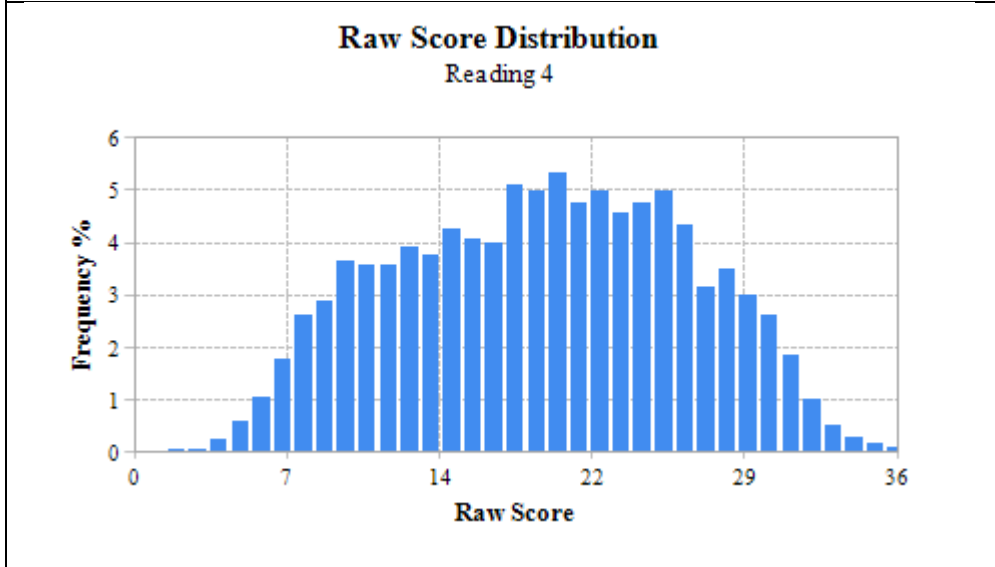
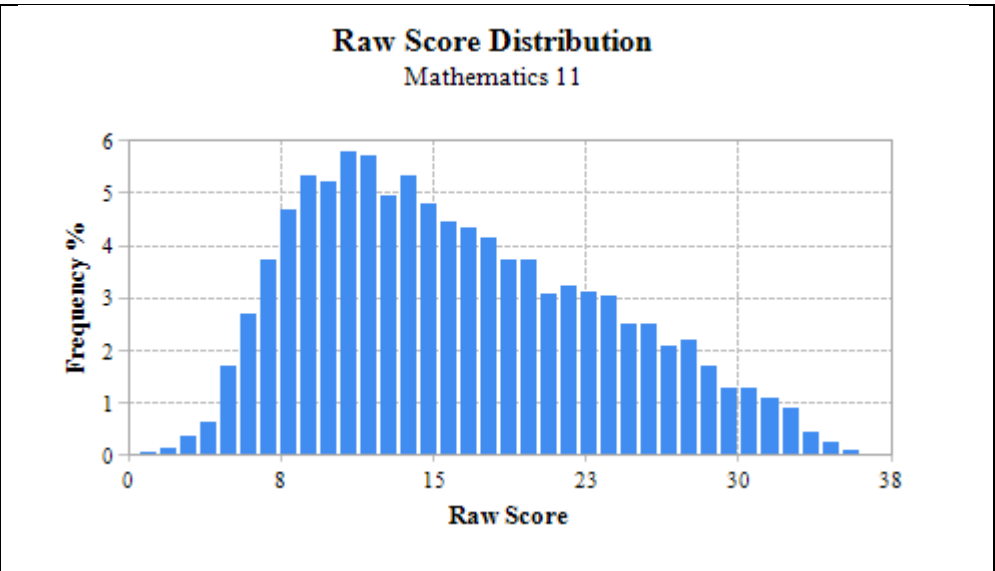
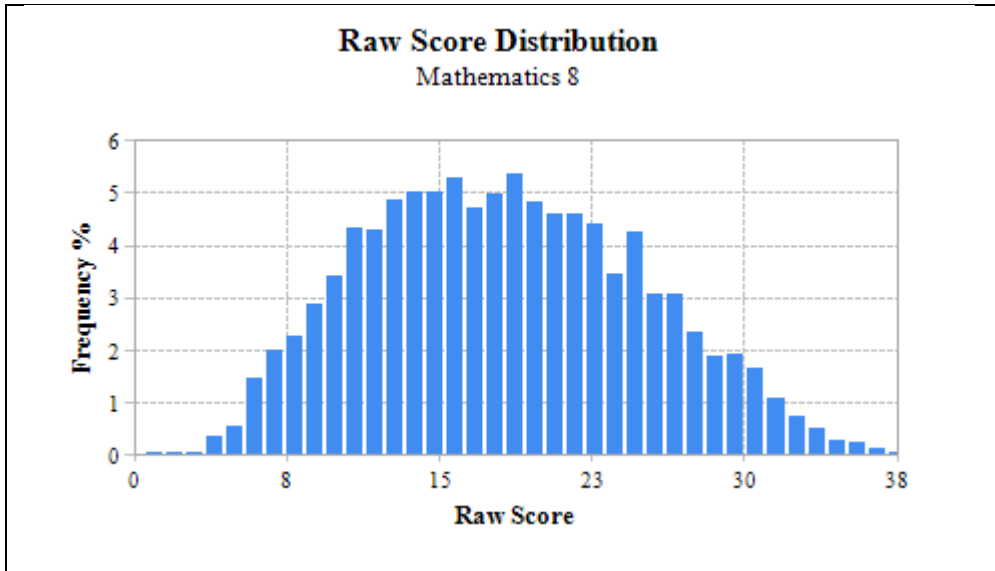
Appendix H provides summary statistics for the operational raw scores. The statistics reported include the number of students tested (N), number of items (L), number of points possible (Pts), minimum score points received (Min), maximum points received (Max), mean score points received (Mean), median score points received (Med), and standard deviation of test scores (SD). 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 provided in Figure 17–1. Most distributions are unimodal and slightly positively skewed in mathematics and slightly negatively skewed in reading.

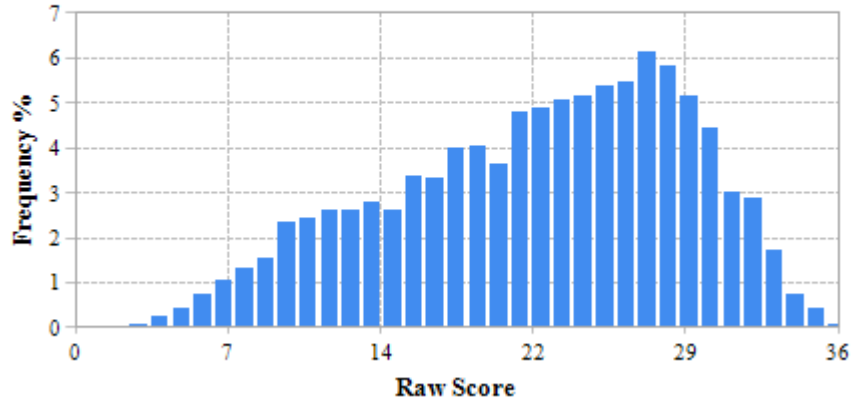
Figure 17–1. Raw Score Distributions





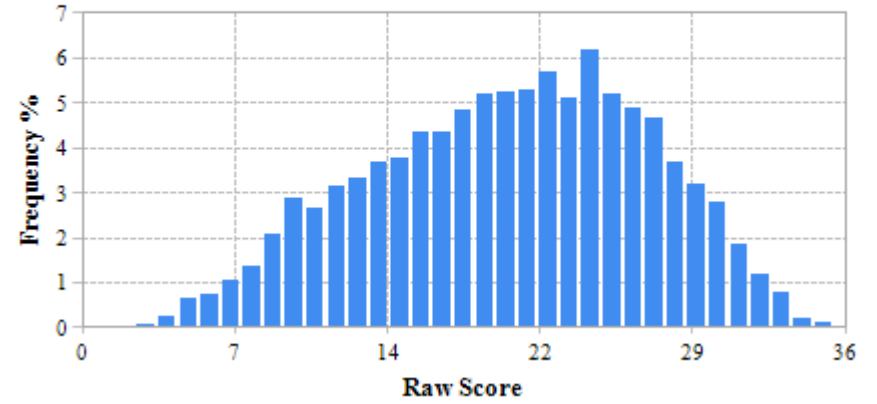
Raw Score Distribution

Reading 6



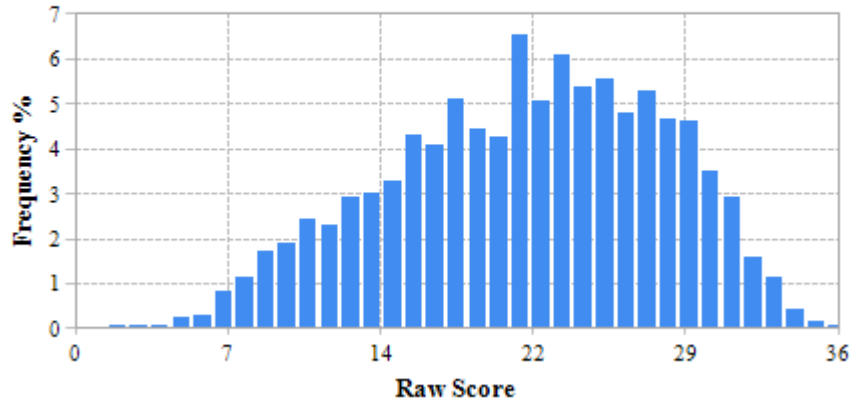
Raw Score Distribution

Reading 7



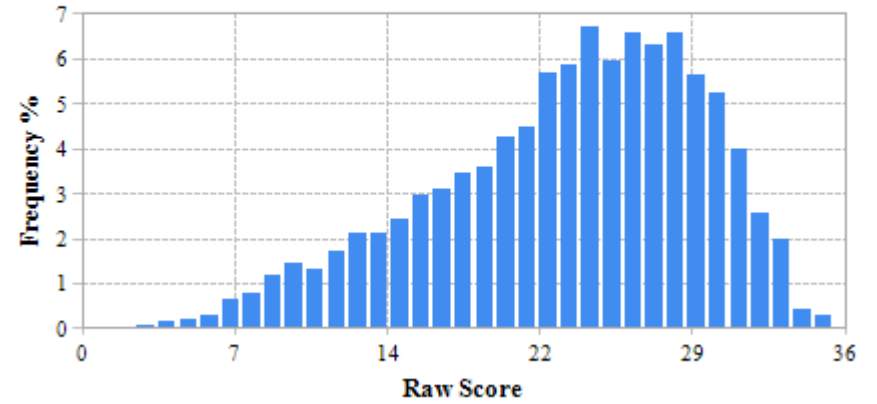
Raw Score Distribution

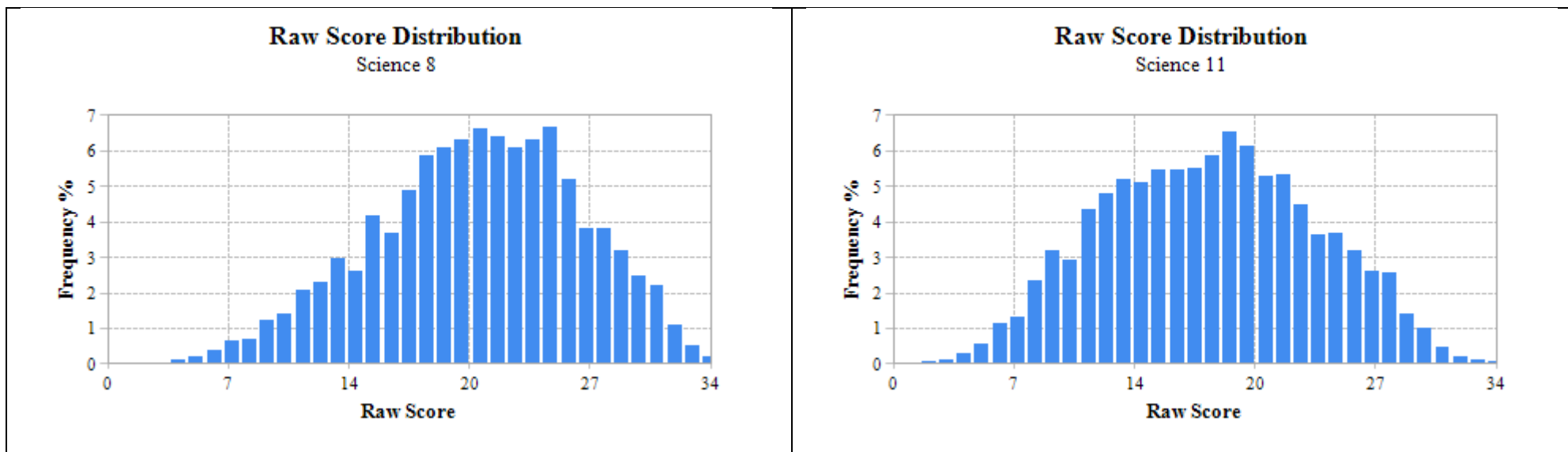
Reading 8



Raw Score Distribution

Reading 11





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 is 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 suggested above, reliability is a complex, nonunitary 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.¹⁰

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

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 one 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.

¹⁰ 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 could be 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 was 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 \times Item Score (X_{pi}) Infinite (Population-Universe) Matrix

Person	Item			
	1	2	... I	... k
1	Y_{11}	Y_{12}	... Y_{1i}	... X_{1k}
2	Y_{21}	Y_{22}	... Y_{2i}	... X_{2k}
.....				
.....				
P	Y_{p1}	Y_{p2}	... Y_{pi}	... X_{pk}
.....				
.....				
N	Y_{N1}	Y_{N2}	... Y_{Ni}	... X_{Nk}

Notes. 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 assessment 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 Coefficient 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., multiple-choice (MC) and open-ended (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, most of the total test score reliabilities reported in Appendix K 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

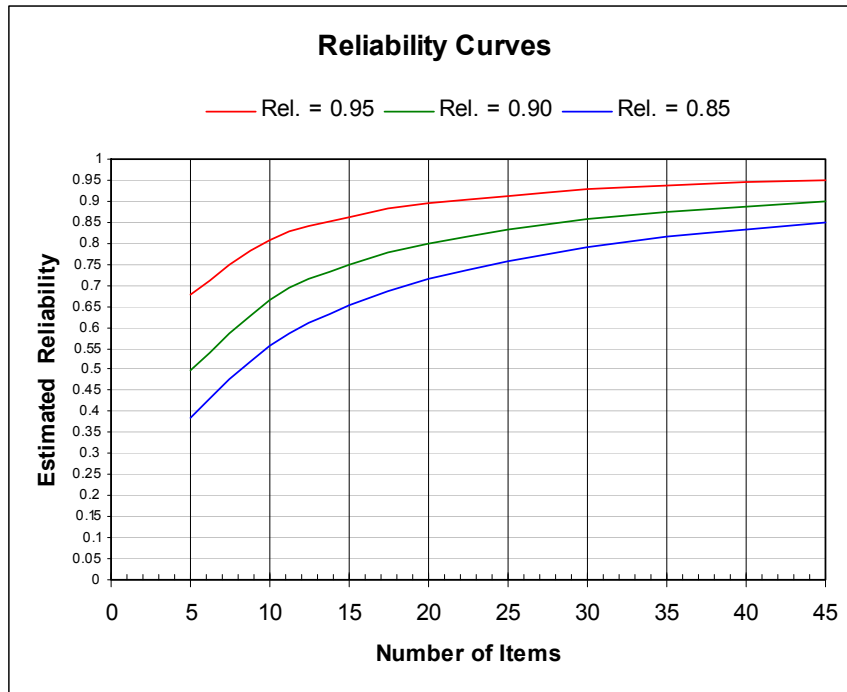
As emphasized in earlier sections, Coefficient Alpha only takes into account measurement error that arises from the selection of items used on a particular test form. There are other sources of random inaccuracy. One is due to the occasion of testing. Other various random conditions that might affect students on any particular testing occasions include illness, fatigue, and anxiety. Also, when a test includes OE items, as the PSSA-M does, another source that can cause random fluctuation is 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. Any internal consistency reliability index could understate the overall problem of measurement error because it ignores such sources of random error.

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 (Reporting Category) 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 is not specifically calculated here. However, as a general rule 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 a few items may not adequately cover the construct of interest. Validity is further discussed in Chapter Nineteen.

STANDARD ERROR OF MEASUREMENT

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 score¹¹, 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 deviation¹² 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.

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

This formula 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.

¹¹ True score is the score the person would receive if the measurement process were perfect.

¹² 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 test 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.¹³ 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. (In the next sections, an alternate approach is described that conditions the SEM on a student’s score estimate.)

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 is that a confidence interval represents the possible score range 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.

¹³ 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 K. 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), 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). Earlier it was noted that reliabilities tend to go up in value with an increase in test length¹⁴ 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 K, 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).¹⁵ 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 individual scores, suggesting that interpretation of strand scores at these aggregate levels might be reasonable in some cases.

¹⁴ 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 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.

¹⁵ 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.

RASCH CONDITIONAL STANDARD ERRORS OF MEASUREMENT

The conditional standard error of measurement (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, $CSEM(\hat{\theta})$ = 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 $CSEM(\hat{\theta})$ by the slope (multiplicative constant, m) of the linear transformation equation used to convert the Rasch ability estimates to scaled scores.

$$CSEM(SS) = CSEM(\hat{\theta}) * m$$

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

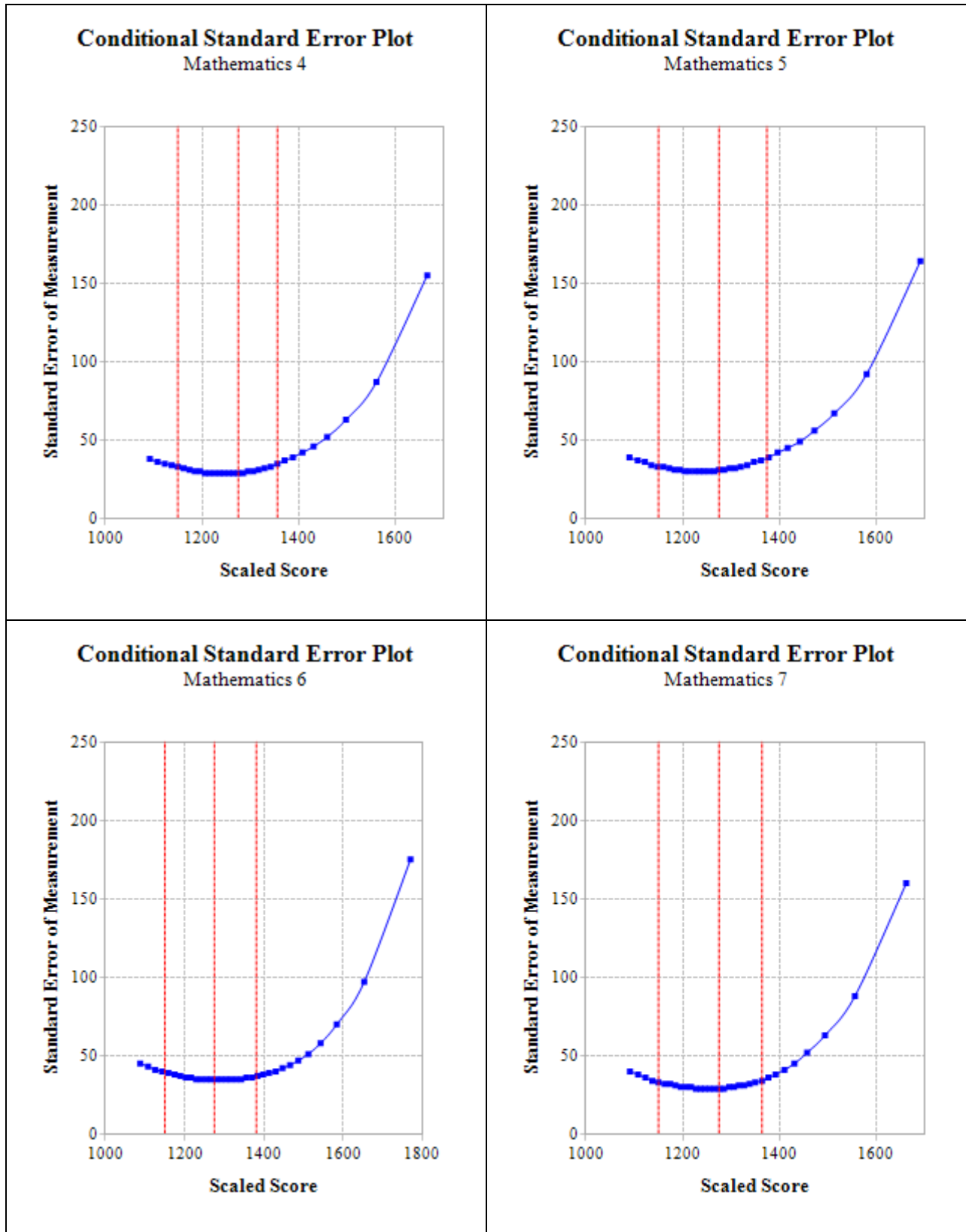
The SEMs documented on the PSSA-M score reports are the Rasch-based conditional standard errors of measurement described above. These are provided by the WINSTEPS scaling program described in Chapter Twelve. As noted earlier, these CSEMs are based on the concept of statistical information. For the purpose 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.¹⁶ Score report content is considered in greater detail in Chapter Sixteen.

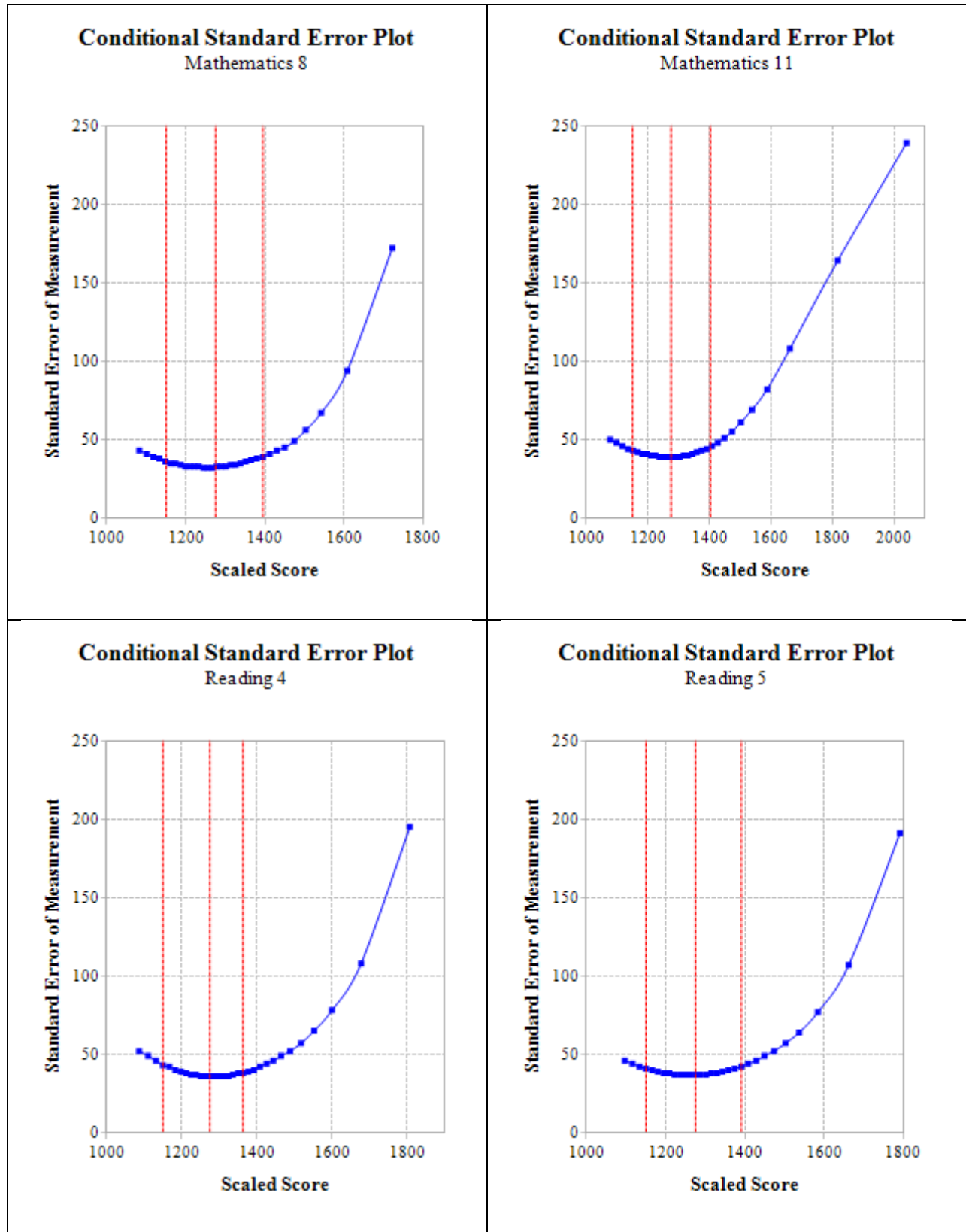
Results and Observations

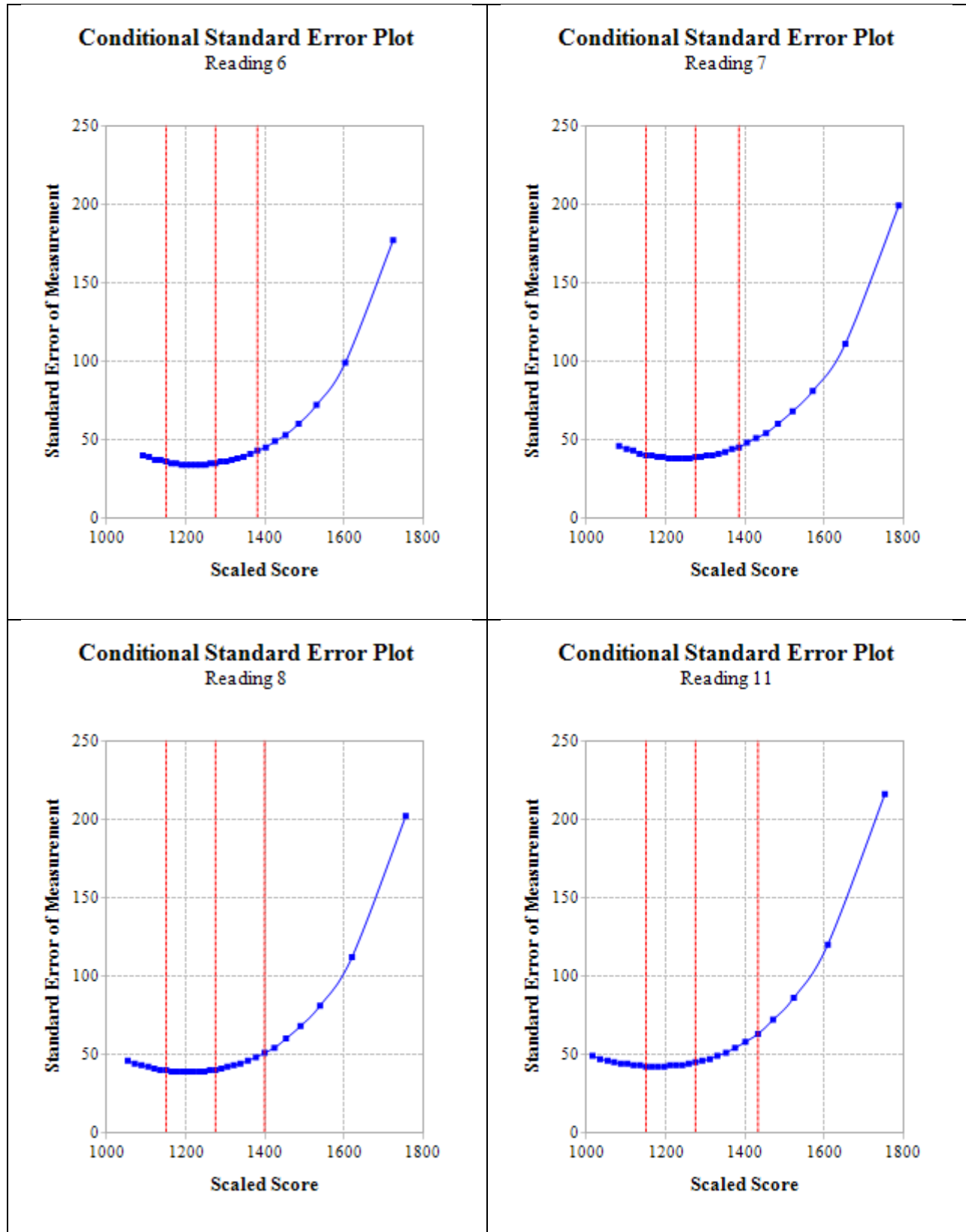
Figure 18–2 shows the Rasch CSEMs associated with each scaled score level. (This information is also provided in tabular form in Appendix N.) Values were derived using the final data file described in Chapter Nine. The values are fairly consistent across a noticeably large range of the scaled scores, 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.

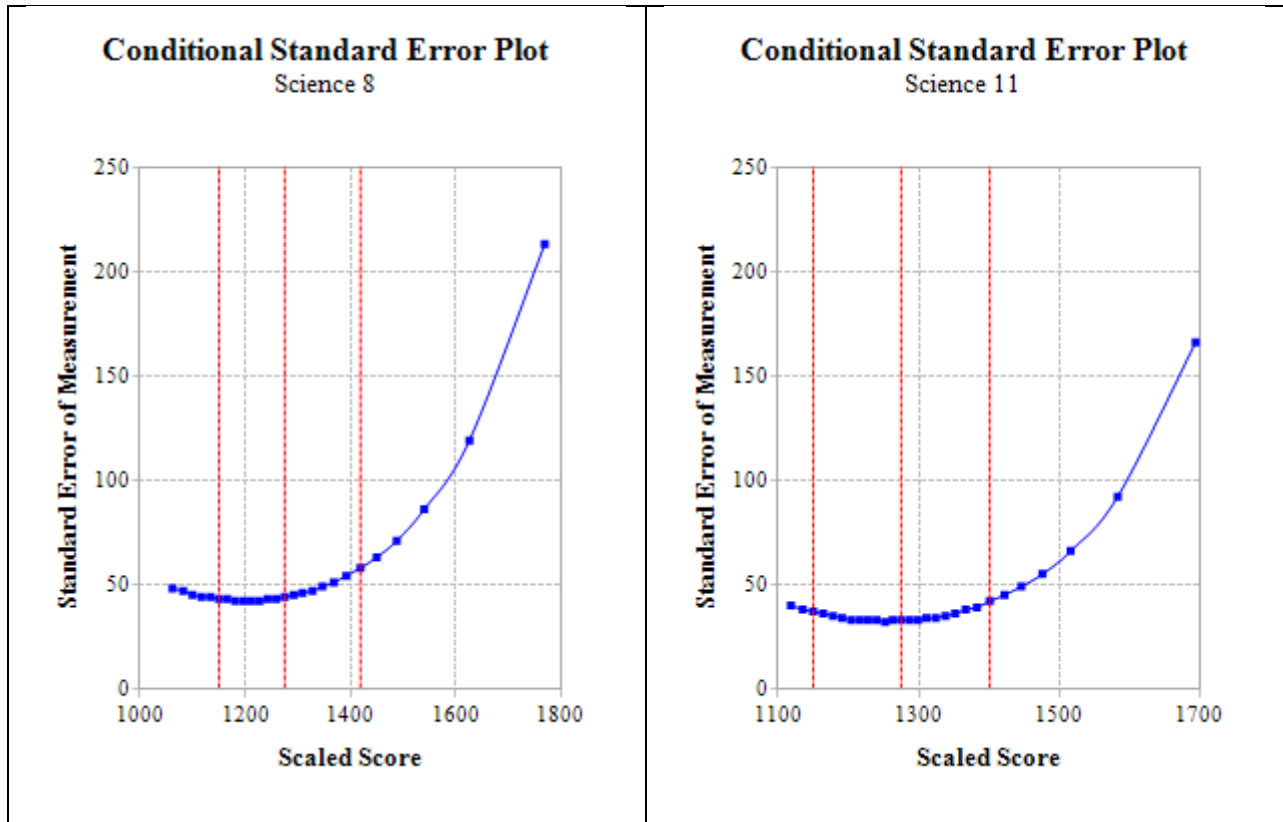
¹⁶ Because IRT CSEMs are based on statistical information, it is questionable if they account for error variance due to items. However, it seems difficult to construct a simple explanation of IRT CSEMs for the general public.

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









DECISION CONSISTENCY

Classification consistency refers to the degree to 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 signifies 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 decision consistency values typically ranged from the high 0.90s to mid .80s with the Basic cut generally yielding the highest values and the Advanced cut yielding the lowest values. Overall consistency values were generally in the low to mid 0.70s. Thus, for the PSSA-M, some individual cut consistencies were as high as the general 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
Mathematics	4	HB	0.65	0.95	0.82	0.87
		LL	0.64	0.95	0.81	0.87
	5	HB	0.66	0.92	0.83	0.91
		LL	0.65	0.91	0.82	0.91
	6	HB	0.65	0.89	0.82	0.92
		LL	0.65	0.89	0.82	0.92
	7	HB	0.68	0.93	0.82	0.93
		LL	0.67	0.92	0.82	0.93
	8	HB	0.66	0.88	0.83	0.94
		LL	0.66	0.88	0.84	0.94
	11	HB	0.64	0.82	0.87	0.94
		LL	0.64	0.82	0.87	0.94
Reading	4	HB	0.65	0.95	0.84	0.85
		LL	0.64	0.94	0.84	0.85
	5	HB	0.66	0.93	0.86	0.86
		LL	0.65	0.93	0.86	0.86
	6	HB	0.64	0.93	0.86	0.85
		LL	0.63	0.93	0.85	0.85
	7	HB	0.62	0.91	0.83	0.87
		LL	0.62	0.91	0.83	0.87
	8	HB	0.63	0.90	0.83	0.89
		LL	0.62	0.90	0.82	0.89
	11	HB	0.64	0.90	0.82	0.92
		LL	0.64	0.90	0.81	0.92
Science	8	HB	0.62	0.90	0.80	0.91
		LL	0.62	0.90	0.80	0.91
	11	HB	0.69	0.93	0.81	0.94
		LL	0.68	0.93	0.81	0.94

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 is 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 for mathematics and science and from the high 60s to mid 80s for reading. Validity indices generally ranged from the low 90s to high 90s for mathematics and science and low 70s to low 90s for reading. 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

Grade	Item	Inter-Rater Agreement %			Percentage Awarded for Each Score Point %					
		Exact	Adjacent	Validity	0	1	2	3	4	B/NS
4	1	97	3	94	15	12	23	18	32	0
	2	97	3	99	16	27	19	28	10	0
5	1	94	6	96	40	11	26	17	5	1
	2	96	4	98	7	15	18	24	34	1
6	1	97	3	97	16	50	25	6	4	1
	2	92	8	99	16	43	29	9	2	1
7	1	89	10	92	18	44	21	8	9	1
	2	92	8	93	13	23	24	24	15	1
8	1	94	6	96	58	14	22	2	3	1
	2	99	1	97	14	22	30	27	6	0
11	1	92	8	94	39	29	15	12	0	5
	2	88	12	92	15	40	22	13	5	5

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

Table 18–8. Inter-Rater Agreement and Percentage Awarded for Each Score Point for OE Items—Reading

Grade	Item	Inter-Rater Agreement %			Percentage Awarded for Each Score Point %				
		Exact	Adjacent	Validity	0	1	2	3	B/NS
4	1	85	15	81	22	26	33	17	1
	2	92	8	91	19	21	19	39	2
5	1	78	22	81	9	33	41	15	1
	2	82	18	84	15	36	30	14	4
6	1	68	32	73	12	42	36	9	1
	2	73	27	80	13	35	34	17	1
7	1	78	21	89	24	29	30	16	1
	2	71	28	81	11	41	39	7	1
8	1	73	27	81	5	24	41	29	1
	2	76	24	84	12	35	37	15	2
11	1	85	15	87	6	34	34	22	3
	2	79	21	77	5	42	31	18	3

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

Table 18–9. Inter-Rater Agreement and Percentage Awarded for Each Score Point for OE Items—Science

Grade	Item	Inter-Rater Agreement %			Percentage Awarded for Each Score Point %			
		Exact	Adjacent	Validity	0	1	2	B/NS
8	1	84	15	95	26	38	34	2
	2	94	6	98	22	41	35	2
11	1	92	8	95	17	61	18	5
	2	90	10	93	47	28	19	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 Item Response Theory (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. This 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 summarizes and synthesizes the evidence based on the *Standards*’ framework. The purposes and intended uses of PSSA-M test scores are reviewed first, 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 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 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 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 administered 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 administered 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 open-ended (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.

Item Response Theory Dimensionality

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

Strand (Reporting Category) Correlations

Correlations and disattenuated correlations among strand (reporting category) 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 strands (denoted by M.A, M.B, M.C, M.D, and M.E), the PSSA-M reading tests have two strands (denoted by R.A and R.B), and the PSSA science tests have four strands (denoted by S.A, S.B, S.C, and S.D).

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

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

M	0.82									
M.A	0.90	0.68								
M.B	0.65	0.47	0.39							
M.C	0.70	0.51	0.33	0.26						
M.D	0.65	0.47	0.38	0.36	0.42					
M.E	0.71	0.52	0.40	0.47	0.40	0.63				
R	0.50	0.44	0.32	0.32	0.33	0.43	0.83			
R.A	0.47	0.42	0.30	0.30	0.31	0.41	0.97	0.78		
R.B	0.45	0.39	0.29	0.28	0.32	0.39	0.84	0.68	0.52	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	

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

M	0.83									
M.A	0.88	0.67								
M.B	0.75	0.59	0.47							
M.C	0.63	0.39	0.38	0.51						
M.D	0.71	0.51	0.47	0.37	0.52					
M.E	0.73	0.49	0.45	0.41	0.47	0.33				
R	0.52	0.41	0.40	0.33	0.43	0.41	0.86			
R.A	0.47	0.37	0.37	0.30	0.40	0.38	0.97	0.82		
R.B	0.51	0.40	0.39	0.33	0.42	0.40	0.86	0.72	0.63	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	

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

M	0.83									
M.A	0.86	0.74								
M.B	0.66	0.45	0.27							
M.C	0.64	0.42	0.34	0.33						
M.D	0.77	0.54	0.42	0.37	0.52					
M.E	0.72	0.48	0.38	0.38	0.48	0.43				
R	0.47	0.35	0.30	0.32	0.38	0.39	0.86			
R.A	0.40	0.30	0.26	0.27	0.33	0.35	0.94	0.79		
R.B	0.47	0.35	0.31	0.31	0.38	0.39	0.93	0.74	0.73	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	

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

M	0.80													
M.A	0.77	0.59												
M.B	0.67	0.40	0.29											
M.C	0.79	0.49	0.42	0.49										
M.D	0.76	0.45	0.39	0.47	0.50									
M.E	0.68	0.39	0.37	0.45	0.42	0.42								
R	0.50	0.35	0.34	0.42	0.38	0.38	0.83							
R.A	0.46	0.31	0.32	0.38	0.35	0.34	0.92	0.72						
R.B	0.47	0.33	0.31	0.40	0.36	0.35	0.92	0.71	0.69					
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B					

Table 19–1e. Correlations among Strands for Grade 8

M	0.83														
M.A	0.69	0.41													
M.B	0.70	0.41	0.35												
M.C	0.78	0.42	0.45	0.49											
M.D	0.84	0.47	0.49	0.53	0.70										
M.E	0.70	0.38	0.39	0.47	0.49	0.47									
R	0.47	0.31	0.28	0.38	0.39	0.40	0.83								
R.A	0.44	0.29	0.26	0.36	0.36	0.37	0.93	0.71							
R.B	0.43	0.27	0.26	0.34	0.36	0.37	0.92	0.71	0.69						
S	0.47	0.32	0.29	0.40	0.36	0.40	0.64	0.60	0.58	0.80					
S.A	0.45	0.31	0.27	0.38	0.36	0.38	0.60	0.56	0.55	0.90	0.66				
S.B	0.33	0.22	0.22	0.28	0.25	0.29	0.50	0.46	0.46	0.76	0.55	0.48			
S.C	0.30	0.20	0.17	0.25	0.22	0.26	0.40	0.38	0.37	0.68	0.48	0.41	0.40		
S.D	0.30	0.21	0.18	0.28	0.22	0.25	0.38	0.37	0.33	0.68	0.47	0.43	0.39	0.42	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	S	S.A	S.B	S.C	S.D	

Table 19–1f. Correlations among Strands for Grade 11

M	0.85														
M.A	0.71	0.53													
M.B	0.73	0.46	0.49												
M.C	0.77	0.48	0.53	0.37											
M.D	0.90	0.52	0.54	0.59	0.70										
M.E	0.72	0.40	0.45	0.48	0.54	0.43									
R	0.51	0.36	0.38	0.38	0.44	0.39	0.83								
R.A	0.48	0.33	0.35	0.36	0.42	0.37	0.93	0.71							
R.B	0.46	0.33	0.35	0.35	0.39	0.36	0.92	0.71	0.70						
S	0.53	0.36	0.40	0.42	0.45	0.41	0.64	0.60	0.59	0.80					
S.A	0.49	0.33	0.36	0.40	0.42	0.37	0.61	0.58	0.56	0.91	0.72				
S.B	0.32	0.20	0.25	0.26	0.27	0.26	0.41	0.38	0.38	0.69	0.47	0.39			
S.C	0.42	0.30	0.32	0.32	0.36	0.34	0.48	0.45	0.44	0.75	0.55	0.43	0.44		
S.D	0.29	0.18	0.23	0.21	0.25	0.23	0.30	0.28	0.28	0.58	0.37	0.30	0.34	0.25	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	S	S.A	S.B	S.C	S.D	

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 weaknesses 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: Grade 4

M	-								
M.A	-	-							
M.B	-	0.91	-						
M.C	-	1.21	1.03	-					
M.D	-	0.87	0.95	1.09	-				
M.E	-	0.80	0.80	1.16	0.77	-			
R	0.61	0.59	0.57	0.68	0.56	0.60	-		
R.A	0.59	0.58	0.55	0.67	0.53	0.58	-	-	
R.B	0.69	0.66	0.65	0.75	0.68	0.67	-	1.07	-
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B

Table 19–2b. Disattenuated Strand Correlations: Grade 5

M	-								
M.A	-	-							
M.B	-	1.06	-						
M.C	-	0.67	0.79	-					
M.D	-	0.86	0.96	0.71	-				
M.E	-	1.04	1.15	0.98	1.12	-			
R	0.61	0.53	0.63	0.50	0.64	0.76	-		
R.A	0.58	0.50	0.60	0.47	0.61	0.72	-	-	
R.B	0.70	0.62	0.72	0.57	0.72	0.88	-	0.99	-
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B

Table 19–2c. Disattenuated Strand Correlations: Grade 6

M	-								
M.A	-	-							
M.B	-	0.99	-						
M.C	-	0.85	1.12	-					
M.D	-	0.88	1.10	0.90	-				
M.E	-	0.85	1.10	1.02	1.01	-			
R	0.55	0.44	0.62	0.59	0.57	0.65	-		
R.A	0.50	0.39	0.56	0.54	0.52	0.60	-	-	
R.B	0.60	0.48	0.68	0.64	0.62	0.69	-	0.99	-
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B

Table 19–2d. Disattenuated Strand Correlations: Grade 7

M	-								
M.A	-	-							
M.B	-	0.98	-						
M.C	-	0.92	1.12	-					
M.D	-	0.83	1.04	0.95	-				
M.E	-	0.79	1.06	0.99	0.93	-			
R	0.61	0.49	0.69	0.66	0.59	0.64	-		
R.A	0.60	0.48	0.70	0.64	0.58	0.62	-	-	
R.B	0.63	0.51	0.68	0.69	0.61	0.66	-	1.00	-
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B

Table 19–2e. Disattenuated Strand Correlations: Grade 8

M	-														
M.A	-	-													
M.B	-	1.07	-												
M.C	-	0.93	1.08	-											
M.D	-	0.88	0.99	0.91	-										
M.E	-	0.86	0.96	0.98	0.86	-									
R	0.57	0.52	0.52	0.60	0.51	0.64	-								
R.A	0.57	0.54	0.52	0.61	0.51	0.63	-	-							
R.B	0.57	0.51	0.53	0.59	0.51	0.64	-	1.01	-						
S	0.58	0.56	0.54	0.64	0.48	0.65	0.79	0.79	0.79	-					
S.A	0.62	0.60	0.56	0.67	0.53	0.68	0.82	0.82	0.82	-	-				
S.B	0.53	0.50	0.53	0.58	0.44	0.60	0.79	0.78	0.80	-	0.97	-			
S.C	0.52	0.51	0.46	0.57	0.43	0.60	0.71	0.71	0.70	-	0.93	0.94	-		
S.D	0.52	0.51	0.48	0.62	0.40	0.57	0.65	0.68	0.62	-	0.90	0.95	0.96	-	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	S	S.A	S.B	S.C	S.D	

Table 19–2f. Disattenuated Strand Correlations: Grade 11

M	-														
M.A	-	-													
M.B	-	0.90	-												
M.C	-	1.08	1.25	-											
M.D	-	0.85	0.92	1.16	-										
M.E	-	0.85	0.97	1.21	0.98	-									
R	0.60	0.54	0.59	0.69	0.57	0.65	-								
R.A	0.61	0.54	0.59	0.71	0.59	0.66	-	-							
R.B	0.60	0.54	0.60	0.68	0.56	0.65	-	1.01	-						
S	0.64	0.55	0.64	0.77	0.60	0.70	0.79	0.80	0.78	-					
S.A	0.62	0.54	0.61	0.77	0.58	0.67	0.79	0.80	0.78	-	-				
S.B	0.56	0.45	0.56	0.69	0.51	0.64	0.72	0.72	0.73	-	0.88	-			
S.C	0.69	0.61	0.69	0.80	0.64	0.77	0.79	0.80	0.79	-	0.97	1.04	-		
S.D	0.62	0.50	0.65	0.69	0.59	0.70	0.67	0.67	0.67	-	0.86	0.97	1.03	-	
	M	M.A	M.B	M.C	M.D	M.E	R	R.A	R.B	S	S.A	S.B	S.C	S.D	

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

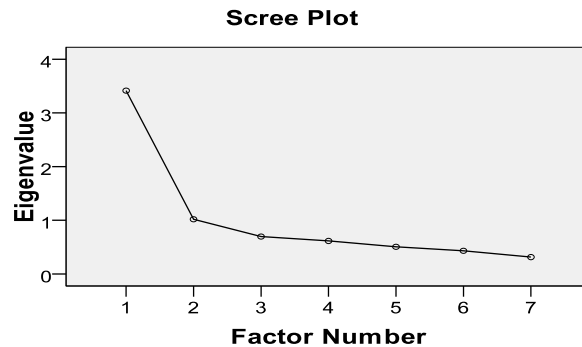
In order to further explore the internal structure of the PSSA-M, an exploratory factor analysis (EFA) of the strand scores was conducted. The PSSA-M final data file (see Chapter Nine) was used to create the observed correlation matrices shown in Tables 19–1a through 19–1f, which in turn were used in 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 about 49 percent of the total variance, while the second factor explained about 15 percent of the total variance. The first two factors had an eigenvalue greater than 1.0, typically suggesting a two-factor solution using the Kaiser criterion. Based on this finding and the prior belief that there should be two distinct factors at Grade 4 (one for mathematics and another for reading), a two-factor solution was further explored.

Table 19–3. Eigenvalues and Explained Variance for Grade 4

Factor	Eigenvalue	%
1	3.42	48.79
2	1.02	14.56
3	0.70	9.96
4	0.62	8.79
5	0.51	7.23
6	0.43	6.18
7	0.31	4.49

Figure 19–1. Scree Plot for Grade 4



The pattern loadings resulting from the two-factor solution are presented in Table 19–4a. The pattern loadings have simple structure which show that the five mathematics domains clearly loaded on the first factor while the two reading domains clearly loaded on the second factor. The respective factor loadings are quite high. The factor correlation matrix shows that the correlation between the two latent factors is 0.61, which is equal to the disattenuated correlation between mathematics and reading.

Table 19–4a. Factor Loadings for Grade 4

Domain	Factor	
	1	2
Mathematics		
M.A	0.78	0.01
M.B	0.57	0.01
M.C	0.67	-0.06
M.D	0.58	0.02
M.E	0.63	0.09
Reading		
R.A	0.02	0.82
R.B	-0.01	0.83
Correlation (F1, F2) = 0.61		

Similar results were found at the other grades. The eigenvalue scree plots consistently indicated a multi-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. The Pattern loadings clearly suggested that the PSSA tests measured different but correlated constructs.

Table 19–4b. Factor Loadings for Grade 5

Domain	Factor	
	1	2
Mathematics		
M.A	0.81	-0.06
M.B	0.74	-0.02
M.C	0.51	0.05
M.D	0.61	0.09
M.E	0.62	0.07
Reading		
R.A	-0.01	0.84
R.B	0.04	0.83
Correlation (F1, F2) = 0.62		

Table 19–4c. Factor Loadings for Grade 6

Domain	Factor	
	1	2
Mathematics		
M.A	0.79	-0.07
M.B	0.59	-0.01
M.C	0.53	0.04
M.D	0.70	0.02
M.E	0.61	0.09
Reading		
R.A	-0.03	0.87
R.B	0.05	0.84
Correlation (F1, F2) = 0.55		

Table 19–4d. Factor Loadings for Grade 7

Domain	Factor	
	1	2
Mathematics		
M.A	0.71	-0.05
M.B	0.58	0.01
M.C	0.69	0.04
M.D	0.66	0.02
M.E	0.58	0.06
Reading		
R.A	0.00	0.84
R.B	0.02	0.83
Correlation (F1, F2) = 0.60		

Table 19–4e. Factor Loadings for Grade 8

Domain	Factor		
	1	2	3
Mathematics			
M.A	0.61	0.03	-0.04
M.B	0.69	-0.04	-0.05
M.C	0.67	0.09	-0.03
M.D	0.79	-0.10	0.05
M.E	0.56	0.06	0.07
Reading			
R.A	0.02	0.10	0.73
R.B	-0.02	-0.04	0.92
Science			
S.A	0.06	0.63	0.15
S.B	-0.03	0.63	0.10
S.C	-0.02	0.64	-0.02
S.D	0.01	0.71	-0.12
Correlation (F1, F2) = 0.55		Correlation (F1, F3) = 0.56	
Correlation (F2, F3) = 0.73			

Table 19–4f. Factor Loadings for Grade 11

Domain	Factor		
	1	2	3
Mathematics			
M.A	0.65	-0.04	0.04
M.B	0.69	0.03	-0.02
M.C	0.77	0.01	-0.04
M.D	0.80	-0.03	0.03
M.E	0.59	0.07	0.02
Reading			
R.A	0.01	-0.02	0.87
R.B	0.01	0.03	0.80
Science			
S.A	0.04	0.59	0.21
S.B	-0.05	0.66	-0.01
S.C	0.04	0.69	0.00
S.D	0.03	0.53	-0.07
Correlation (F1, F2) = 0.61		Correlation (F1, F3) = 0.59	
Correlation (F2, F3) = 0.74			

Taken as a whole, all the internal structure evidence presented above generally indicates that related elements of each of the PSSA-M tests correlate in the intended manner. Different PSSA-M subject area tests seem to measure different constructs. Additionally, the strands within each subject area have stronger relationships than those across subject strands. This further supports using a total score to report student performance in the different subject areas.

The strand scores present more of a mixed message. Since the strands in each subject area were designed to measure distinct components of the subject area, it is reasonable to expect that the inter-subject strand correlations should be positive and strong, but ideally, not extremely high. However, the disattenuated correlations imply that some strands are essentially measuring the same constructs. Consequently, there may be less support for providing results for some strand scores beyond the total score. While there is content rationale underlying the creation of the strand scores, the empirical correlations illustrate that caution is required when using the strand scores as a way to identify individual student's strengths and weaknesses. Certainly, instructional programs should not be based on strand score information alone but in conjunction with other sources of evidence available (e.g., teacher observations, 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 apparent 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 review by content and bias specialists. The statistical results can help organize the review so 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 a 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-HAENZEL 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 and 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 and 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.¹⁷

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
Mathematics	4	11	17	2	0	0	0	30	11	18	1	0	0	0	30
	5	11	19	0	0	0	0	30	15	15	0	0	0	0	30
	6	13	14	1	2	0	0	30	15	15	0	0	0	0	30
	7	19	8	0	1	0	2	30	15	13	1	1	0	0	30
	8	18	10	0	1	0	1	30	19	11	0	0	0	0	30
	11	15	12	1	2	0	0	30	14	15	0	1	0	0	30
Reading	4	14	16	0	0	0	0	30	10	20	0	0	0	0	30
	5	15	15	0	0	0	0	30	8	22	0	0	0	0	30
	6	13	17	0	0	0	0	30	12	16	0	2	0	0	30
	7	16	13	1	0	0	0	30	12	17	0	1	0	0	30
	8	13	15	1	1	0	0	30	8	22	0	0	0	0	30
	11	13	16	0	1	0	0	30	12	16	0	2	0	0	30
Sci.	8	11	16	1	1	0	1	30	10	19	0	1	0	0	30
	11	15	11	1	2	1	0	30	13	17	0	0	0	0	30

¹⁷ 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
Mathematics	4	2	0	0	0	0	0	2	1	1	0	0	0	0	2
	5	0	1	1	0	0	0	2	0	2	0	0	0	0	2
	6	1	1	0	0	0	0	2	0	2	0	0	0	0	2
	7	1	1	0	0	0	0	2	0	2	0	0	0	0	2
	8	2	0	0	0	0	0	2	1	1	0	0	0	0	2
	11	1	1	0	0	0	0	2	0	2	0	0	0	0	2
Reading	4	2	0	0	0	0	0	2	0	2	0	0	0	0	2
	5	1	1	0	0	0	0	2	1	1	0	0	0	0	2
	6	2	0	0	0	0	0	2	2	0	0	0	0	0	2
	7	2	0	0	0	0	0	2	1	1	0	0	0	0	2
	8	2	0	0	0	0	0	2	2	0	0	0	0	0	2
	11	2	0	0	0	0	0	2	2	0	0	0	0	0	2
Sci.	8	0	1	1	0	0	0	2	0	2	0	0	0	0	2
	11	0	1	0	1	0	0	2	0	1	0	1	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).

Table 19–6. Percentage of Students Scoring in the Proficient-M or Advanced-M Category

		PSSA-M 2010	PSSA-M 2011	PSSA 2010	PSSA 2011
Mathematics	4	59.4	53.7	84.8	85.2
	5	51.1	45.2	74.4	76.3
	6	48.1	42.8	78.0	78.9
	7	41.2	41.0	78.0	78.6
	8	40.8	38.3	75.1	76.9
	11	33.2	32.4	59.6	60.4
Reading	4		63.9		73.4
	5		61.4		67.2
	6		56.2		69.9
	7		50.6		76.0
	8		46.1		81.7
	11		45.5		69.2
Sci.	8		47.8		58.3
	11		46.6		40.8

Lane and Stone (2002) also summarized possible unintended outcomes:

- Narrowing of curriculum and instruction to focus on only the specific standards assessed and ignore 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 provides accurate and clear test score and report information to 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.

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.

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 (see 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 were 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.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME]. (1999). *Standards for educational and psychological testing*. 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 ed.). 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). [Computer Software] Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement & Assessment. CASMA: education.uiowa.edu/casma.
- 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). *Technical 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, A., & 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*. Minneapolis, MN: University of Minnesota. 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., pp.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.education.state.pa.us>. Also available from <http://www.pacode.com/secure/data/022/Chapter4/s4.51.html>.
- Pennsylvania Department of Education. (2004). *Mathematics item and scoring sampler*. Retrieved December 13, 2004, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2004). *Reading item and scoring sampler*. Retrieved December 13, 2004, from <http://www.education.state.pa.us.us>

- Pennsylvania Department of Education. (2004, April). *Assessment anchors and eligible content*. Retrieved December 13, 2004, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2004, November). *Mathematics assessment handbook*. Retrieved December 13, 2004, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2004, November). *Reading assessment handbook*. Retrieved December 13, 2004, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005, December). *2005–2006 Mathematics assessment handbook*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005, December). *2005–2006 Reading assessment handbook*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005). *2005–2006 Mathematics item and scoring sampler*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005). *2005–2006 Reading item and scoring sampler*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005, December). *2005–2006 Writing assessment handbook*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2005). *2005–2006 Writing item and scoring sampler*. Retrieved September 14, 2005, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006). *2006–2007 Mathematics item and scoring sampler*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006). *2006–2007 Reading item and scoring sampler*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006). *2006–2007 Writing item and scoring sampler*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006, December). *2006–2007 Writing assessment handbook*. Retrieved January 30, 2006, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006). *2006–2007 Science item and scoring sampler*. Retrieved March 15, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2006, November). *Science assessment handbook*. Retrieved March 15, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2007, January). *2006–2007 Mathematics assessment handbook*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2007, January). *2006–2007 Reading assessment handbook*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2007, January). *2007 Accommodations guidelines for students with IEPs, students with 504 plans, English language learners, and all students*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2007). *Assessment anchors and eligible content*. 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*. Retrieved January 30, 2007, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2007, March). *PSSA reading and mathematics directions for administration manual*. Retrieved April 2, 2007, from <http://www.education.state.pa.us.us>.
- Pennsylvania Department of Education. (2007). *2008 PSSA accommodations guidelines for students with IEPs and students with 504 plans*. Retrieved March 4, 2008, from <http://www.education.state.pa.us.us>.
- Pennsylvania Department of Education. (2008). *PSSA 2008 Handbook for assessment coordinators and administrators: Grades 3–8 and 11 reading and mathematics*. Retrieved March 4, 2008, from <http://www.education.state.pa.us.us>.
- Pennsylvania Department of Education. (2008). *2008–2009 Mathematics item and scoring sampler*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2008). *2008–2009 Reading item and scoring sampler*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2008). *2008–2009 Science item and scoring sampler*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2008). *2008–2009 Writing item and scoring sampler*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2009). *PSSA accommodations guidelines for students with IEPs and students with 504 plans*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.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. (2009). *2008–2009 Assessment handbook*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2009). *The 2008–2009 PSSA handbook for assessment coordinators: Writing, reading and mathematics, science*. Retrieved February 10, 2009, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2010). *2009–2010 Assessment handbook*. Retrieved February 24, 2010 from <http://www.education.state.pa.us>
- Pennsylvania Department of Education. (2010). *PSSA and PSSA-M Accommodations guidelines for students with IEPs and students with 504 plans, revised 1-11-2010*. Retrieved February 24, 2010, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2010). *The 2009-2010 PSSA handbook for assessment coordinators: Writing, reading and mathematics, science*. Retrieved February 24, 2010, from <http://www.education.state.pa.us.us>
- Pennsylvania Department of Education. (2011). *PSSA, PSSA-M, Keystone (paper/pencil) accommodations guidelines for students with IEPs and students with 504 plans, revised 1-12-2011*. Retrieved February 25, 2011 from <http://www.education.state.pa.us.us>

- Pennsylvania Department of Education. (2011). *2010-2011 PSSA handbook for assessment coordinators: Reading and mathematics, writing, science*. Retrieved February 25, 2011, from <http://www.education.state.pa.us.us>
- Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, 8(2), 111–120.
- Raïche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, 19:1, 1012.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.
- Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. W. (in press). *Accommodations for English Language Learners*. San Francisco, CA: WestEd.
- 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.
- Smith, R. & Miao, C. (1994). Assessing unidimensionality for Rasch measurement. Chapter 18 in M. Wilson (Ed.) *Objective Measurement: Theory into Practice*. Vol. 2. Norwood NJ: Ablex.
- 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*. Thousand Oaks: Sage.
- 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* (NISE Research Monograph No. 18). Madison, WI: University of Wisconsin–Madison, 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)*. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research.
- WINSTEPS (2000). *WINSTEPS[®] Rasch measurement*. Copyright John M. Linacre.
- Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: 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

PENNSYLVANIA DEPARTMENT OF EDUCATION
About the Mathematics Assessment Anchors*

Introduction

This is a brief introduction to the Mathematics Assessment Anchors for the PSSA-M. 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 *General Introduction* provided on the website and the *Frequently Asked Questions*.

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

Appendix A: Assessment Anchor Explanations

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)

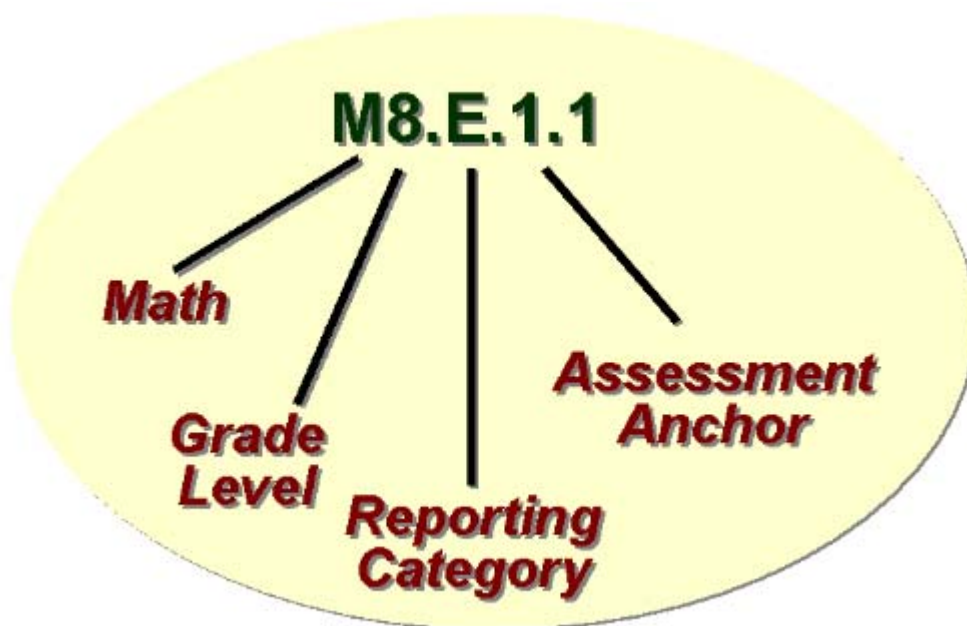
Important Patterns

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.

How to Read the Assessment Anchors

The Mathematics Assessment Anchors begin with an “M” to distinguish them from the Reading Assessment Anchors “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 labeling 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 label 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.

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).

PENNSYLVANIA DEPARTMENT OF EDUCATION
About the Reading Assessment Anchors

Introduction

This is a brief introduction to the Reading Assessment Anchors for the PSSA-M. 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 *General Introduction* provided on the website and the *Frequently Asked Questions*.

How the Assessment Anchors Connect to the Standards

The PA Academic Standards for Reading, Writing, Speaking and Listening are:

- 1.1 Learning to Read Independently
- 1.2 Reading Critically in All Content Areas
- 1.3 Reading, Analyzing and Interpreting Literature
- 1.4 Types of Writing
- 1.5 Quality of Writing
- 1.6 Speaking and Listening
- 1.7 Characteristics and Function of the English Language
- 1.8 Research

In the past, the Reading PSSA assessed standards 1.1, 1.2, 1.3, 1.7 and 1.8 in Grades 5, 8 and 11. The Writing PSSA assessed standards 1.4 and 1.5. Speaking and Listening have always been assessed through local assessments. *Because of the shift to create a clearer and more focused test using the Assessment Anchors, the 2005 PSSA will only assess the first three reading standards.* Learning to read independently and critically and the ability to analyze and interpret are at the heart of what students must be able to do to be good readers in today's society. Standards 1.7 and 1.8 are not specific to reading and for the most part these standards are better assessed at the district level.

How the Assessment Anchors Are Organized

Instead of having five reporting categories, the Assessment Anchors will have two:

Reporting Category	Standard
A. Comprehension and Reading Skills	1.1 (Learning to Read Independently) and 1.2 (Reading Critically in All Content Areas)
B. Interpretation and Analysis of Fiction and Nonfiction Text	1.1 (Learning to Read Independently) and 1.2 (Reading Critically in All Content Areas) and 1.3 Reading, Analyzing and Interpreting Literature)

Important Patterns

There are additional patterns within each Reporting Category. Each Reporting Category includes some basic elements that are consistent across all of the grade levels.

A. *Comprehension and Reading Skills*

Comprehension and Reading Skills has two basic elements:

- A.1 Fiction
- A.2 Nonfiction

B. *Interpretation and Analysis of Fiction and Nonfiction Text*

Interpretation and Analysis of Fiction and Nonfiction Text has three basic elements:

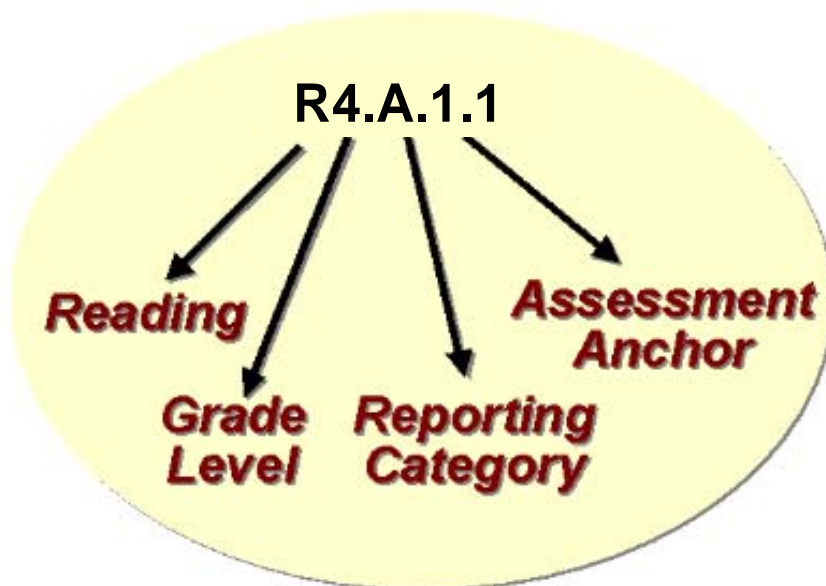
- B.1 Components within text **or** components within and across texts
- B.2 Literary Devices
- B.3 Concepts and Organization of Nonfiction Text

The Anchors generally target the same comprehension skills from Grades 4 through 8 and 11, although the depth of knowledge required to comprehend the text grows in complexity over the years. In addition, the expectation is that the level of texts themselves will grow in complexity.

How to Read the Assessment Anchors

The Reading Assessment Anchors begin with “R” to distinguish them from the Mathematics Assessment Anchors, which begin with “M”. The number after the “R” in the label is the grade level (e.g., R4 would be Reading at fourth grade). The second letter in the labeling system is the Reporting Category (A or B). The same reporting categories continue across all Grades 4 through 8 and 11. The final number in the label 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, R4.A.1.1 is a Reading Assessment Anchor (R stands for Reading) at 4th grade (4). The A indicates that this Anchor is in the Comprehension and Reading Skills Reporting Category and the 1.1 means that it is the first Assessment Anchor in that Reporting Category. (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 crosswalk the Anchors to the Standards.

PENNSYLVANIA DEPARTMENT OF EDUCATION
Overview of Reading Assessment Anchors

GRADE 4

R4.A. Comprehension and Reading Skills

- R4.A.1 Understand Fiction Appropriate to Grade level
- R4.A.2 Understand Nonfiction Appropriate to Grade Level

R4.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R4.B.1 Understand Components Within and Between Texts
- R4.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R4.B.3 Understand Concepts and Organization of Nonfictional Text

GRADE 5

R5.A. Comprehension and Reading Skills

- R5.A.1 Understand Fiction Appropriate to Grade level
- R5.A.2 Understand Nonfiction Appropriate to Grade Level

R5.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R5.B.1 Understand Components Within and Between Texts
- R5.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R5.B.3 Understand Concepts and Organization of Nonfictional Text

GRADE 6

R6.A. Comprehension and Reading Skills

- R6.A.1 Understand Fiction Appropriate to Grade level
- R6.A.2 Understand Nonfiction Appropriate to Grade Level

R6.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R6.B.1 Understand Components Within and Between Texts
- R6.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R6.B.3 Understand Concepts and Organization of Nonfictional Text

GRADE 7

R7.A. Comprehension and Reading Skills

- R7.A.1 Understand Fiction Appropriate to Grade level
- R7.A.2 Understand Nonfiction Appropriate to Grade Level

R7.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R7.B.1 Understand Components Within and Between Texts
- R7.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R7.B.3 Understand Concepts and Organization of Nonfictional Text

GRADE 8

R8.A. Comprehension and Reading Skills

- R8.A.1 Understand Fiction Appropriate to Grade level
- R8.A.2 Understand Nonfiction Appropriate to Grade Level

R8.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R8.B.1 Understand Components Within and Between Texts
- R8.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R8.B.3 Understand Concepts and Organization of Nonfictional Text

GRADE 11

R11.A. Comprehension and Reading Skills

- R11.A.1 Understand Fiction Appropriate to Grade level
- R11.A.2 Understand Nonfiction Appropriate to Grade Level

R11.B. Interpretation and Analysis of Fictional and Nonfictional Text

- R11.B.1 Understand Components Within and Between Texts
- R11.B.2 Understand Literary Devices in Fictional and Nonfictional Text
- R11.B.3 Understand Concepts and Organization of Nonfictional Text

Pennsylvania Science

About the Science Assessment Anchors

Introduction

The Pennsylvania Science Assessment is based on the Academic Standards adopted by the State Board of Education in January of 2002. The standards are comprised of two documents: Science and Technology Standards and Environment and Ecology Standards. These documents contain seventeen important categories that describe what students need to know. The purpose of the Assessment Anchors is to articulate essential and assessable elements, and to provide clarity for instruction and for the focus of the state assessment in grades 8 and 11.

How the Assessment Anchors Connect to the Standards

The Pennsylvania Academic Standards for Science are:

- | | |
|--|---|
| 3.1 Unifying Themes | 4.1 Watersheds and Wetlands |
| 3.2 Inquiry and Design | 4.2 Renewable and Nonrenewable Resources |
| 3.3 Biological Sciences | 4.3 Environmental Health |
| 3.4 Physical Science, Chemistry,
and Physics | 4.4 Agriculture and Society |
| 3.5 Earth Sciences | 4.5 Integrated Pest Management |
| 3.6 Technology Education | 4.6 Ecosystems and their Interactions |
| 3.7 Technological Devices | 4.7 Threatened, Endangered and Extinct Species |
| 3.8 Science, Technology and
Human Endeavors | 4.8 Humans and the Environment |
| | 4.9 Environmental Laws and Regulations |

All of the Science Standards categories are included in the Assessment Anchors, but the anchors tighten the focus of what is assessed. The Assessment Anchors clarify what is expected from grade span to grade span (K-4, 5-7, and 8-10). In addition, the Assessment Anchors have fewer Reporting Categories to help create more reliable scores (meaning that there are more items per reporting category making interpretations about what students actually know more reliable). Rather than reporting student results in all 17 standards, the reports will be organized into four reporting categories.

How the Assessment Anchors are Organized

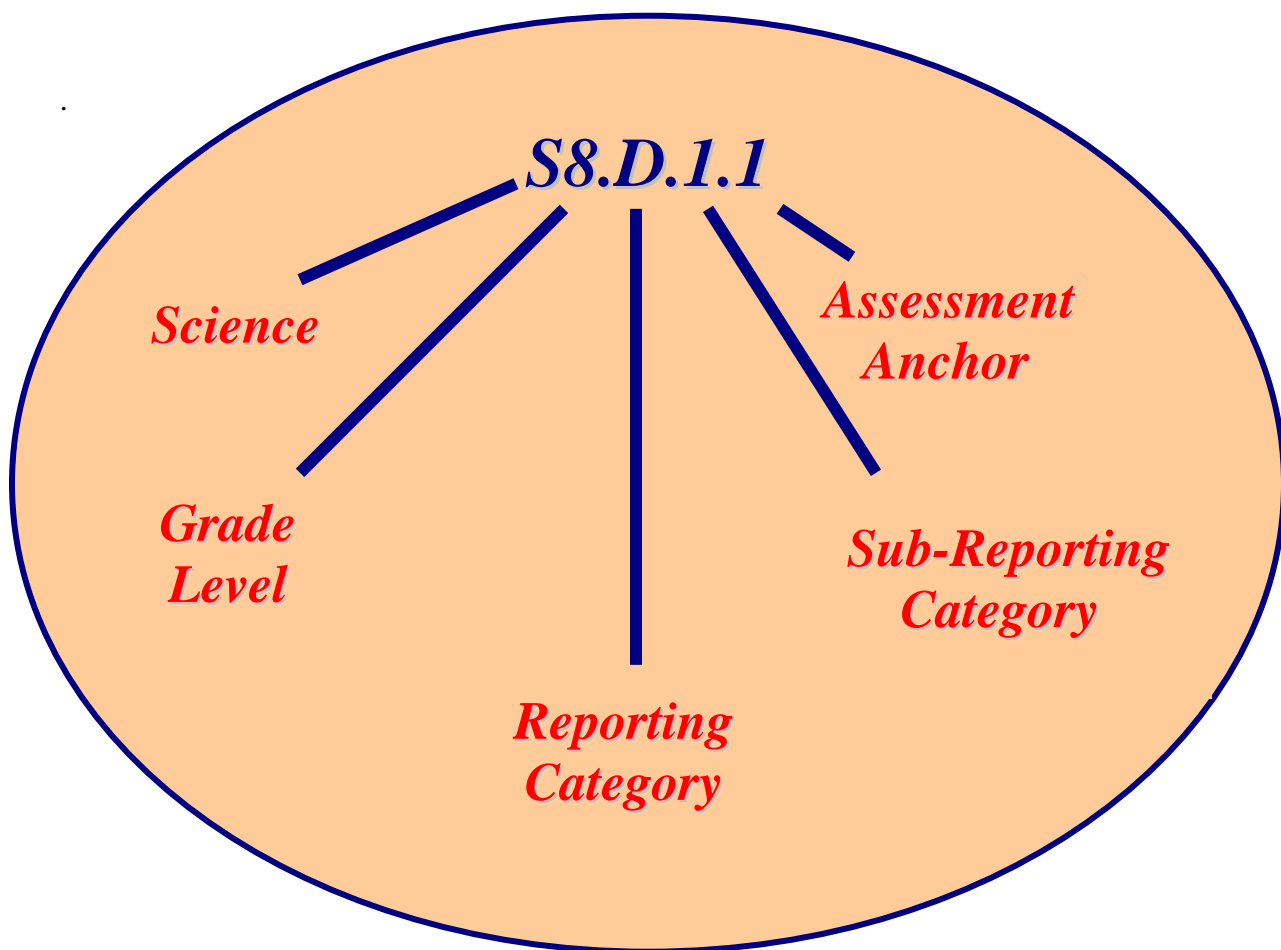
The four reporting categories are similar to those used by the National Assessment of Educational Progress (NEAP) and The Third International Mathematics and Science Study (TIMSS). The four categories for the assessment anchors are included in these major assessments, but are organized differently. Below are the four student reporting categories for the assessment anchors for the PSSA and the PSSA-M in Science and the related standards.

Appendix A: Assessment Anchor Explanations

Reporting Categories	Connections to the Standards
A. The Nature of Science	3.1 Unifying Themes of Science 3.2 Inquiry and Design 3.6 Technology Education 3.7 Technological Devices 3.8 Science, Technology, and Human Endeavors 4.4 Agriculture and Society 4.6 Ecosystems and their Interactions 4.7 Threatened, Endangered, and Extinct Species 4.8 Humans and the Environment
B. Biological Sciences	3.1 Unifying Themes of Science 3.3 Biological Sciences 4.2 Renewable and Nonrenewable Resources 4.3 Environmental Health 4.6 Ecosystems and Their Interactions 4.7 Threatened, Endangered, and Extinct Species
C. Physical Sciences	3.2 Inquiry and Design 3.4 Physical Science, Chemistry, and Physics 3.6 Earth Sciences
D. Earth and Space Sciences	3.2 Inquiry and Design 3.4 Physical Science, Chemistry, and Physics 3.5 Earth Sciences 3.7 Technological Devices 4.1 Watersheds and Wetlands 4.2 Renewable and Nonrenewable Resources 4.8 Humans and the Environment

How to Read the Assessment Anchors

All of the Science Assessment Anchors begin with an “S” to indicate science. The number after the “S” in the label is the grade level (e.g., S8 would be Science at eighth grade). The second letter in the labeling system is the Reporting Category (A through D) followed by the sub-reporting category number. The same reporting categories continue across all Grade levels, 4, 8, and 11. The final number in the label is the actual Assessment Anchor number (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. (*See example below.*)



For example, **S8.D.1.1** is the code for the first science (S) assessment anchor for Grade 8 in the reporting category of (D) Earth and Space Sciences, and the sub-category of Earth Features and Processes That Change Earth and Its Resources.

Overview of Science Assessment Anchors

**Note that on this overview document, the grade level does not appear in the reporting categories because these occur at all grade levels (8 and 11).*

SA. The Nature of Science

SA.1 Reasoning and Analysis

SA.2 Processes, Procedures, and Tools of Scientific Investigations

SA.3 Systems, Models, and Patterns

SB. Biological Sciences

SB.1 Structure and Function of Organisms

SB.2 Continuity of Life

SB.3 Ecological Behavior and Systems

SC. Physical Sciences

SC.1 Structure, Properties, and Interaction of Matter and Energy

SC.2 Forms, Sources, Conversion, and Transfer of Energy

SC.3 Principles of Motion and Force

SD. Earth and Space Sciences

SD.1 Earth Features and Processes that Change Earth and Its Resources

SD.2 Weather, Climate, and Atmospheric Processes

SD.3 Composition and Structure of the Universe

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

OTOff task

IL.....Illegible

LOE.....Response in a language other than English

This document is available on the PDE website.

Note: The PSSA General Description of Mathematics Scoring Guidelines also applies to the PSSA-M mathematics assessments.

PENNSYLVANIA DEPARTMENT OF EDUCATION
PSSA

General Scoring Guidelines for Open-Ended Reading Items

3 Points

- The response provides a complete answer to the task (e.g., a statement that offers a correct answer as well as text-based support).
- The response provides specific, appropriate and accurate details (e.g., naming, describing, explaining, or comparing) or examples.

2 Points

- The response provides a partial answer to the task (e.g., indicates some awareness of the task and at least one text-based detail).
- The response attempts to provide sufficient, appropriate details (e.g., naming, describing, explaining, or comparing) or examples; may contain minor inaccuracies.

1 Point

- The response provides an incomplete answer to the task (e.g., indicating either a misunderstanding of the task or no text-based details).
- The response provides insufficient or inappropriate details or examples that have a major effect on accuracy.
- The response consists entirely of relevant copied text.

0 Points

- The response provides insufficient material for scoring.
- The response is inaccurate in all aspects.

Categories within zero reported separately:

- **BLK (blank) = no response or written refusal to respond or too brief to determine response.**
- **OT = off task/topic.**
- **LOE = response in a language other than English.**
- **IL = illegible.**

Note: The PSSA General Scoring Guidelines for Open-Ended Reading Items also apply to the PSSA-M reading assessments.

PENNSYLVANIA DEPARTMENT OF EDUCATION

PSSA
SCIENCE

DESCRIPTION OF SCORING GUIDELINES FOR 2-POINT OPEN-ENDED ITEMS:

General Description of Science Scoring Guidelines:

- 2 – The response demonstrates a *thorough* understanding of the scientific content, concepts, and procedures required by the task/s.**

The response provides a clear, complete, and correct response as required by the task/s. Response may contain a minor blemish (e.g., misspelled words) or omission in work or explanation that does not detract from demonstrating a thorough understanding.

- 1 – The response demonstrates a *partial* understanding of the scientific content, concepts, and procedures required by the task/s.**

The response is somewhat correct with partial understanding of the required scientific content, concepts, and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

- 0 – The response provides insufficient evidence to demonstrate any understanding of the scientific content, concepts, and procedures as required by the task/s for that grade level.**

Response may show only information copied or rephrased from the question or insufficient correct information to receive a score of 1.

Special Categories within zero reported separately:

BLK – Blank, entirely erased or written refusal to respond

OT – Off Task

IL – Illegible

LOE – Response in a language other than English

Note: The PSSA Description of Scoring Guidelines for 2-Point Open-Ended Items for Science also applies to the PSSA-M science assessments.

Appendix C:

2011 PSSA-M Tally Sheets

Comparing the 2011 PSSA Core with the 2011 PSSA-M Core

Mathematics Grade 4

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	1	
A.1.1.1	1	1	1	1	
A.1.1.2	1	1	2	0	
A.1.1.3	1	1	2	1	
A.1.1.4	2	1	2	1	
A.1.2.1	2	1	2	1	
A.1.2.2	1	1	1	1	
A.1.3.1	2	1	2	1	
A.1.3.2	4	1	2	1	
A.2	1	0	1	0	
A.2.1.1	2	0	0	0	
A.2.1.2	2	1	2	1	
A.3	0	1	1	0	
A.3.1.1	3	1	2	1	
A.3.1.2	2	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	2	1	2	1	
B.1	0	0	1	0	
B.1.1.1	1	1	1	1	
B.1.1.2	1	1	1	1	
B.1.1.3	2	0	1	1	
B.1.1.4	2	1	0	1	
B.2	0	0	0	0	
B.2.1.1	2	0	1	1	
B.2.2.1	1	2	2	0	
C.1	0	0	0	1	
C.1.1.1	1	0	1	0	
C.1.1.2	0	0	0	1	
C.1.2.1	1	0	1	0	
C.1.2.2	2	1	2	0	
C.2	0	0	0	0	
C.2.1.1	2	1	2	0	
C.3	1	1	1	0	
C.3.1.1	1	0	1	0	
D.1	1	0	1	0	
D.1.1.1	1	0	1	0	

G4 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	14	22%	8	25%
A.2	5	8%	1	3%
A.3	10	16%	4	13%
B.1	6	10%	3	9%
B.2	3	5%	2	6%
C.1	4	6%	1	3%
C.2	2	3%	1	3%
C.3	2	3%	1	3%
D.1	5	8%	3	9%
D.2	3	5%	2	6%
E.1	6	10%	4	13%
E.3	3	5%	2	6%

G4 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	29	46%	13	41%
B	9	14%	5	16%
C	8	13%	3	9%
D	8	13%	5	16%
E	9	14%	6	19%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.1.1.2	1	1	1	0	
D.1.1.3	1	1	1	0	
D.1.2.1	0	0	0	0	
D.1.2.2	1	1	1	0	
D.2	0	0	0	1	
D.2.1.1	2	1	0	1	
D.2.2.1	0	1	1	0	
D.2.2.2	1	0	1	0	
E.1	0	0	1	0	
E.1.1.1	2	2	2	1	
E.1.2.1	3	1	1	1	
E.1.2.2	1	1	2	1	
E.3	0	0	0	0	
E.3.1.1	3	2	2	2	
Totals	63	32	54	26	

Mathematics Grade 5

Anchor or Eligible Content	2011 PSSA Core	2011 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	1	1	1	1	
A.1.3.2	1	0	0	1	
A.1.3.3	2	0	2	0	
A.1.4.1	1	1	2	0	
A.1.4.2	2	1	1	1	
A.1.5.1	1	1	2	1	
A.1.6.1	3	0	1	1	
A.1.6.2	1	1	1	1	
A.2	1	1	1	1	
A.2.1.1	2	1	0	1	
A.2.1.2	2	1	1	1	
A.2.1.3	1	0	0	1	
A.3	0	0	0	0	
A.3.1.1	2	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	0	0	1	0	
B.1.2.1	0	1	1	0	
B.1.2.2	1	1	0	2	
B.1.3.1	1	1	1	0	
B.1.3.2	1	0	0	1	
B.2	0	0	1	0	
B.2.1.1	1	0	1	0	
B.2.2.1	1	1	0	1	
B.2.2.2	1	1	1	1	
B.2.2.3	0	1	1	0	
C.1	0	0	1	0	
C.1.1.1	2	1	2	0	
C.1.1.2	1	0	0	1	
C.1.2.1	2	1	1	0	
C.2	0	0	0	1	

G5 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	17	27%	8	25%
A.2	6	10%	3	9%
A.3	5	8%	2	6%
B.1	4	6%	3	9%
B.2	3	5%	3	9%
C.1	5	8%	2	6%
C.2	5	8%	3	9%
D.1	4	6%	3	9%
D.2	4	6%	2	6%
E.1	2	3%	1	3%
E.2	3	5%	1	3%
E.3	5	8%	1	3%

G5 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	28	44%	13	41%
B	7	11%	6	19%
C	10	16%	5	16%
D	8	13%	5	16%
E	10	16%	3	9%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
C.2.1.1	2	1	2	0	
C.2.1.2	3	2	2	0	
D.1	1	0	1	0	
D.1.1.1	1	1	2	0	
D.1.1.2	1	1	1	1	
D.1.2.1	1	1	1	1	
D.2	0	0	0	0	
D.2.1.1	1	1	2	1	
D.2.1.2	3	1	1	2	
E.1	0	1	1	0	
E.1.1.1	2	0	1	0	
E.2	0	0	0	1	
E.2.1.1	2	1	1	0	
E.2.1.2	1	0	2	0	
E.3	0	0	0	0	
E.3.1.1	2	0	2	1	
E.3.1.2	3	1	1	0	
Totals	63	32	53	27	

Mathematics Grade 6

Anchor or Eligible Content	2011 PSSA Core	2011 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	1	1	1	1	
A.1.1.3	1	2	1	1	
A.1.1.4	1	1	1	1	
A.1.2.1	2	1	1	1	
A.1.3.1	1	1	2	0	
A.1.3.2	1	1	2	1	
A.1.3.3	1	0	1	0	
A.1.4.1	2	2	1	1	
A.2	1	0	0	1	
A.2.1.1	3	1	2	0	
A.3	0	0	1	0	
A.3.1.1	2	0	0	0	non-calculator only
A.3.2.1	2	0	1	1	
B.1	0	0	1	0	
B.1.1.1	3	1	2	1	
B.2	1	1	0	1	
B.2.1.1	0	1	1	0	
B.2.1.2	0	0	1	0	
B.2.1.3	1	0	1	0	
B.2.2.1	2	0	1	0	
B.2.3.1	1	0	1	0	
C.1	0	0	0	0	
C.1.1.1	2	0	1	1	
C.1.1.2	1	1	2	0	
C.1.1.3	2	2	1	1	
C.1.1.4	1	1	1	0	
C.1.2.1	2	0	1	1	
C.1.2.2	2	1	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	2	1	1	1	
D.1.2.1	2	2	2	1	
D.2	0	1	1	1	

G6 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	11	17%	10	31%
A.2	4	6%	1	3%
A.3	4	6%	0	0%
B.1	3	5%	1	3%
B.2	5	8%	2	6%
C.1	10	16%	5	16%
C.3	3	5%	1	3%
D.1	4	6%	3	9%
D.2	9	14%	2	6%
E.1	4	6%	4	13%
E.2	2	3%	1	3%
E.3	4	6%	2	6%

G6 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	19	30%	11	34%
B	8	13%	3	9%
C	13	21%	6	19%
D	13	21%	5	16%
E	10	16%	7	22%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.2.1.1	3	0	2	1	
D.2.1.2	2	1	3	0	
D.2.2.1	4	0	1	1	
E.1	1	0	0	0	
E.1.1.1	1	1	2	1	
E.1.1.2	1	1	1	1	
E.1.1.3	1	2	1	2	
E.2	0	0	1	0	
E.2.1.1	2	1	2	1	
E.3	0	0	0	0	
E.3.1.1	2	1	1	1	
E.3.1.2	2	1	2	0	
Totals	63	32	53	27	

Mathematics Grade 7

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	0	
A.1.1.1	3	1	1	1	
A.1.2.1	3	1	2	1	
A.1.2.2	0	0	1	0	
A.2	0	0	1	0	
A.2.1.1	2	1	1	1	
A.2.2.1	2	1	1	0	
A.2.2.2	1	1	1	0	
A.2.2.3	1	0	0	2	
A.2.2.4	0	1	0	1	
A.2.2.5	0	1	1	0	
A.2.2.6	1	0	1	0	
A.3	0	0	0	0	
A.3.1.1	2	0	0	0	non-calculator only
A.3.2.1	2	0	1	1	
A.3.2.2	0	2	1	1	
B.1	1	1	0	1	
B.1.1.1	2	0	1	0	
B.2	0	0	1	0	
B.2.1.1	1	0	1	1	
B.2.1.2	1	0	2	0	
B.2.1.3	0	1	2	0	
B.2.2.1	1	1	1	1	
B.2.2.2	1	0	0	0	
C.1	1	0	0	1	
C.1.1.1	1	2	1	1	
C.1.1.2	3	1	2	0	
C.1.1.3	0	0	1	0	
C.1.2.1	1	0	1	1	
C.1.2.2	1	1	2	0	
C.3	0	1	1	0	
C.3.1.1	2	0	1	0	
C.3.1.2	1	0	1	1	
D.1	0	0	0	0	
D.1.1.1	3	1	3	1	
D.2	1	0	1	0	

G7 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	6	10%	2	6%
A.2	7	11%	5	16%
A.3	4	6%	2	6%
B.1	3	5%	1	3%
B.2	4	6%	2	6%
C.1	7	11%	4	13%
C.3	3	5%	1	3%
D.1	3	5%	1	3%
D.2	7	11%	5	16%
D.3	6	10%	3	9%
E.1	3	5%	2	6%
E.2	2	3%	2	6%
E.3	6	10%	1	3%
E.4	2	3%	1	3%

G7 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	17	27%	9	28%
B	7	11%	3	9%
C	10	16%	5	16%
D	16	25%	9	28%
E	13	21%	6	19%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 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	2	2	1	
D.2.2.1	2	2	3	1	
D.3	0	0	1	0	
D.3.1.1	5	2	1	3	
D.3.1.2	1	1	1	1	
E.1	0	0	0	0	
E.1.1.1	3	2	1	1	
E.2	0	0	0	1	
E.2.1.1	2	2	2	0	
E.2.1.2	0	0	1	0	
E.3	0	0	0	0	
E.3.1.1	2	1	2	0	
E.3.1.2	2	0	1	1	
E.3.1.3	2	0	1	0	
E.4	0	0	1	0	
E.4.1.1	2	1	1	1	
Totals	63	32	54	27	

Mathematics Grade 8

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	0	0	0	0	
A.1.1.1	2	2	1	2	
A.1.1.2	0	0	1	0	
A.2	1	0	1	0	
A.2.1.1	1	1	1	1	
A.2.2.1	2	1	2	1	
A.2.2.2	1	1	2	0	
A.3	0	0	0	0	
A.3.1.1	1	0	1	1	
A.3.1.2	1	1	1	0	
A.3.2.1	2	0	0	0	non-calculator only
A.3.3.1	1	1	1	1	
B.1	0	0	1	0	
B.1.1.1	1	0	1	0	
B.1.1.2	1	1	1	0	
B.1.1.3	1	0	0	0	
B.1.1.4	2	0	0	1	
B.2	0	1	1	1	
B.2.1.1	1	0	1	0	
B.2.1.2	1	1	1	0	
B.2.1.3	1	0	0	0	
B.2.2.1	1	0	1	0	
B.2.2.2	1	0	1	0	
B.2.2.3	1	0	0	1	
C.1	0	0	0	1	
C.1.1.1	2	1	2	0	
C.1.1.2	2	1	2	1	
C.1.1.3	2	1	2	1	
C.1.2.1	3	1	2	0	
C.3	0	1	1	0	
C.3.1.1	3	1	1	1	
D.1	1	0	0	0	
D.1.1.1	1	0	2	1	
D.1.1.2	1	1	1	1	
D.1.1.3	2	1	1	1	
D.2	0	0	1	0	

G8 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	2	3%	2	6%
A.2	5	8%	3	9%
A.3	5	8%	2	6%
B.1	5	8%	1	3%
B.2	6	10%	2	6%
C.1	9	14%	4	13%
C.3	3	5%	2	6%
D.1	5	8%	2	6%
D.2	7	11%	5	16%
D.4	5	8%	3	9%
E.1	2	3%	2	6%
E.3	5	8%	2	6%
E.4	4	6%	2	6%

G11 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	12	19%	7	22%
B	11	17%	3	9%
C	12	19%	6	19%
D	17	27%	10	31%
E	11	17%	6	19%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
D.2.1.1	2	1	1	1	
D.2.1.2	1	1	1	1	
D.2.1.3	1	1	2	0	
D.2.2.1	2	1	1	2	
D.2.2.2	1	1	1	1	
D.4	0	0	0	0	
D.4.1.1	1	1	1	1	
D.4.1.2	1	1	1	1	
D.4.1.3	3	1	2	0	
E.1	0	0	1	0	
E.1.1.1	1	0	1	0	
E.1.1.2	0	1	1	1	
E.1.1.3	1	1	2	0	
E.3	1	0	0	1	
E.3.1.1	3	1	1	0	
E.3.2.1	1	1	1	1	
E.4	0	0	0	0	
E.4.1.1	3	1	1	0	
E.4.1.2	1	1	2	1	
Totals	63	32	54	27	

Mathematics Grade 11

Anchor or Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2009 PSSA-M FT	2010 PSSA-M FT	Comments
A.1	1	0	0	0	
A.1.1.1	1	0	1	0	
A.1.1.2	0	0	0	0	
A.1.1.3	0	1	0	1	
A.1.2.1	0	0	1	0	
A.1.3.1	0	1	1	0	
A.1.3.2	1	0	0	0	
A.2	0	0	1	0	
A.2.1.1	0	0	1	0	
A.2.1.2	0	1	1	0	
A.2.1.3	0	1	0	1	
A.2.2.1	1	1	0	1	
A.2.2.2	0	0	0	0	
A.3	0	0	0	0	
A.3.1.1	3	0	1	1	
A.3.2.1	1	0	0	0	non-calculator only
B.2	0	0	1	0	
B.2.1.1	4	1	1	1	
B.2.2.1	2	1	1	0	
B.2.2.2	1	1	0	1	
B.2.2.3	1	0	1	0	
B.2.2.4	2	2	1	1	
B.2.3.1	1	0	2	1	
C.1	0	0	1	0	
C.1.1.1	1	0	1	0	
C.1.1.2	1	1	1	0	
C.1.2.1	0	0	1	0	
C.1.2.2	0	0	0	1	
C.1.2.3	2	1	1	0	
C.1.3.1	0	0	1	0	
C.1.4.1	2	0	1	0	
C.3	1	1	0	1	
C.3.1.1	2	0	1	0	
C.3.1.2	1	0	1	1	
D.1	0	0	1	0	
D.1.1.1	1	1	1	1	

G11 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1	3	5%	2	6%
A.2	1	3%	3	9%
A.3	4	6%	0	0%
B.2	11	17%	5	16%
C.1	6	10%	2	6%
C.3	4	6%	1	3%
D.1	4	6%	1	3%
D.2	14	22%	7	22%
D.3	4	6%	6	19%
D.4	2	3%	2	6%
E.1	1	2%	2	6%
E.2	4	6%	1	3%
E.3	2	3%	0	0%
E.4	3	5%	0	0%

G11 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	8	13%	5	16%
B	11	17%	5	16%
C	10	16%	3	9%
D	24	38%	16	50%
E	10	16%	3	9%

Appendix C: 2011 PSSA-M Tally Sheets

Anchor or Eligible Content	2011 PSSA Core	2011 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	0	0	1	0	
D.2.1.2	2	0	1	1	
D.2.1.3	2	0	1	0	
D.2.1.4	2	2	2	1	
D.2.1.5	2	1	1	0	
D.2.2.1	2	2	2	0	
D.2.2.2	3	2	1	1	
D.2.2.3	0	0	1	1	
D.3	0	0	1	0	
D.3.1.1	0	1	1	1	
D.3.1.2	0	1	1	0	
D.3.2.1	2	1	1	1	
D.3.2.2	0	2	1	1	
D.3.2.3	2	1	2	0	
D.4	0	0	0	0	
D.4.1.1	2	2	2	2	
E.1	0	1	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	3	0	1	0	
E.2.1.2	1	1	1	1	
E.2.1.3	0	0	0	0	
E.3	0	0	0	1	
E.3.1.1	1	0	0	0	
E.3.1.2	1	0	1	1	
E.3.2.1	0	0	1	0	
E.4	0	0	0	0	
E.4.1.1	0	0	1	0	
E.4.1.2	2	0	0	0	
E.4.2.1	0	0	1	0	
E.4.2.2	1	0	1	0	
Totals	63	32	54	27	

Reading Grade 4

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	0	1	1
A.1.1.2	2	1	6
A.1.2.1	2	1	3
A.1.2.2	0	0	3
A.1.3.1	8	3	21
A.1.4.1	4	6	27
A.1.5.1	1	1	3
A.1.6.1	2	2	4
A.2.1.1	0	0	1
A.2.1.2	0	0	0
A.2.2.1	1	0	3
A.2.2.2	1	1	3
A.2.3.1	4	2	12
A.2.4.1	4	4	34
A.2.5.1	1	0	1
A.2.6.1	1	1	5
B.1.1.1	6	5	27
B.1.2.1	0	0	0
B.2.1.1	1	2	3
B.2.1.2	0	0	6
B.2.1.3	1	0	3
B.3.1.1	1	0	3
B.3.2.1	0	0	1
B.3.3.1	1	1	5
B.3.3.2	0	1	4
B.3.3.3	3	0	0
B.3.3.4	0	0	0
Totals	44	32	179

G4 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	2	4%	2	6%
A.1.2	2	4%	1	3%
A.1.3	10	19%	3	8%
A.1.4	4	8%	8	22%
A.1.5	1	2%	1	3%
A.1.6	2	4%	2	6%
A.2.1	0	0%	0	0%
A.2.2	2	4%	1	3%
A.2.3	4	8%	2	6%
A.2.4	6	12%	4	11%
A.2.5	1	2%	0	0%
A.2.6	1	2%	1	3%
B.1.1	10	19%	7	19%
B.1.2	0	0%	0	0%
B.2.1	2	4%	2	6%
B.3.1	1	2%	0	0%
B.3.2	0	0%	0	0%
B.3.3	4	8%	2	6%

G4 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	35	67%	25	69%
B	17	33%	11	31%

Reading Grade 5

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	0	1	2
A.1.1.2	1	1	5
A.1.2.1	1	0	1
A.1.2.2	0	0	4
A.1.3.1	6	2	18
A.1.3.2	0	0	1
A.1.4.1	6	7	13
A.1.5.1	1	0	2
A.1.6.1	1	1	5
A.1.6.2	0	0	0
A.2.1.1	0	0	4
A.2.1.2	1	1	1
A.2.2.1	0	0	5
A.2.2.2	0	0	5
A.2.3.1	5	3	16
A.2.3.2	1	0	3
A.2.4.1	3	5	27
A.2.5.1	0	1	2
A.2.6.1	0	1	5
A.2.6.2	0	0	1
B.1.1.1	6	4	27
B.1.2.1	0	0	9
B.2.1.1	0	1	2
B.2.1.2	2	1	5
B.2.1.3	2	0	1
B.2.1.4	1	0	2
B.2.2.1	1	1	4
B.2.2.2	0	0	1
B.3.1.1	2	2	5
B.3.2.1	0	0	1
B.3.2.2	0	0	0
B.3.3.1	1	0	1
B.3.3.2	0	0	1
B.3.3.3	1	0	0
B.3.3.4	2	0	2
Totals	44	32	181

G5 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	1	2%	2	6%
A.1.2	1	2%	0	0%
A.1.3	6	12%	2	6%
A.1.4	6	12%	7	19%
A.1.5	3	6%	0	0%
A.1.6	1	2%	1	3%
A.2.1	1	2%	1	3%
A.2.2	0	0%	0	0%
A.2.3	10	19%	3	8%
A.2.4	3	6%	5	14%
A.2.5	0	0%	3	8%
A.2.6	0	0%	1	3%
B.1.1	8	15%	6	17%
B.1.2	0	0%	0	0%
B.2.1	5	10%	2	6%
B.2.2	1	2%	1	3%
B.3.1	2	4%	2	6%
B.3.2	0	0%	0	0%
B.3.3	4	8%	0	0%

G5 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	32	62%	25	69%
B	20	38%	11	31%

Reading Grade 6

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	1	0	1
A.1.1.2	1	1	3
A.1.2.1	0	1	4
A.1.2.2	1	0	3
A.1.3.1	5	2	15
A.1.3.2	0	0	0
A.1.4.1	2	1	16
A.1.5.1	1	0	1
A.1.6.1	1	1	4
A.1.6.2	0	0	0
A.2.1.1	0	0	0
A.2.1.2	0	0	1
A.2.2.1	1	0	1
A.2.2.2	1	3	7
A.2.3.1	2	2	9
A.2.3.2	2	1	2
A.2.4.1	8	6	31
A.2.5.1	1	0	2
A.2.6.1	1	0	2
A.2.6.2	0	0	1
B.1.1.1	4	5	29
B.1.2.1	3	1	3
B.2.1.1	0	0	2
B.2.1.2	2	2	7
B.2.1.3	0	1	1
B.2.1.4	2	0	1
B.2.2.1	0	0	1
B.2.2.2	0	0	1
B.3.1.1	0	3	8
B.3.2.1	1	1	1
B.3.2.2	0	0	0
B.3.3.1	0	0	5
B.3.3.2	1	1	4
B.3.3.3	3	0	1
B.3.3.4	0	0	0
Totals	44	32	167

G6 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	2	4%	1	3%
A.1.2	1	2%	1	3%
A.1.3	5	10%	2	6%
A.1.4	2	4%	1	3%
A.1.5	1	2%	0	0%
A.1.6	1	2%	1	3%
A.2.1	0	0%	0	0%
A.2.2	2	4%	3	8%
A.2.3	8	15%	3	8%
A.2.4	8	15%	6	17%
A.2.5	1	2%	0	0%
A.2.6	1	2%	0	0%
B.1.1	6	12%	9	25%
B.1.2	5	10%	1	3%
B.2.1	4	8%	3	8%
B.2.2	0	0%	0	0%
B.3.1	0	0%	3	8%
B.3.2	1	2%	1	3%
B.3.3	4	8%	1	3%

G6 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	32	62%	18	50%
B	20	38%	18	50%

Reading Grade 7

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	0	0	2
A.1.1.2	2	1	2
A.1.2.1	1	0	1
A.1.2.2	1	0	0
A.1.3.1	4	4	13
A.1.3.2	0	0	0
A.1.4.1	1	1	8
A.1.5.1	0	1	2
A.1.6.1	1	0	4
A.1.6.2	0	0	0
A.2.1.1	1	1	5
A.2.1.2	0	0	1
A.2.2.1	1	1	4
A.2.2.2	1	0	5
A.2.3.1	5	0	17
A.2.3.2	1	0	2
A.2.4.1	3	5	29
A.2.5.1	2	1	4
A.2.6.1	1	1	5
A.2.6.2	0	0	3
B.1.1.1	6	9	32
B.1.2.1	2	0	5
B.2.1.1	4	2	14
B.2.1.2	2	1	2
B.2.2.1	0	1	3
B.2.2.2	1	0	0
B.3.1.1	1	0	2
B.3.2.1	1	0	2
B.3.3.1	1	1	4
B.3.3.2	1	0	2
B.3.3.3	0	1	4
B.3.3.4	0	1	2
Totals	44	32	179

G7 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	2	4%	1	3%
A.1.2	2	4%	0	0%
A.1.3	6	12%	4	11%
A.1.4	1	2%	1	3%
A.1.5	0	0%	1	3%
A.1.6	1	2%	0	0%
A.2.1	1	2%	1	3%
A.2.2	2	4%	1	3%
A.2.3	8	15%	0	0%
A.2.4	3	6%	5	14%
A.2.5	4	8%	3	8%
A.2.6	1	2%	1	3%
B.1.1	6	12%	11	31%
B.1.2	4	8%	0	0%
B.2.1	6	12%	3	8%
B.2.2	1	2%	1	3%
B.3.1	1	2%	0	0%
B.3.2	1	2%	0	0%
B.3.3	2	4%	3	8%

G7 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	31	60%	18	50%
B	21	40%	18	50%

Reading Grade 8

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	1	0	2
A.1.1.2	0	1	6
A.1.2.1	0	0	2
A.1.2.2	2	1	2
A.1.3.1	4	2	12
A.1.3.2	0	0	1
A.1.4.1	3	2	9
A.1.5.1	2	0	3
A.1.6.1	1	0	4
A.1.6.2	1	0	0
A.2.1.1	1	0	1
A.2.1.2	0	2	5
A.2.2.1	1	1	1
A.2.2.2	1	0	3
A.2.3.1	3	3	11
A.2.3.2	0	1	2
A.2.4.1	2	3	17
A.2.5.1	0	0	1
A.2.6.1	0	1	3
A.2.6.2	1	0	1
B.1.1.1	5	4	23
B.1.2.1	5	2	6
B.2.1.1	2	2	12
B.2.1.2	1	2	3
B.2.2.1	1	0	4
B.2.2.2	1	1	1
B.3.1.1	1	2	7
B.3.2.1	0	0	4
B.3.3.1	1	1	5
B.3.3.2	0	0	3
B.3.3.3	2	0	0
B.3.3.4	2	1	2
Totals	44	32	156

G8 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	1	2%	1	3%
A.1.2	2	4%	1	3%
A.1.3	6	12%	2	6%
A.1.4	3	6%	2	6%
A.1.5	2	4%	0	0%
A.1.6	2	4%	0	0%
A.2.1	1	2%	2	6%
A.2.2	2	4%	1	3%
A.2.3	3	6%	6	17%
A.2.4	2	4%	3	8%
A.2.5	0	0%	0	0%
A.2.6	1	2%	1	3%
B.1.1	7	13%	6	17%
B.1.2	9	17%	2	6%
B.2.1	3	6%	4	11%
B.2.2	2	4%	1	3%
B.3.1	1	2%	2	6%
B.3.2	0	0%	0	0%
B.3.3	5	10%	2	6%

G8 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	25	48%	19	53%
B	27	52%	17	47%

Reading Grade 11

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	1	1	1
A.1.1.2	0	0	6
A.1.2.1	1	0	0
A.1.2.2	0	0	4
A.1.3.1	2	2	13
A.1.3.2	0	0	1
A.1.4.1	2	2	8
A.1.5.1	0	0	0
A.1.6.1	0	0	2
A.1.6.2	0	0	0
A.2.1.1	0	0	0
A.2.1.2	1	0	4
A.2.2.1	1	2	4
A.2.2.2	2	2	7
A.2.3.1	5	2	14
A.2.3.2	1	0	2
A.2.4.1	4	5	30
A.2.5.1	0	0	1
A.2.6.1	0	1	3
A.2.6.2	2	0	2
B.1.1.1	3	3	23
B.1.2.1	6	2	6
B.2.1.1	2	3	12
B.2.1.2	2	2	5
B.2.2.1	1	0	2
B.2.2.2	0	0	0
B.3.1.1	2	2	8
B.3.2.1	3	0	2
B.3.2.2	0	1	1
B.3.3.1	0	1	2
B.3.3.2	3	0	3
B.3.3.3	0	1	3
B.3.3.4	0	0	1
Totals	44	32	170

G11 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	1	2%	1	3%
A.1.2	1	2%	0	0%
A.1.3	2	4%	2	6%
A.1.4	2	4%	2	6%
A.1.5	0	0%	0	0%
A.1.6	0	0%	0	0%
A.2.1	1	2%	0	0%
A.2.2	3	6%	4	11%
A.2.3	6	12%	4	11%
A.2.4	4	8%	5	14%
A.2.5	0	0%	0	0%
A.2.6	2	4%	1	3%
B.1.1	5	10%	5	14%
B.1.2	10	19%	2	6%
B.2.1	6	12%	5	14%
B.2.2	1	2%	0	0%
B.3.1	2	4%	2	6%
B.3.2	3	6%	1	3%
B.3.3	3	6%	2	6%

G11 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	22	42%	19	53%
B	30	58%	17	47%

Science Grade 8

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	0	1	1
A.1.1.2	1	2	2
A.1.1.3	0	0	2
A.1.1.4	2	0	1
A.1.2.1	1	0	3
A.1.2.2	0	0	2
A.1.2.3	2	1	2
A.1.2.4	1	0	1
A.1.3.1	1	0	2
A.1.3.2	2	1	7
A.1.3.3	1	0	2
A.1.3.4	1	1	2
A.2.1.1	1	0	2
A.2.1.2	0	0	1
A.2.1.3	1	0	1
A.2.1.4	2	0	0
A.2.1.5	2	1	1
A.2.1.6	1	0	0
A.2.2.1	2	2	7
A.2.2.2	1	0	3
A.2.2.3	1	1	3
A.3.1.1	1	0	1
A.3.1.2	1	2	5
A.3.1.3	1	2	2
A.3.1.4	2	0	3
A.3.1.5	0	0	0
A.3.2.1	0	0	3
A.3.2.2	1	1	1
A.3.2.3	1	1	3
A.3.3.1	1	0	3
A.3.3.2	1	0	1
B.1.1.1	0	1	2
B.1.1.2	1	2	5
B.1.1.3	1	0	3
B.1.1.4	1	0	1
B.2.1.1	0	0	1
B.2.1.2	0	0	0
B.2.1.3	0	0	1
B.2.1.4	1	1	1

G8 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	3	4%	3	9%
A.1.2	5	7%	1	3%
A.1.3	6	9%	3	9%
A.2.1	7	10%	1	3%
A.2.2	4	6%	3	9%
A.3.1	5	7%	4	12%
A.3.2	2	3%	2	6%
A.3.3	2	3%	0	0%
B.1.1	3	4%	3	9%
B.2.1	1	1%	1	3%
B.2.2	2	3%	3	9%
B.3.1	3	4%	0	0%
B.3.2	0	0%	0	0%
B.3.3	2	3%	0	0%
C.1.1	4	6%	1	3%
C.2.1	2	3%	2	6%
C.2.2	3	4%	1	3%
C.3.1	2	3%	1	3%
D.1.1	1	1%	0	0%
D.1.2	2	3%	1	3%
D.1.3	5	7%	2	6%
D.2.1	0	0%	0	0%
D.3.1	4	6%	2	6%

G8 Reporting Category Summary

	2011 PSSA Core		2011 PSSA-M Core	
A	34	50%	17	50%
B	11	16%	7	21%
C	11	16%	5	15%
D	12	18%	5	15%

Appendix C: 2011 PSSA-M Tally Sheets

B.2.1.5	0	0	0
B.2.2.1	1	1	2
B.2.2.2	1	1	1
B.3.1.1	0	0	1
B.3.1.2	0	0	0
B.3.1.3	2	0	1
B.3.2.1	0	0	0
B.3.2.2	0	0	0
B.3.2.3	0	0	1
B.3.3.1	0	0	0
B.3.3.2	1	0	2
B.3.3.3	1	0	1
B.3.3.4	0	0	0
C.1.1.1	1	0	1
C.1.1.2	1	1	2
C.1.1.3	2	0	3
C.2.1.1	1	0	2
C.2.1.2	1	1	3
C.2.1.3	0	1	1
C.2.2.1	1	1	1
C.2.2.2	1	0	2
C.2.2.3	1	0	1
C.3.1.1	1	0	3
C.3.1.2	0	0	1
C.3.1.3	1	1	2
D.1.1.1	0	0	2
D.1.1.2	0	0	2
D.1.1.3	1	0	0
D.1.1.4	0	0	1
D.1.2.1	1	1	1
D.1.2.2	1	0	0
D.1.3.1	1	1	4
D.1.3.2	2	1	2
D.1.3.3	1	0	2
D.1.3.4	0	0	0
D.2.1.1	0	0	1
D.2.1.2	0	0	2
D.2.1.3	0	0	1
D.3.1.1	0	1	2
D.3.1.2	1	1	2
D.3.1.3	2	0	1
Totals	63	32	135

Science Grade 11

Eligible Content	2011 PSSA Core	2011 PSSA-M Core	2010 PSSA-M FT
A.1.1.1	1	0	3
A.1.1.2	2	1	5
A.1.1.3	1	1	7
A.1.1.4	1	0	2
A.1.1.5	1	1	5
A.1.2.1	1	1	3
A.1.2.2	0	0	1
A.1.3.1	2	1	4
A.1.3.2	1	1	4
A.1.3.3	1	1	2
A.1.3.4	0	1	2
A.2.1.1	2	0	1
A.2.1.2	0	0	1
A.2.1.3	1	0	3
A.2.1.4	0	1	3
A.2.1.5	2	0	3
A.2.2.1	2	2	5
A.2.2.2	5	0	0
A.3.1.1	1	1	2
A.3.1.2	1	2	2
A.3.1.3	0	1	2
A.3.1.4	1	0	1
A.3.2.1	1	0	0
A.3.2.2	1	0	0
A.3.2.3	0	1	1
A.3.3.1	1	0	1
A.3.3.2	0	0	0
A.3.3.3	2	0	4
B.1.1.1	1	0	0
B.1.1.2	0	0	0
B.1.1.3	1	0	3
B.2.1.1	0	0	0
B.2.1.2	1	0	1
B.2.1.3	0	0	0
B.2.1.4	1	0	0
B.2.2.1	0	0	0
B.2.2.2	0	0	2
B.2.2.3	2	0	2

G11 Anchor Summary

	2011 PSSA Core		2011 PSSA-M Core	
A.1.1	9	12%	3	9%
A.1.2	1	1%	1	3%
A.1.3	5	7%	4	12%
A.2.1	7	9%	1	3%
A.2.2	7	9%	2	6%
A.3.1	3	4%	5	15%
A.3.2	3	4%	1	3%
A.3.3	3	4%	0	0%
B.1.1	3	4%	0	0%
B.2.1	2	3%	0	0%
B.2.2	2	3%	0	0%
B.3.1	3	4%	3	9%
B.3.2	1	1%	2	6%
B.3.3	1	1%	1	3%
C.1.1	8	11%	2	6%
C.2.1	2	3%	0	0%
C.2.2	0	0%	3	9%
C.3.1	4	5%	1	3%
D.1.1	3	4%	2	6%
D.1.2	1	1%	0	0%
D.1.3	2	3%	1	3%
D.2.1	2	3%	0	0%
D.3.1	2	3%	2	6%

	2011 PSSA Core		2011 PSSA-M Core	
A	38	51%	17	50%
B	12	16%	6	18%
C	14	19%	6	18%
D	10	14%	5	15%

Appendix C: 2011 PSSA-M Tally Sheets

B.3.1.1	0	0	1
B.3.1.2	1	1	4
B.3.1.3	1	1	1
B.3.1.4	0	1	1
B.3.1.5	0	0	0
B.3.2.1	0	0	0
B.3.2.2	1	1	2
B.3.2.3	0	1	2
B.3.3.1	1	0	2
B.3.3.2	0	0	0
B.3.3.3	0	1	1
C.1.1.1	1	1	3
C.1.1.2	1	0	0
C.1.1.3	0	0	3
C.1.1.4	0	0	0
C.1.1.5	1	1	3
C.1.1.6	3	0	1
C.2.1.1	1	0	2
C.2.1.2	0	0	1
C.2.1.3	0	0	1
C.2.1.4	0	0	1
C.2.2.1	0	0	0
C.2.2.2	0	1	1
C.2.2.3	0	1	2
C.3.1.1	3	0	0
C.3.1.2	0	1	1
C.3.1.3	0	0	3
C.3.1.4	0	0	0
C.3.1.5	1	0	1
C.3.1.6	0	0	0
D.1.1.1	1	0	2
D.1.1.2	2	1	3
D.1.1.3	0	1	2
D.1.2.1	0	0	0
D.1.2.2	1	0	1
D.1.3.1	0	0	1
D.1.3.2	2	1	2
D.1.3.3	0	0	1

Appendix C: 2011 PSSA-M Tally Sheets

D.2.1.1	2	0	1
D.2.1.2	0	0	1
D.2.1.3	0	0	1
D.2.1.4	0	0	2
D.3.1.1	0	0	1
D.3.1.2	2	2	3
D.3.1.3	0	0	2
Totals	62	32	135

Appendix D:
Item and Test Development Process

Item and Test Development Process for the PSSA-M (for 2010-2011 only)

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 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, and 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	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 (for 2010-2011 only)

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 (<i>Item Development and Educational Assessment System</i>). 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 (<i>Item Development and Educational Assessment System</i>) (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 (for 2010-2011 only)

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, presents 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 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 or Embedded Field Test Forms	DRC item and test development specialists select test items for inclusion on the mathematics embedded field test forms and the reading and science 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 mathematics items are included in three embedded field test forms per grade. The reading items are arranged in six unique standalone field test forms for each grade. The science items are arranged in five unique standalone field test forms 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 (for 2010-2011 only)

Step	Description
19. Review Results of the Field Test	Following the administration of a field test form and the subsequent ranging 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. Psychometricians 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 <i>Item Development and Educational Assessment System, IDEAS</i> .
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 (for 2010-2011 only)

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.
25. Construct Test Forms	Items and test components are assembled into forms using the form construction and typesetting function of DRC's <i>Item Development and Educational Assessment System</i> , IDEAS. Forms are reviewed internally for style and formatting requirements.
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.
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.

Guidelines for Item Revision and Enhancement

Overview

The PSSA-M is 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) are 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 reading and science field tests (spring 2010), PSSA-M assessment items, tasks, etc. were revised and/or enhanced to maintain the integrity of the grade-level content; however, revisions and/or enhancements were purposefully and necessarily made in the operationalization of the grade-level content in order to address the specific access needs of the students who are eligible for the PSSA-M.

Three main areas of consideration 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 are 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 are 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 are 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 relied 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 is supplemented with PDE’s analyses of the 2010 student performance data (e.g., p-values, point biserials, and omission rates). In addition, Cognitive Interviews were also conducted prior to the spring 2010 field test.

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 (For example, students are provided the same passage/prompt 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 or prompt. 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 or prompt
 - 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 include, but are not 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 construction.
- The verb should follow the subject (subject and verb should be adjacent to each other)—use common construction.

Appendix D: Item and Test Development Process

- 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 that 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).

Appendix E:
PSSA-M Item Review Cards

Appendix E: PSSA-M Item Review Cards

<p>1. A bag has 10 marbles in it. The marbles are described below:</p> <ul style="list-style-type: none"> • 3 green marbles • 3 blue marbles • 4 white marbles <p>Luci selects 1 marble from the bag without looking.</p> <p>What is the probability Luci selects a white marble?</p> <p><input type="radio"/> $\frac{1}{10}$</p> <p><input type="radio"/> $\frac{1}{4}$</p> <p><input type="radio"/> $\frac{4}{10}$</p> <p><input type="radio"/> $\frac{4}{6}$</p>	<p>PSSA-M Item Card</p> <p>Item ID</p> <p>Content Area</p> <p>Mathematics</p> <p>Passage ID</p> <p>Passage Title</p> <p>Grade</p> <p>5</p> <p>AACS Standards</p> <p>E.3.1.2</p> <p>Item Type</p> <p>Multiple Choice</p> <p>Points</p> <p>1</p> <p>Depth of Knowledge</p> <p>2</p> <p>Est Difficulty</p> <p>Low</p> <p>Key</p> <p>C</p> <p>Calculator</p> <p>C</p> <p>Focus</p> <p>Probability</p>
--	--

Appendix E: PSSA-M Item Review Cards

<p>Kelly has a toy animal collection. Kelly's toy animal collection is larger than Ann's by more than 20 animals.</p> <p>6. Write an inequality that can be used to show the number of animals a in Kelly's collection.</p> <div data-bbox="293 422 1268 527" style="border: 1px solid black; padding: 5px;"><p>Inequality: _____</p></div>	
--	--

Appendix E: PSSA-M Item Review Cards

<p>1. Simplify</p> $\frac{15 - 2^2 + 10 - 20}{2}$ <p>(Hint: Remember to use order of operations.)</p> <p> <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 10 <input type="radio"/> 12 </p>	<p>PSSA-M Data Card</p> <p>Item ID</p> <p>Content Area</p> <p>Mathematics</p> <p>Passage ID</p> <p>Passage Title</p> <p>Grade</p> <p>8</p> <p>Standards</p> <p>AACS: A.2.1.1</p> <p>Item Type</p> <p>Multiple Choice</p> <p>Points</p> <p>1</p> <p>Depth of Knowledge</p> <p>1</p> <p>Est Difficulty</p> <p>Low</p> <p>Calculator</p> <p>Yes</p> <p>Key</p> <p>A</p> <p>Focus</p> <p>Order of operations</p>
---	---

Appendix E: PSSA-M Item Review Cards

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

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: <ol style="list-style-type: none"> 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

Appendix F: Item Rating Sheet and Item Review Criteria Guidelines

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? <i>A thorough, independent bias review</i> (separate from this meeting) <i>will be completed for all items.</i>	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

Appendix F: Item Rating Sheet and Item Review Criteria Guidelines

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:
2011 Test Book Section Layout Plans

2011 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 100-120 minutes. [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	1	1–common (core) item	50–60
2	Mathematics	15	15–common (core) items	1	1–common (core) item	50–60

Notes:

- 1) There is 1 form.
- 2) The ruler items may fall in Sections 1 or 2.
- 3) All items in the PSSA-M mathematics test allow for calculator use.

2011 Modified Reading Test Book Section Layout for Grades 4, 5, 7, 8 & 11

Reading Core

Core/common MC items	30
2 core 3 pt OE items	6
Total	36 points

The estimated testing time for reading is approximately 100-120 minutes. [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	Reading	18	18–common (core) items	1	1–common (core) item	60–70
2	Reading	12	12–common (core) items	1	1–common (core) item	40–50

2011 Modified Reading Test Book Section Layout for Grade 6

Reading Core

Core/common MC items	30
2 core 3 pt OE items	6
Total	36 points

The estimated testing time for reading is approximately 100-120 minutes. [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	Reading	19	19–common (core) items	1	1–common (core) item	60–70
2	Reading	11	11–common (core) items	1	1–common (core) item	40–50

2011 Modified Science Test Book Section Layout for Grades 8 & 11

Science Core

Core/common MC items	30
2 core 2 pt OE items	4
Total	34 points

The estimated testing time for science is approximately 60-80 minutes. [Timing assumes 5 min per OE and 2 min per MC.]

Section	Content	Number of MC	MC Item Breakdown	Number of OE	OE Item Breakdown	Section Time (in minutes)
1	Science	15	15–common (core) items	1	1–common (core) item	30–40
2	Science	15	15–common (core) items	1	1–common (core) item	30–40

Appendix H:
Mean Raw Scores by Form

Column Heading	Definition
Form	Form
N	N students
L	Length
Pts	Points possible
Min	Minimum
Max	Maximum
Mean	Mean
Med	Median
<i>SD</i>	Standard deviation

Appendix H: Mean Raw Scores by Form

		Form	N	L	Pts	Min	Max	Mean	Med	SD
Mathematics	4	0	2375	32	38	2	38	22.7	23.0	6.95
	5	0	3366	32	38	3	38	21.2	22.0	7.12
	6	0	3600	32	38	2	37	17.6	17.0	6.54
	7	0	3972	32	38	4	38	19.9	20.0	6.74
	8	0	4114	32	38	1	38	18.5	18.0	6.87
	11	0	4269	32	38	1	36	16.4	15.0	7.28
Reading	4	0	3388	32	36	2	36	19.2	19.0	6.87
	5	0	3947	32	36	2	36	20.5	21.0	7.19
	6	0	3983	32	36	3	36	21.8	23.0	6.98
	7	0	3974	32	36	3	35	20.1	21.0	6.54
	8	0	3647	32	36	2	36	21.3	22.0	6.46
	11	0	3919	32	36	3	35	22.9	24.0	6.34
Sci.	8	0	3252	32	34	4	34	20.9	21.0	5.82
	11	0	3540	32	34	2	34	17.8	18.0	5.99

Appendix I:

Item Statistics

Column Heading	Definition
PubID	Public ID
Std	Standard
DOK	Depth of knowledge
N	N
PVal	P-Value
P()	Proportion selecting given response (-=blank)
PtBis	Point biserial
PT()	Point biserial of repsonse
Meas	Rasch item measure
MeasSE	Rasch item measure standard error
t	t fit statistic
MS	Mean square fit statistic

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Math	4	0276	A.1.3.1	1	2375	0.46	0.09	0.38	0.08	0.46	0.00	0.29	-0.21	-0.30	-0.19	0.29	1.1078	0.0450	2.3	1.0	2.7	1.1
Math	4	0413	C.1.2.2	1	2375	0.74	0.03	0.08	0.74	0.15	0.00	0.30	-0.26	-0.24	0.30	-0.19	-0.2913	0.0502	0.3	1.0	0.2	1.0
Math	4	0799	B.1.1.1	1	2375	0.65	0.09	0.20	0.06	0.65	0.00	0.33	-0.35	-0.23	-0.22	0.33	0.2074	0.0466	0.0	1.0	-0.5	1.0
Math	4	1730	A.1.2.2	1	2375	0.56	0.24	0.07	0.12	0.56	0.00	0.30	-0.26	-0.25	-0.24	0.30	0.6300	0.0451	1.9	1.0	1.2	1.0
Math	4	1814	A.3.1.1	1	2375	0.47	0.22	0.47	0.19	0.12	0.00	0.22	-0.21	0.22	-0.16	-0.20	1.0499	0.0450	6.5	1.1	7.7	1.2
Math	4	1986	E.1.1.1	1	2375	0.87	0.03	0.04	0.87	0.06	0.00	0.37	-0.23	-0.24	0.37	-0.26	-1.2888	0.0642	-3.0	0.9	-4.1	0.7
Math	4	2333	A.1.1.3	1	2375	0.35	0.09	0.35	0.21	0.34	0.00	0.29	-0.21	0.29	-0.17	-0.39	1.6139	0.0467	1.5	1.0	3.0	1.1
Math	4	2449	E.1.2.2	1	2375	0.89	0.05	0.04	0.02	0.89	0.00	0.38	-0.27	-0.23	-0.23	0.38	-1.4177	0.0669	-3.4	0.9	-4.1	0.7
Math	4	2842	A.3.2.2	1	2375	0.85	0.03	0.04	0.07	0.85	0.00	0.38	-0.24	-0.22	-0.29	0.38	-1.1094	0.0609	-3.1	0.9	-4.0	0.8
Math	4	3063	A.3.1.2	1	2375	0.55	0.11	0.28	0.55	0.06	0.00	0.36	-0.30	-0.31	0.36	-0.35	0.6860	0.0450	-1.4	1.0	-1.4	1.0
Math	4	4042	D.1.2.2	2	2375	0.52	0.28	0.52	0.10	0.10	0.00	0.27	-0.19	0.27	-0.28	-0.26	0.8293	0.0449	3.9	1.1	4.0	1.1
Math	4	4529	D.2.2.1	1	2375	0.81	0.05	0.81	0.04	0.10	0.00	0.39	-0.23	0.39	-0.25	-0.31	-0.7799	0.0558	-3.3	0.9	-3.7	0.8
Math	4	4538	E.3.1.1	2	2375	0.84	0.84	0.04	0.05	0.06	0.00	0.29	0.29	-0.20	-0.25	-0.14	-1.0300	0.0596	-1.0	1.0	-1.3	0.9
Math	4	4652	B.2.2.1	1	2375	0.41	0.14	0.41	0.30	0.14	0.00	0.24	-0.40	0.24	-0.13	-0.22	1.3241	0.0455	4.9	1.1	5.2	1.1
Math	4	5318	A.1.1.4	1	2375	0.47	0.27	0.15	0.47	0.11	0.00	0.36	-0.38	-0.26	0.36	-0.33	1.0519	0.0450	-1.7	1.0	-1.1	1.0
Math	4	5359	B.2.2.1	1	2375	0.54	0.21	0.18	0.07	0.54	0.00	0.33	-0.22	-0.41	-0.18	0.33	0.7160	0.0450	0.6	1.0	0.8	1.0
Math	4	5771	D.1.1.3	2	2375	0.35	0.20	0.31	0.13	0.35	0.00	0.30	-0.29	-0.30	-0.29	0.30	1.6465	0.0469	0.5	1.0	2.5	1.1
Math	4	5854	E.1.2.1	1	2375	0.87	0.06	0.87	0.03	0.04	0.00	0.43	-0.36	0.43	-0.20	-0.27	-1.2201	0.0629	-4.4	0.8	-3.8	0.8
Math	4	5889	E.3.1.1	2	2375	0.72	0.19	0.04	0.72	0.04	0.00	0.40	-0.37	-0.21	0.40	-0.24	-0.2130	0.0495	-3.6	0.9	-3.3	0.9
Math	4	6225	E.1.1.1	2	2375	0.73	0.09	0.10	0.73	0.08	0.00	0.41	-0.29	-0.30	0.41	-0.32	-0.2495	0.0498	-3.7	0.9	-4.2	0.8
Math	4	6461	A.1.3.2	1	2375	0.40	0.31	0.10	0.40	0.19	0.00	0.40	-0.39	-0.32	0.40	-0.41	1.3879	0.0457	-4.0	0.9	-2.1	1.0
Math	4	6745	A.1.1.1	1	2375	0.53	0.13	0.30	0.53	0.04	0.00	0.39	-0.31	-0.39	0.39	-0.18	0.7538	0.0449	-3.0	1.0	-2.9	0.9
Math	4	6890	B.1.1.4	2	2375	0.45	0.20	0.45	0.18	0.17	0.00	0.28	-0.27	0.28	-0.27	-0.24	1.1659	0.0451	2.6	1.0	4.5	1.1
Math	4	7504	A.2.1.2	2	2375	0.73	0.07	0.73	0.05	0.15	0.00	0.38	-0.30	0.38	-0.30	-0.26	-0.2519	0.0498	-2.3	1.0	-3.3	0.9
Math	4	7572	D.2.1.1	1	2375	0.34	0.19	0.08	0.38	0.34	0.00	0.18	-0.16	-0.25	-0.18	0.18	1.6706	0.0470	6.0	1.1	6.8	1.2
Math	4	8440	A.1.1.2	1	2375	0.73	0.25	0.73	0.01	0.02	0.00	0.36	-0.33	0.36	-0.18	-0.20	-0.2203	0.0495	-1.8	1.0	-1.7	0.9
Math	4	8571	B.1.1.2	1	2375	0.50	0.28	0.50	0.13	0.09	0.00	0.25	-0.17	0.25	-0.28	-0.29	0.9266	0.0449	5.0	1.1	4.9	1.1
Math	4	8876	A.1.2.1	1	2375	0.52	0.11	0.13	0.23	0.52	0.00	0.28	-0.29	-0.23	-0.23	0.28	0.8015	0.0449	3.2	1.1	2.2	1.1
Math	4	9101	C.2.1.1	1	2375	0.83	0.83	0.06	0.06	0.05	0.00	0.28	0.28	-0.20	-0.14	-0.21	-0.9540	0.0583	-0.6	1.0	-0.7	1.0
Math	4	9590	D.1.1.2	2	2375	0.67	0.67	0.09	0.07	0.17	0.00	0.38	0.38	-0.37	-0.22	-0.27	0.0638	0.0474	-2.2	1.0	-3.1	0.9
Math	5	0364	C.1.1.1		3366	0.52	0.23	0.52	0.21	0.04	0.00	0.20	-0.23	0.20	-0.11	-0.17	0.6183	0.0376	9.5	1.1	9.0	1.2
Math	5	0895	A.3.1.1		3366	0.48	0.48	0.24	0.16	0.12	0.00	0.31	0.31	-0.34	-0.18	-0.27	0.7821	0.0377	2.1	1.0	2.2	1.0
Math	5	0973	C.2.1.2		3366	0.63	0.06	0.63	0.08	0.23	0.00	0.39	-0.16	0.39	-0.23	-0.40	0.0538	0.0388	-3.7	1.0	-3.0	0.9
Math	5	1317	A.1.4.2	1	3366	0.52	0.19	0.03	0.52	0.27	0.00	0.27	-0.30	-0.18	0.27	-0.20	0.6211	0.0376	4.9	1.1	4.0	1.1
Math	5	2129	C.2.1.1		3366	0.65	0.65	0.09	0.19	0.07	0.00	0.32	0.32	-0.24	-0.25	-0.26	-0.0250	0.0391	0.1	1.0	0.8	1.0
Math	5	2205	C.1.2.1	1	3366	0.60	0.22	0.03	0.16	0.60	0.00	0.26	-0.26	-0.17	-0.17	0.26	0.2224	0.0383	4.5	1.1	4.3	1.1
Math	5	2759	A.1.4.1		3366	0.75	0.75	0.15	0.03	0.06	0.00	0.35	0.35	-0.30	-0.21	-0.22	-0.5833	0.0426	-2.4	1.0	-3.5	0.9
Math	5	2966	E.3.1.2		3366	0.55	0.21	0.17	0.55	0.07	0.00	0.31	-0.27	-0.23	0.31	-0.25	0.4638	0.0378	2.1	1.0	1.5	1.0
Math	5	4012	E.2.1.1	1	3366	0.34	0.12	0.37	0.34	0.17	0.00	0.32	-0.40	-0.23	0.32	-0.45	1.4588	0.0395	0.0	1.0	1.0	1.0
Math	5	4397	C.2.1.2		3366	0.54	0.06	0.54	0.22	0.17	0.00	0.37	-0.17	0.37	-0.29	-0.41	0.4862	0.0377	-1.9	1.0	-1.9	1.0
Math	5	4421	D.2.1.2	1	3366	0.60	0.23	0.60	0.12	0.05	0.00	0.33	-0.30	0.33	-0.26	-0.21	0.2266	0.0382	0.0	1.0	0.3	1.0
Math	5	5504	D.2.1.1		3366	0.79	0.79	0.06	0.06	0.09	0.00	0.41	0.41	-0.20	-0.33	-0.31	-0.8301	0.0449	-5.1	0.9	-6.0	0.8

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Math	5	6385	A.2.1.2		3366	0.44	0.06	0.10	0.40	0.44	0.00	0.25	-0.30	-0.16	-0.23	0.25	0.9736	0.0379	5.1	1.1	5.0	1.1
Math	5	6562	A.1.6.2	1	3366	0.60	0.27	0.06	0.07	0.60	0.00	0.43	-0.45	-0.23	-0.19	0.43	0.2180	0.0383	-6.3	0.9	-5.7	0.9
Math	5	6889	D.1.2.1		3366	0.73	0.13	0.07	0.06	0.73	0.00	0.41	-0.30	-0.31	-0.32	0.41	-0.4645	0.0417	-5.7	0.9	-4.6	0.9
Math	5	7032	B.1.3.1	1	3366	0.41	0.39	0.41	0.15	0.05	0.00	0.22	-0.17	0.22	-0.26	-0.27	1.1329	0.0382	6.9	1.1	7.1	1.2
Math	5	7209	A.3.1.2		3366	0.26	0.16	0.26	0.19	0.38	0.00	0.29	-0.37	0.29	-0.35	-0.28	1.8823	0.0421	0.0	1.0	1.5	1.1
Math	5	7746	D.1.1.2		3366	0.65	0.18	0.13	0.65	0.03	0.00	0.27	-0.24	-0.20	0.27	-0.19	-0.0461	0.0392	2.4	1.0	3.5	1.1
Math	5	7803	D.1.1.1		3366	0.77	0.11	0.77	0.05	0.07	0.00	0.39	-0.31	0.39	-0.24	-0.26	-0.6596	0.0433	-4.2	0.9	-4.1	0.9
Math	5	7981	B.2.2.2		3366	0.54	0.02	0.20	0.54	0.23	0.00	0.31	-0.26	-0.40	0.31	-0.15	0.4848	0.0377	1.7	1.0	0.8	1.0
Math	5	8210	A.1.1.1	1	3366	0.70	0.15	0.06	0.09	0.70	0.00	0.31	-0.28	-0.21	-0.18	0.31	-0.2631	0.0404	-0.1	1.0	-0.6	1.0
Math	5	8213	A.1.2.1		3366	0.53	0.39	0.05	0.53	0.03	0.00	0.28	-0.25	-0.22	0.28	-0.19	0.5474	0.0377	3.9	1.1	3.3	1.1
Math	5	8216	A.1.5.1	2	3366	0.29	0.27	0.29	0.14	0.30	0.00	0.27	-0.34	0.27	-0.22	-0.27	1.7468	0.0411	0.3	1.0	4.7	1.2
Math	5	8452	B.2.2.1		3366	0.88	0.88	0.06	0.03	0.03	0.00	0.28	0.28	-0.22	-0.15	-0.16	-1.5151	0.0541	-2.2	0.9	-1.9	0.9
Math	5	9130	A.1.3.1	1	3366	0.74	0.74	0.08	0.13	0.05	0.00	0.32	0.32	-0.24	-0.23	-0.20	-0.4971	0.0419	-0.7	1.0	-1.3	1.0
Math	5	9399	B.1.2.2	2	3366	0.36	0.14	0.36	0.23	0.27	0.00	0.31	-0.31	0.31	-0.24	-0.37	1.3873	0.0391	0.0	1.0	1.5	1.0
Math	5	9407	B.1.2.1		3366	0.53	0.11	0.25	0.10	0.53	0.00	0.50	-0.46	-0.44	-0.40	0.50	0.5544	0.0377	-9.9	0.9	-9.4	0.8
Math	5	9492	A.2.1.1		3366	0.41	0.30	0.41	0.15	0.14	0.00	0.32	-0.32	0.32	-0.25	-0.34	1.1286	0.0382	0.2	1.0	1.6	1.0
Math	5	9508	B.2.2.3		3366	0.78	0.03	0.78	0.02	0.17	0.00	0.42	-0.24	0.42	-0.22	-0.37	-0.7518	0.0441	-5.5	0.9	-6.1	0.8
Math	5	9559	A.1.2.2	1	3366	0.59	0.20	0.59	0.09	0.12	0.00	0.37	-0.32	0.37	-0.30	-0.28	0.2524	0.0382	-2.4	1.0	-2.1	1.0
Math	6	1057	E.3.1.1	2	3600	0.32	0.36	0.32	0.22	0.10	0.00	0.09	-0.06	0.09	-0.05	-0.24	1.0633	0.0387	9.9	1.2	9.9	1.4
Math	6	1139	D.1.2.1	2	3600	0.43	0.22	0.23	0.43	0.12	0.00	0.29	-0.19	-0.37	0.29	-0.21	0.5057	0.0367	2.6	1.0	2.3	1.1
Math	6	1326	E.2.1.1	2	3600	0.32	0.14	0.32	0.23	0.31	0.00	0.24	-0.19	0.24	-0.11	-0.37	1.0545	0.0387	4.0	1.1	4.4	1.1
Math	6	1559	A.1.1.1		3600	0.39	0.04	0.39	0.24	0.33	0.00	0.48	-0.18	0.48	-0.52	-0.47	0.7194	0.0372	-9.9	0.9	-8.9	0.8
Math	6	2585	A.1.1.3		3600	0.63	0.06	0.15	0.16	0.63	0.00	0.39	-0.33	-0.37	-0.24	0.39	-0.4186	0.0373	-5.5	0.9	-4.3	0.9
Math	6	3147	B.2.1.1	1	3600	0.49	0.49	0.26	0.10	0.15	0.00	0.21	0.21	-0.23	-0.17	-0.10	0.2434	0.0363	8.7	1.1	8.2	1.2
Math	6	3204	A.2.1.1		3600	0.64	0.20	0.11	0.64	0.06	0.00	0.37	-0.23	-0.34	0.37	-0.34	-0.4580	0.0375	-4.0	0.9	-3.2	0.9
Math	6	3334	A.1.3.1		3600	0.44	0.05	0.44	0.38	0.13	0.00	0.48	-0.42	0.48	-0.43	-0.49	0.4523	0.0366	-9.9	0.9	-9.8	0.8
Math	6	3571	E.1.1.3		3600	0.76	0.76	0.10	0.06	0.07	0.00	0.36	0.36	-0.25	-0.21	-0.27	-1.1368	0.0415	-4.5	0.9	-4.5	0.9
Math	6	3851	A.1.4.1		3600	0.60	0.60	0.09	0.25	0.06	0.00	0.33	0.33	-0.25	-0.32	-0.13	-0.2809	0.0369	-1.1	1.0	-1.8	1.0
Math	6	3996	C.1.2.2	1	3600	0.70	0.16	0.04	0.70	0.11	0.00	0.21	-0.19	-0.17	0.21	-0.11	-0.7563	0.0388	3.9	1.1	4.9	1.1
Math	6	4631	A.1.1.2		3600	0.36	0.12	0.36	0.08	0.43	0.00	0.46	-0.32	0.46	-0.30	-0.52	0.8314	0.0376	-8.9	0.9	-8.0	0.8
Math	6	4732	E.1.1.2	2	3600	0.73	0.73	0.09	0.09	0.09	0.00	0.33	0.33	-0.24	-0.22	-0.24	-0.9715	0.0402	-3.0	0.9	-2.9	0.9
Math	6	4955	E.1.1.3		3600	0.54	0.06	0.54	0.09	0.31	0.00	0.33	-0.34	0.33	-0.25	-0.26	0.0115	0.0364	0.2	1.0	-0.3	1.0
Math	6	5070	E.3.1.2		3600	0.47	0.17	0.27	0.47	0.08	0.00	0.30	-0.29	-0.26	0.30	-0.23	0.3063	0.0364	2.0	1.0	2.8	1.1
Math	6	6005	D.1.2.1		3600	0.42	0.11	0.26	0.20	0.42	0.00	0.37	-0.35	-0.32	-0.36	0.37	0.5332	0.0367	-3.0	1.0	-2.6	1.0
Math	6	6084	D.2.1.2		3600	0.82	0.03	0.04	0.82	0.11	0.00	0.36	-0.09	-0.23	0.36	-0.33	-1.5299	0.0454	-5.3	0.9	-5.7	0.8
Math	6	6381	D.1.1.1		3600	0.46	0.29	0.46	0.16	0.08	0.00	0.33	-0.32	0.33	-0.27	-0.25	0.3552	0.0364	0.2	1.0	0.7	1.0
Math	6	6556	C.3.1.1		3600	0.78	0.78	0.01	0.20	0.01	0.00	0.27	0.27	-0.17	-0.24	-0.16	-1.2214	0.0422	-1.5	1.0	-0.1	1.0
Math	6	6821	C.1.1.4		3600	0.28	0.18	0.32	0.28	0.22	0.00	0.19	-0.21	-0.19	0.19	-0.19	1.2824	0.0401	4.6	1.1	7.4	1.2
Math	6	7142	A.1.3.2	1	3600	0.49	0.23	0.19	0.49	0.09	0.00	0.29	-0.27	-0.21	0.29	-0.29	0.2075	0.0363	2.7	1.0	1.7	1.0
Math	6	7320	C.1.1.2		3600	0.36	0.20	0.30	0.14	0.36	0.00	0.19	-0.06	-0.29	-0.15	0.19	0.8661	0.0378	7.7	1.1	8.1	1.2
Math	6	7331	B.1.1.1		3600	0.34	0.23	0.34	0.33	0.09	0.00	0.20	-0.29	0.20	-0.13	-0.21	0.9634	0.0382	5.9	1.1	7.1	1.2
Math	6	7760	A.1.1.4	1	3600	0.49	0.43	0.04	0.49	0.05	0.00	0.46	-0.48	-0.14	0.46	-0.25	0.2409	0.0363	-9.9	0.9	-7.4	0.9

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Math	6	8164	C.1.1.3	1	3600	0.63	0.17	0.63	0.08	0.12	0.00	0.29	-0.30	0.29	-0.18	-0.16	-0.4362	0.0374	1.1	1.0	0.2	1.0
Math	6	8585	E.1.1.1		3600	0.55	0.25	0.55	0.13	0.07	0.00	0.32	-0.32	0.32	-0.13	-0.34	-0.0670	0.0365	0.2	1.0	0.4	1.0
Math	6	9012	A.1.1.3	1	3600	0.44	0.28	0.44	0.18	0.09	0.00	0.29	-0.26	0.29	-0.24	-0.31	0.4692	0.0366	3.0	1.0	2.6	1.1
Math	6	9193	C.1.1.3		3600	0.49	0.20	0.49	0.10	0.21	0.00	0.29	-0.24	0.29	-0.35	-0.19	0.2204	0.0363	2.9	1.0	3.7	1.1
Math	6	9569	A.1.2.1		3600	0.26	0.04	0.09	0.61	0.26	0.00	0.33	-0.19	-0.23	-0.37	0.33	1.3658	0.0407	-1.7	1.0	-0.7	1.0
Math	6	9609	A.1.4.1	2	3600	0.23	0.59	0.08	0.23	0.09	0.00	0.39	-0.40	-0.39	0.39	-0.40	1.5396	0.0420	-4.7	0.9	-3.6	0.9
Math	7	0015	D.2.2.1	2	3972	0.62	0.62	0.09	0.23	0.05	0.00	0.26	0.26	-0.20	-0.23	-0.15	-0.1846	0.0349	1.6	1.0	1.7	1.0
Math	7	0420	D.3.1.1	2	3972	0.49	0.49	0.20	0.23	0.08	0.00	0.21	0.21	-0.13	-0.23	-0.19	0.4231	0.0341	6.1	1.1	5.8	1.1
Math	7	0426	E.1.1.1	1	3972	0.49	0.15	0.11	0.49	0.24	0.00	0.30	-0.22	-0.30	0.30	-0.27	0.3981	0.0341	-0.2	1.0	-0.5	1.0
Math	7	1250	C.1.2.2	1	3972	0.60	0.60	0.13	0.08	0.18	0.00	0.36	0.36	-0.30	-0.21	-0.33	-0.0878	0.0346	-4.9	0.9	-4.3	0.9
Math	7	1272	A.2.2.5	2	3972	0.54	0.12	0.54	0.24	0.10	0.00	0.25	-0.28	0.25	-0.16	-0.23	0.1842	0.0342	3.2	1.0	2.7	1.0
Math	7	2365	B.2.1.3	1	3972	0.42	0.14	0.42	0.40	0.05	0.00	0.31	-0.35	0.31	-0.29	-0.13	0.7305	0.0346	-1.1	1.0	0.7	1.0
Math	7	2596	C.1.1.1	1	3972	0.49	0.49	0.15	0.19	0.16	0.00	0.37	0.37	-0.32	-0.34	-0.33	0.3890	0.0341	-5.4	0.9	-4.9	0.9
Math	7	3027	E.2.1.1	1	3972	0.76	0.13	0.05	0.76	0.05	0.00	0.27	-0.23	-0.12	0.27	-0.19	-0.9376	0.0391	-1.3	1.0	-1.2	1.0
Math	7	3642	C.1.1.2	1	3972	0.52	0.11	0.19	0.52	0.18	0.00	0.34	-0.25	-0.42	0.34	-0.17	0.2492	0.0341	-2.8	1.0	-3.0	1.0
Math	7	4302	D.2.1.1	1	3972	0.60	0.60	0.21	0.09	0.11	0.00	0.34	0.34	-0.26	-0.35	-0.22	-0.0960	0.0347	-3.4	1.0	-3.7	0.9
Math	7	4613	E.2.1.1	1	3972	0.73	0.14	0.73	0.09	0.04	0.00	0.30	-0.28	0.30	-0.14	-0.22	-0.7424	0.0376	-2.6	1.0	-1.3	1.0
Math	7	4647	A.3.2.2	1	3972	0.57	0.11	0.13	0.57	0.18	0.00	0.31	-0.29	-0.21	0.31	-0.26	0.0410	0.0344	-0.9	1.0	-0.3	1.0
Math	7	4775	A.2.2.1	2	3972	0.47	0.47	0.33	0.11	0.09	0.00	0.28	0.28	-0.24	-0.21	-0.34	0.4881	0.0342	1.3	1.0	0.6	1.0
Math	7	4970	A.1.2.1		3972	0.55	0.18	0.15	0.13	0.55	0.00	0.41	-0.33	-0.35	-0.36	0.41	0.1534	0.0342	-8.3	0.9	-7.3	0.9
Math	7	5474	A.1.1.1	1	3972	0.38	0.24	0.38	0.34	0.04	0.00	0.38	-0.38	0.38	-0.42	-0.12	0.9317	0.0351	-5.1	0.9	-5.4	0.9
Math	7	5649	A.2.1.1	1	3972	0.56	0.14	0.10	0.19	0.56	0.00	0.24	-0.28	-0.23	-0.09	0.24	0.0675	0.0343	4.0	1.1	3.2	1.1
Math	7	5816	E.4.1.1	2	3972	0.38	0.16	0.29	0.38	0.17	0.00	0.17	-0.21	-0.03	0.17	-0.31	0.8954	0.0350	7.6	1.1	7.8	1.1
Math	7	6311	D.2.1.2	1	3972	0.53	0.10	0.27	0.53	0.10	0.00	0.31	-0.29	-0.30	0.31	-0.15	0.2298	0.0342	-0.7	1.0	-1.1	1.0
Math	7	6409	D.2.2.1	2	3972	0.54	0.15	0.54	0.11	0.20	0.00	0.22	-0.18	0.22	-0.23	-0.15	0.1672	0.0342	5.6	1.1	4.5	1.1
Math	7	6501	E.1.1.1	2	3972	0.81	0.02	0.10	0.81	0.07	0.00	0.30	-0.13	-0.28	0.30	-0.15	-1.2380	0.0419	-3.3	0.9	-3.3	0.9
Math	7	6512	B.2.2.1	2	3972	0.45	0.45	0.18	0.29	0.08	0.00	0.19	0.19	-0.24	-0.11	-0.20	0.5855	0.0343	7.6	1.1	7.6	1.1
Math	7	6813	A.2.2.2	2	3972	0.61	0.10	0.16	0.61	0.13	0.00	0.33	-0.26	-0.25	0.33	-0.27	-0.1242	0.0347	-2.8	1.0	-2.2	1.0
Math	7	7600	D.2.1.2	1	3972	0.44	0.16	0.25	0.15	0.44	0.00	0.33	-0.38	-0.21	-0.33	0.33	0.6143	0.0343	-1.9	1.0	-0.9	1.0
Math	7	8019	A.2.2.4		3972	0.63	0.63	0.09	0.07	0.21	0.00	0.35	0.35	-0.28	-0.28	-0.27	-0.2492	0.0351	-4.4	0.9	-4.2	0.9
Math	7	8626	D.3.1.2	2	3972	0.43	0.45	0.43	0.06	0.07	0.00	0.28	-0.28	0.28	-0.22	-0.22	0.6931	0.0345	1.0	1.0	1.5	1.0
Math	7	8883	A.3.2.2	1	3972	0.57	0.57	0.23	0.13	0.07	0.00	0.24	0.24	-0.26	-0.12	-0.15	0.0341	0.0344	3.9	1.1	3.1	1.1
Math	7	8924	C.1.1.1	1	3972	0.65	0.08	0.65	0.17	0.10	0.00	0.30	-0.26	0.30	-0.18	-0.26	-0.3602	0.0356	-1.0	1.0	-0.8	1.0
Math	7	9543	D.3.1.1	2	3972	0.53	0.12	0.53	0.24	0.11	0.00	0.23	-0.27	0.23	-0.11	-0.24	0.2218	0.0342	5.1	1.1	3.8	1.1
Math	7	9581	D.1.1.1	2	3972	0.60	0.02	0.60	0.16	0.22	0.00	0.26	-0.16	0.26	-0.17	-0.26	-0.0831	0.0346	1.9	1.0	1.5	1.0
Math	7	9763	E.3.1.1	2	3972	0.43	0.43	0.10	0.23	0.25	0.00	0.35	0.35	-0.30	-0.39	-0.27	0.6920	0.0345	-3.9	1.0	-3.1	1.0
Math	8	0312	E.1.1.3	1	4114	0.69	0.08	0.12	0.69	0.11	0.00	0.40	-0.21	-0.29	0.40	-0.37	-0.6541	0.0361	-6.6	0.9	-6.2	0.9
Math	8	0831	A.1.1.1	1	4114	0.62	0.02	0.62	0.17	0.19	0.00	0.41	-0.18	0.41	-0.33	-0.37	-0.2943	0.0347	-7.1	0.9	-6.0	0.9
Math	8	1173	A.3.3.1	2	4114	0.30	0.25	0.19	0.30	0.26	0.00	0.19	-0.18	-0.21	0.19	-0.21	1.2694	0.0368	5.8	1.1	6.7	1.2
Math	8	1671	A.2.2.1	1	4114	0.39	0.28	0.19	0.39	0.14	0.00	0.16	-0.12	-0.16	0.16	-0.19	0.8000	0.0348	9.9	1.1	9.9	1.2
Math	8	1856	A.1.1.1	1	4114	0.57	0.57	0.11	0.25	0.06	0.00	0.28	0.28	-0.28	-0.21	-0.19	-0.0604	0.0341	3.0	1.0	2.0	1.0
Math	8	1909	C.1.1.1	2	4114	0.53	0.15	0.05	0.27	0.53	0.00	0.31	-0.29	-0.20	-0.26	0.31	0.1232	0.0339	1.4	1.0	1.2	1.0

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Math	8	2357	D.2.2.2	1	4114	0.68	0.10	0.68	0.08	0.15	0.00	0.46	-0.32	0.46	-0.29	-0.42	-0.5620	0.0357	-9.9	0.9	-9.6	0.8
Math	8	3388	A.2.2.2	1	4114	0.71	0.04	0.71	0.11	0.14	0.00	0.30	-0.13	0.30	-0.32	-0.18	-0.7292	0.0365	-1.3	1.0	-0.6	1.0
Math	8	3409	C.1.2.1	2	4114	0.39	0.15	0.10	0.39	0.35	0.00	0.26	-0.25	-0.21	0.26	-0.26	0.7647	0.0347	4.1	1.1	3.8	1.1
Math	8	3989	D.2.1.1	1	4114	0.72	0.09	0.14	0.72	0.06	0.00	0.45	-0.31	-0.36	0.45	-0.29	-0.7820	0.0368	-9.3	0.9	-9.5	0.8
Math	8	4016	C.1.1.2	2	4114	0.53	0.08	0.12	0.53	0.27	0.00	0.25	-0.25	-0.23	0.25	-0.19	0.1456	0.0339	5.4	1.1	4.6	1.1
Math	8	4269	E.4.1.1	1	4114	0.42	0.42	0.23	0.20	0.15	0.00	0.26	0.26	-0.25	-0.26	-0.20	0.6335	0.0343	4.5	1.1	4.4	1.1
Math	8	4472	D.1.1.3	2	4114	0.43	0.24	0.18	0.43	0.15	0.00	0.38	-0.26	-0.39	0.38	-0.42	0.6002	0.0342	-4.0	1.0	-3.1	1.0
Math	8	4626	C.1.1.3	1	4114	0.58	0.10	0.17	0.14	0.58	0.00	0.24	-0.26	-0.19	-0.14	0.24	-0.1174	0.0342	5.3	1.1	5.8	1.1
Math	8	4746	D.2.1.3	1	4114	0.57	0.24	0.14	0.57	0.05	0.00	0.41	-0.36	-0.34	0.41	-0.28	-0.0344	0.0341	-6.8	0.9	-6.2	0.9
Math	8	4884	E.4.1.2	2	4114	0.40	0.09	0.45	0.40	0.06	0.00	0.30	-0.23	-0.29	0.30	-0.25	0.7401	0.0346	1.7	1.0	1.9	1.0
Math	8	4942	D.1.1.2	1	4114	0.45	0.23	0.21	0.45	0.10	0.00	0.43	-0.43	-0.34	0.43	-0.38	0.4843	0.0341	-8.1	0.9	-6.9	0.9
Math	8	5113	B.2.1.2	1	4114	0.44	0.31	0.44	0.18	0.07	0.00	0.22	-0.22	0.22	-0.19	-0.15	0.5501	0.0342	7.6	1.1	6.7	1.1
Math	8	5228	D.4.1.3	2	4114	0.57	0.57	0.16	0.11	0.16	0.00	0.42	0.42	-0.36	-0.28	-0.36	-0.0344	0.0341	-7.1	0.9	-7.1	0.9
Math	8	5426	D.4.1.2	1	4114	0.51	0.27	0.11	0.11	0.51	0.00	0.27	-0.19	-0.28	-0.22	0.27	0.2217	0.0339	4.7	1.1	3.8	1.1
Math	8	5548	B.1.1.2	1	4114	0.46	0.20	0.16	0.18	0.46	0.00	0.37	-0.34	-0.35	-0.31	0.37	0.4651	0.0340	-3.2	1.0	-3.4	0.9
Math	8	5773	A.3.1.2	1	4114	0.46	0.10	0.18	0.26	0.46	0.00	0.26	-0.19	-0.28	-0.21	0.26	0.4369	0.0340	5.3	1.1	5.4	1.1
Math	8	6149	E.1.1.2	2	4114	0.85	0.06	0.04	0.85	0.05	0.00	0.33	-0.22	-0.20	0.33	-0.23	-1.6298	0.0445	-4.5	0.9	-5.5	0.8
Math	8	7024	D.2.1.2	1	4114	0.51	0.51	0.24	0.16	0.09	0.00	0.45	0.45	-0.43	-0.37	-0.32	0.2273	0.0339	-9.9	0.9	-8.3	0.9
Math	8	7159	C.3.1.1	1	4114	0.71	0.71	0.28	0.01	0.01	0.00	0.38	0.38	-0.36	-0.15	-0.15	-0.7265	0.0365	-5.0	0.9	-5.2	0.9
Math	8	7470	D.2.2.1	2	4114	0.35	0.11	0.39	0.14	0.35	0.00	0.13	-0.14	-0.08	-0.23	0.13	0.9646	0.0353	9.9	1.2	9.9	1.2
Math	8	7593	E.3.1.1	2	4114	0.51	0.12	0.28	0.09	0.51	0.00	0.38	-0.25	-0.38	-0.31	0.38	0.2161	0.0339	-4.5	1.0	-3.4	1.0
Math	8	8177	A.2.1.1	1	4114	0.66	0.06	0.25	0.66	0.02	0.00	0.24	-0.23	-0.18	0.24	-0.16	-0.4904	0.0354	3.3	1.1	3.0	1.1
Math	8	8244	E.3.2.1	2	4114	0.33	0.35	0.09	0.33	0.22	0.01	0.19	-0.26	-0.24	0.19	-0.07	1.0712	0.0358	6.6	1.1	7.5	1.2
Math	8	9065	D.4.1.1	1	4114	0.49	0.49	0.16	0.16	0.19	0.00	0.35	0.35	-0.30	-0.31	-0.31	0.2977	0.0339	-2.0	1.0	-1.8	1.0
Math	11	0141	D.2.2.1		4269	0.39	0.39	0.35	0.13	0.13	0.00	0.43	0.43	-0.36	-0.46	-0.41	0.6235	0.0344	-5.5	0.9	-4.1	0.9
Math	11	0843	E.2.1.2		4269	0.29	0.15	0.37	0.29	0.19	0.00	0.33	-0.20	-0.38	0.33	-0.39	1.1488	0.0368	0.1	1.0	0.7	1.0
Math	11	1786	A.2.1.3		4269	0.45	0.05	0.07	0.45	0.42	0.00	0.39	-0.36	-0.36	0.39	-0.34	0.3159	0.0337	-2.9	1.0	-3.0	1.0
Math	11	2051	B.2.2.4	2	4269	0.49	0.11	0.17	0.49	0.23	0.00	0.43	-0.39	-0.39	0.43	-0.32	0.1363	0.0335	-6.2	0.9	-5.4	0.9
Math	11	2751	D.3.2.2	2	4269	0.41	0.41	0.28	0.20	0.11	0.00	0.34	0.34	-0.31	-0.35	-0.24	0.5111	0.0341	0.6	1.0	0.6	1.0
Math	11	2790	D.3.2.1	2	4269	0.46	0.12	0.46	0.28	0.14	0.00	0.31	-0.32	0.31	-0.24	-0.31	0.3006	0.0337	2.6	1.0	2.1	1.0
Math	11	3072	A.2.2.1	1	4269	0.39	0.13	0.33	0.15	0.39	0.00	0.37	-0.39	-0.37	-0.28	0.37	0.6269	0.0344	-1.2	1.0	-1.5	1.0
Math	11	3580	D.4.1.1		4269	0.38	0.28	0.17	0.16	0.38	0.00	0.26	-0.18	-0.30	-0.29	0.26	0.6441	0.0345	5.7	1.1	5.6	1.1
Math	11	3719	A.1.1.3		4269	0.33	0.33	0.19	0.44	0.03	0.00	0.41	0.41	-0.39	-0.45	-0.16	0.9086	0.0355	-4.0	0.9	-3.5	0.9
Math	11	4499	C.1.2.3		4269	0.54	0.19	0.19	0.54	0.08	0.00	0.30	-0.26	-0.26	0.30	-0.22	-0.0727	0.0335	2.6	1.0	1.8	1.0
Math	11	4902	C.1.1.2	1	4269	0.43	0.25	0.22	0.43	0.10	0.00	0.30	-0.31	-0.23	0.30	-0.28	0.4185	0.0339	3.8	1.1	3.8	1.1
Math	11	5137	B.2.2.4	2	4269	0.78	0.78	0.08	0.09	0.05	0.00	0.32	0.32	-0.24	-0.22	-0.16	-1.3203	0.0386	-3.8	0.9	-4.6	0.9
Math	11	5217	D.2.1.4	1	4269	0.48	0.12	0.21	0.19	0.48	0.00	0.33	-0.27	-0.26	-0.30	0.33	0.1959	0.0336	1.7	1.0	1.0	1.0
Math	11	5461	D.4.1.1	1	4269	0.35	0.26	0.24	0.14	0.35	0.00	0.39	-0.42	-0.40	-0.28	0.39	0.7950	0.0350	-2.8	1.0	-2.3	1.0
Math	11	5483	D.2.1.5		4269	0.46	0.17	0.17	0.20	0.46	0.00	0.33	-0.27	-0.31	-0.31	0.33	0.2853	0.0337	1.1	1.0	1.1	1.0
Math	11	5527	D.1.1.1	2	4269	0.46	0.16	0.46	0.12	0.26	0.00	0.35	-0.37	0.35	-0.31	-0.26	0.2624	0.0336	0.0	1.0	-0.1	1.0
Math	11	5794	D.2.1.4		4269	0.62	0.12	0.16	0.62	0.10	0.00	0.11	-0.07	-0.08	0.11	-0.10	-0.4499	0.0341	9.9	1.2	9.9	1.3
Math	11	6192	B.2.2.2		4269	0.49	0.19	0.49	0.14	0.17	0.00	0.22	-0.30	0.22	-0.27	0.01	0.1601	0.0335	9.5	1.1	8.0	1.1

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Math	11	6443	D.3.2.2	1	4269	0.55	0.13	0.20	0.55	0.12	0.00	0.35	-0.28	-0.34	0.35	-0.19	-0.1182	0.0335	-0.7	1.0	-0.6	1.0
Math	11	7067	D.3.1.1		4269	0.37	0.18	0.37	0.21	0.24	0.00	0.28	-0.11	0.28	-0.27	-0.39	0.7079	0.0347	4.6	1.1	3.8	1.1
Math	11	7239	D.3.2.3		4269	0.48	0.48	0.13	0.13	0.26	0.00	0.41	0.41	-0.37	-0.34	-0.35	0.1699	0.0335	-5.0	0.9	-4.8	0.9
Math	11	7304	E.1.1.2		4269	0.44	0.44	0.30	0.16	0.10	0.00	0.34	0.34	-0.37	-0.23	-0.25	0.3522	0.0338	0.7	1.0	0.5	1.0
Math	11	8047	D.2.2.2		4269	0.44	0.23	0.44	0.22	0.09	0.01	0.29	-0.20	0.29	-0.30	-0.27	0.3566	0.0338	4.3	1.1	3.9	1.1
Math	11	8058	A.1.3.1		4269	0.58	0.20	0.58	0.04	0.18	0.00	0.28	-0.29	0.28	-0.18	-0.17	-0.2787	0.0337	3.6	1.0	3.7	1.1
Math	11	8093	A.2.1.2		4269	0.48	0.06	0.22	0.48	0.23	0.00	0.34	-0.27	-0.36	0.34	-0.25	0.1938	0.0336	0.3	1.0	1.0	1.0
Math	11	8473	D.3.1.2		4269	0.45	0.29	0.07	0.19	0.45	0.00	0.24	-0.11	-0.34	-0.31	0.24	0.3445	0.0338	8.1	1.1	7.8	1.1
Math	11	8528	B.2.2.1		4269	0.46	0.14	0.07	0.33	0.46	0.00	0.58	-0.47	-0.40	-0.58	0.58	0.2864	0.0337	-9.9	0.8	-9.9	0.7
Math	11	8792	D.2.2.2	1	4269	0.39	0.39	0.19	0.29	0.12	0.00	0.39	0.39	-0.37	-0.35	-0.40	0.6212	0.0344	-3.2	1.0	-1.5	1.0
Math	11	8811	B.2.1.1		4269	0.71	0.13	0.13	0.71	0.04	0.00	0.27	-0.15	-0.24	0.27	-0.18	-0.9316	0.0360	0.3	1.0	0.0	1.0
Math	11	9545	D.2.2.1	1	4269	0.41	0.12	0.31	0.16	0.41	0.00	0.41	-0.27	-0.38	-0.45	0.41	0.5190	0.0341	-4.4	0.9	-3.7	0.9
Reading	4	0321	A.1.6.1	2	3388	0.48	0.48	0.19	0.15	0.17	0.00	0.20	0.20	-0.20	-0.16	-0.17	0.2736	0.0374	8.5	1.1	8.2	1.2
Reading	4	0409	B.1.1.1	1	3388	0.85	0.08	0.03	0.85	0.03	0.00	0.21	-0.11	-0.18	0.21	-0.15	-1.8124	0.0503	-0.8	1.0	1.2	1.1
Reading	4	0906	A.1.1.2	1	3388	0.54	0.14	0.22	0.54	0.11	0.00	0.40	-0.40	-0.33	0.40	-0.32	-0.0005	0.0375	-5.0	0.9	-3.9	0.9
Reading	4	0932	A.2.4.1	1	3388	0.53	0.17	0.11	0.53	0.18	0.00	0.40	-0.29	-0.41	0.40	-0.32	0.0473	0.0374	-4.5	0.9	-3.9	0.9
Reading	4	0970	A.2.4.1	1	3388	0.34	0.23	0.18	0.24	0.34	0.01	0.31	-0.33	-0.36	-0.28	0.31	0.9497	0.0393	-0.9	1.0	0.6	1.0
Reading	4	0990	A.1.3.1	2	3388	0.41	0.41	0.19	0.13	0.28	0.00	0.12	0.12	-0.07	-0.22	-0.09	0.5959	0.0380	9.9	1.2	9.9	1.2
Reading	4	1556	A.1.1.1	2	3388	0.69	0.69	0.09	0.11	0.11	0.00	0.48	0.48	-0.37	-0.37	-0.35	-0.7320	0.0398	-9.3	0.9	-8.9	0.8
Reading	4	2109	B.2.1.1	2	3388	0.38	0.25	0.17	0.38	0.20	0.00	0.22	-0.30	-0.18	0.22	-0.14	0.7154	0.0383	5.3	1.1	6.1	1.1
Reading	4	2368	A.1.4.1	1	3388	0.65	0.65	0.18	0.06	0.10	0.00	0.40	0.40	-0.33	-0.28	-0.30	-0.5524	0.0389	-5.1	0.9	-5.4	0.9
Reading	4	3067	B.1.1.1	1	3388	0.48	0.15	0.48	0.22	0.14	0.00	0.32	-0.33	0.32	-0.26	-0.28	0.2503	0.0374	0.9	1.0	1.0	1.0
Reading	4	3156	A.1.3.1	2	3388	0.60	0.60	0.12	0.12	0.16	0.00	0.35	0.35	-0.32	-0.28	-0.26	-0.2978	0.0380	-1.9	1.0	-1.1	1.0
Reading	4	4059	B.1.1.1	1	3388	0.56	0.10	0.28	0.07	0.56	0.00	0.44	-0.42	-0.33	-0.43	0.44	-0.0882	0.0376	-7.2	0.9	-6.3	0.9
Reading	4	4128	A.2.4.1	1	3388	0.65	0.15	0.65	0.10	0.10	0.00	0.40	-0.31	0.40	-0.34	-0.28	-0.5245	0.0388	-4.9	0.9	-3.7	0.9
Reading	4	4566	A.2.3.1	2	3388	0.49	0.49	0.24	0.12	0.16	0.00	0.38	0.38	-0.31	-0.42	-0.30	0.2258	0.0374	-3.6	1.0	-3.3	0.9
Reading	4	4747	A.1.5.1	2	3388	0.41	0.22	0.18	0.18	0.41	0.00	0.32	-0.34	-0.27	-0.30	0.32	0.5818	0.0379	-0.5	1.0	0.3	1.0
Reading	4	5351	A.1.4.1	1	3388	0.72	0.16	0.07	0.04	0.72	0.00	0.34	-0.31	-0.17	-0.22	0.34	-0.9297	0.0410	-2.4	1.0	-2.7	0.9
Reading	4	5419	A.2.2.2	2	3388	0.72	0.09	0.10	0.10	0.72	0.00	0.48	-0.36	-0.35	-0.36	0.48	-0.8839	0.0407	-9.2	0.8	-9.5	0.7
Reading	4	5661	A.1.6.1	2	3388	0.44	0.11	0.20	0.44	0.25	0.00	0.23	-0.36	-0.18	0.23	-0.17	0.4632	0.0377	6.1	1.1	6.0	1.1
Reading	4	5671	A.1.2.1	1	3388	0.64	0.12	0.64	0.15	0.08	0.00	0.34	-0.26	0.34	-0.29	-0.25	-0.4909	0.0387	-1.6	1.0	-2.6	0.9
Reading	4	7454	B.1.1.1	1	3388	0.57	0.21	0.57	0.11	0.11	0.00	0.46	-0.40	0.46	-0.37	-0.39	-0.1695	0.0377	-9.2	0.9	-8.1	0.9
Reading	4	7467	A.1.4.1	1	3388	0.69	0.11	0.12	0.08	0.69	0.00	0.30	-0.22	-0.19	-0.28	0.30	-0.7444	0.0399	-0.3	1.0	-0.1	1.0
Reading	4	7756	A.1.4.1	2	3388	0.63	0.63	0.22	0.08	0.07	0.00	0.20	0.20	-0.17	-0.16	-0.11	-0.4502	0.0385	6.8	1.1	6.5	1.2
Reading	4	8495	A.2.4.1	1	3388	0.49	0.49	0.22	0.12	0.17	0.00	0.32	0.32	-0.29	-0.40	-0.18	0.1931	0.0374	0.3	1.0	0.4	1.0
Reading	4	8698	A.2.3.1	2	3388	0.48	0.17	0.48	0.19	0.16	0.01	0.31	-0.24	0.31	-0.24	-0.36	0.2763	0.0374	1.1	1.0	1.3	1.0
Reading	4	8953	A.1.4.1	1	3388	0.59	0.16	0.16	0.59	0.09	0.00	0.46	-0.40	-0.36	0.46	-0.37	-0.2502	0.0379	-8.9	0.9	-8.1	0.9
Reading	4	8995	B.2.1.1	1	3388	0.41	0.30	0.41	0.16	0.13	0.00	0.25	-0.24	0.25	-0.21	-0.24	0.6058	0.0380	4.3	1.1	5.1	1.1
Reading	4	9110	B.3.3.1	2	3388	0.30	0.33	0.30	0.17	0.19	0.00	0.10	-0.04	0.10	-0.19	-0.12	1.1621	0.0405	8.2	1.2	9.9	1.3
Reading	4	9433	B.3.3.2	2	3388	0.36	0.14	0.37	0.36	0.12	0.00	0.26	-0.25	-0.25	0.26	-0.23	0.8112	0.0387	2.8	1.0	4.4	1.1
Reading	4	9619	A.1.3.1	2	3388	0.32	0.14	0.34	0.19	0.32	0.00	0.19	-0.23	-0.13	-0.27	0.19	1.0465	0.0398	4.9	1.1	7.1	1.2
Reading	4	9899	A.2.6.1	2	3388	0.59	0.59	0.13	0.13	0.14	0.00	0.33	0.33	-0.26	-0.29	-0.26	-0.2712	0.0379	-0.6	1.0	-0.1	1.0

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Reading	5	0087	A.2.4.1	2	3947	0.68	0.17	0.04	0.10	0.68	0.00	0.41	-0.22	-0.37	-0.44	0.41	-0.4605	0.0375	-3.3	1.0	-0.9	1.0
Reading	5	0275	A.1.4.1	1	3947	0.63	0.18	0.11	0.63	0.07	0.00	0.40	-0.32	-0.31	0.40	-0.34	-0.2280	0.0366	-2.1	1.0	-2.8	0.9
Reading	5	0296	A.2.3.1	2	3947	0.55	0.14	0.15	0.16	0.55	0.00	0.34	-0.24	-0.26	-0.34	0.34	0.1682	0.0356	3.1	1.0	1.8	1.0
Reading	5	1019	B.3.1.1	1	3947	0.49	0.20	0.12	0.49	0.18	0.00	0.36	-0.34	-0.33	0.36	-0.28	0.4832	0.0355	1.6	1.0	2.1	1.0
Reading	5	1470	A.1.4.1	1	3947	0.76	0.76	0.12	0.07	0.06	0.00	0.33	0.33	-0.19	-0.29	-0.25	-0.9160	0.0403	-0.1	1.0	0.1	1.0
Reading	5	1648	A.2.4.1	1	3947	0.54	0.13	0.16	0.17	0.54	0.00	0.40	-0.45	-0.22	-0.38	0.40	0.2411	0.0356	-1.8	1.0	-1.8	1.0
Reading	5	2327	B.2.1.1	2	3947	0.47	0.23	0.09	0.47	0.20	0.00	0.30	-0.26	-0.36	0.30	-0.25	0.5570	0.0355	5.3	1.1	5.6	1.1
Reading	5	2851	A.1.4.1	1	3947	0.55	0.25	0.07	0.13	0.55	0.00	0.46	-0.44	-0.41	-0.30	0.46	0.1991	0.0356	-5.9	0.9	-5.9	0.9
Reading	5	3292	A.2.3.1	2	3947	0.57	0.57	0.13	0.10	0.20	0.00	0.38	0.38	-0.34	-0.34	-0.28	0.1123	0.0357	0.3	1.0	0.5	1.0
Reading	5	4160	A.2.3.1	2	3947	0.74	0.13	0.05	0.08	0.74	0.00	0.55	-0.44	-0.37	-0.42	0.55	-0.8193	0.0396	-9.9	0.8	-9.9	0.7
Reading	5	4360	B.1.1.1	2	3947	0.49	0.23	0.11	0.49	0.17	0.00	0.32	-0.23	-0.40	0.32	-0.25	0.4955	0.0355	4.7	1.1	5.2	1.1
Reading	5	4490	A.1.3.1	2	3947	0.45	0.14	0.45	0.29	0.12	0.00	0.37	-0.36	0.37	-0.35	-0.31	0.6943	0.0356	-0.8	1.0	1.5	1.0
Reading	5	4888	A.1.4.1	1	3947	0.85	0.04	0.06	0.04	0.85	0.00	0.45	-0.30	-0.33	-0.27	0.45	-1.6079	0.0471	-6.8	0.8	-7.9	0.6
Reading	5	5234	A.1.4.1	2	3947	0.64	0.64	0.20	0.08	0.08	0.00	0.29	0.29	-0.17	-0.31	-0.29	-0.2647	0.0367	4.6	1.1	5.5	1.1
Reading	5	5428	B.3.1.1	1	3947	0.57	0.57	0.14	0.14	0.15	0.00	0.45	0.45	-0.42	-0.39	-0.32	0.0974	0.0358	-5.0	0.9	-4.7	0.9
Reading	5	6307	A.1.1.2	1	3947	0.36	0.32	0.36	0.20	0.11	0.00	0.31	-0.36	0.31	-0.30	-0.24	1.1457	0.0367	2.1	1.0	3.1	1.1
Reading	5	6714	A.1.3.1	2	3947	0.41	0.22	0.29	0.41	0.08	0.00	0.21	-0.27	-0.11	0.21	-0.32	0.8635	0.0359	9.9	1.2	9.9	1.3
Reading	5	7009	A.1.4.1	1	3947	0.63	0.09	0.63	0.24	0.03	0.00	0.35	-0.22	0.35	-0.32	-0.23	-0.1981	0.0365	1.8	1.0	0.1	1.0
Reading	5	7914	B.2.1.2	1	3947	0.62	0.62	0.14	0.10	0.13	0.00	0.36	0.36	-0.29	-0.30	-0.27	-0.1852	0.0364	1.0	1.0	0.5	1.0
Reading	5	7976	A.2.1.2	2	3947	0.75	0.75	0.07	0.09	0.09	0.00	0.39	0.39	-0.26	-0.21	-0.37	-0.9049	0.0402	-3.0	0.9	-3.3	0.9
Reading	5	8052	A.2.4.1	2	3947	0.62	0.09	0.10	0.19	0.62	0.00	0.42	-0.44	-0.38	-0.25	0.42	-0.1452	0.0363	-3.1	1.0	-3.3	0.9
Reading	5	8059	B.1.1.1	2	3947	0.59	0.07	0.17	0.59	0.17	0.00	0.40	-0.44	-0.37	0.40	-0.23	0.0048	0.0359	-1.2	1.0	-1.5	1.0
Reading	5	8224	A.2.6.1	1	3947	0.41	0.16	0.20	0.23	0.41	0.00	0.28	-0.17	-0.30	-0.31	0.28	0.8913	0.0360	5.4	1.1	6.2	1.1
Reading	5	8438	A.2.4.1	1	3947	0.47	0.19	0.47	0.17	0.17	0.00	0.36	-0.39	0.36	-0.34	-0.24	0.5718	0.0355	0.3	1.0	1.9	1.0
Reading	5	8605	B.2.2.1	2	3947	0.35	0.32	0.09	0.24	0.35	0.00	0.24	-0.10	-0.41	-0.36	0.24	1.2020	0.0369	6.8	1.1	7.2	1.2
Reading	5	8843	A.2.4.1	1	3947	0.66	0.16	0.09	0.09	0.66	0.00	0.48	-0.39	-0.34	-0.41	0.48	-0.3911	0.0372	-7.4	0.9	-8.0	0.8
Reading	5	9136	A.1.6.1	2	3947	0.60	0.60	0.12	0.19	0.10	0.00	0.38	0.38	-0.39	-0.30	-0.24	-0.0369	0.0360	-0.1	1.0	0.0	1.0
Reading	5	9445	A.1.4.1	2	3947	0.61	0.21	0.08	0.61	0.10	0.00	0.35	-0.25	-0.32	0.35	-0.33	-0.0863	0.0362	1.6	1.0	2.1	1.1
Reading	5	9710	B.1.1.1	2	3947	0.80	0.06	0.07	0.06	0.80	0.00	0.40	-0.23	-0.31	-0.28	0.40	-1.2316	0.0430	-4.6	0.9	-3.7	0.8
Reading	5	9866	A.1.1.1	2	3947	0.64	0.07	0.09	0.64	0.20	0.00	0.38	-0.35	-0.30	0.38	-0.27	-0.2515	0.0367	-0.3	1.0	0.1	1.0
Reading	6	0086	B.3.3.2	1	3983	0.54	0.54	0.19	0.13	0.12	0.01	0.28	0.28	-0.25	-0.24	-0.22	0.4559	0.0354	7.1	1.1	6.5	1.1
Reading	6	0386	A.2.4.1	2	3983	0.67	0.67	0.10	0.13	0.10	0.00	0.49	0.49	-0.50	-0.29	-0.39	-0.2042	0.0373	-7.5	0.9	-6.8	0.8
Reading	6	0695	A.1.6.1	2	3983	0.69	0.07	0.17	0.07	0.69	0.00	0.40	-0.35	-0.28	-0.29	0.40	-0.2897	0.0377	-1.8	1.0	-2.1	0.9
Reading	6	0771	B.1.2.1	2	3983	0.42	0.14	0.42	0.21	0.22	0.00	0.29	-0.35	0.29	-0.40	-0.12	1.0368	0.0356	5.1	1.1	6.3	1.1
Reading	6	0829	B.2.1.3	1	3983	0.56	0.56	0.17	0.20	0.08	0.00	0.28	0.28	-0.22	-0.22	-0.28	0.3992	0.0355	7.4	1.1	6.3	1.1
Reading	6	1065	B.1.1.1	1	3983	0.62	0.21	0.62	0.07	0.10	0.00	0.39	-0.35	0.39	-0.30	-0.29	0.0774	0.0362	-1.0	1.0	-0.9	1.0
Reading	6	1134	A.2.2.2	2	3983	0.77	0.77	0.07	0.08	0.07	0.00	0.40	0.40	-0.31	-0.30	-0.25	-0.7566	0.0407	-3.2	0.9	-3.2	0.9
Reading	6	1186	A.2.3.1	2	3983	0.72	0.15	0.09	0.72	0.04	0.00	0.43	-0.30	-0.40	0.43	-0.28	-0.4555	0.0387	-4.2	0.9	-4.8	0.9
Reading	6	1839	A.1.4.1	1	3983	0.88	0.88	0.02	0.07	0.03	0.00	0.45	0.45	-0.29	-0.34	-0.27	-1.6197	0.0504	-6.0	0.8	-8.1	0.6
Reading	6	2645	B.3.1.1	1	3983	0.54	0.54	0.13	0.15	0.15	0.02	0.37	0.37	-0.35	-0.31	-0.29	0.4583	0.0354	0.3	1.0	-0.3	1.0
Reading	6	3395	A.1.3.1	2	3983	0.49	0.49	0.36	0.02	0.13	0.00	0.32	0.32	-0.32	-0.30	-0.23	0.7301	0.0353	3.6	1.1	3.5	1.1
Reading	6	4545	B.1.1.1	2	3983	0.67	0.21	0.67	0.07	0.04	0.00	0.38	-0.29	0.38	-0.30	-0.32	-0.1838	0.0372	-0.8	1.0	-1.1	1.0

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Reading	6	4868	B.3.1.1	1	3983	0.54	0.17	0.15	0.54	0.13	0.00	0.42	-0.37	-0.36	0.42	-0.35	0.4534	0.0354	-3.7	1.0	-3.1	0.9
Reading	6	5358	A.2.4.1	1	3983	0.49	0.22	0.19	0.10	0.49	0.01	0.33	-0.24	-0.27	-0.46	0.33	0.7386	0.0353	3.4	1.1	2.7	1.1
Reading	6	5505	B.1.1.1	2	3983	0.48	0.12	0.15	0.25	0.48	0.00	0.39	-0.32	-0.29	-0.42	0.39	0.7803	0.0353	-1.5	1.0	-0.5	1.0
Reading	6	6142	A.2.4.1	2	3983	0.65	0.10	0.65	0.16	0.09	0.01	0.36	-0.32	0.36	-0.24	-0.33	-0.0858	0.0368	0.5	1.0	-0.1	1.0
Reading	6	7092	A.1.3.1	2	3983	0.76	0.05	0.02	0.76	0.17	0.00	0.30	-0.27	-0.23	0.30	-0.21	-0.7273	0.0405	1.4	1.0	2.0	1.1
Reading	6	7233	B.3.1.1	1	3983	0.67	0.67	0.12	0.12	0.09	0.00	0.47	0.47	-0.38	-0.37	-0.35	-0.1716	0.0372	-6.2	0.9	-6.1	0.9
Reading	6	7344	A.1.2.1	1	3983	0.67	0.67	0.12	0.11	0.09	0.00	0.40	0.40	-0.32	-0.33	-0.27	-0.2056	0.0373	-1.8	1.0	-2.6	0.9
Reading	6	7460	A.2.4.1	2	3983	0.42	0.42	0.16	0.35	0.07	0.01	0.26	0.26	-0.26	-0.21	-0.39	1.0743	0.0356	7.2	1.1	7.0	1.2
Reading	6	7469	A.2.2.2	2	3983	0.70	0.70	0.21	0.05	0.04	0.01	0.40	0.40	-0.30	-0.36	-0.32	-0.3248	0.0379	-2.2	1.0	-2.2	0.9
Reading	6	8166	B.2.1.2	1	3983	0.68	0.09	0.10	0.12	0.68	0.00	0.48	-0.36	-0.41	-0.36	0.48	-0.2509	0.0375	-7.2	0.9	-6.2	0.8
Reading	6	8349	B.2.1.2	1	3983	0.71	0.10	0.71	0.14	0.05	0.00	0.40	-0.24	0.40	-0.39	-0.24	-0.4003	0.0383	-2.3	1.0	-2.5	0.9
Reading	6	8577	A.2.3.1	2	3983	0.70	0.11	0.70	0.12	0.06	0.00	0.43	-0.33	0.43	-0.35	-0.29	-0.3745	0.0382	-3.9	0.9	-4.3	0.9
Reading	6	8600	B.3.2.1	2	3983	0.36	0.18	0.36	0.28	0.19	0.00	0.14	-0.06	0.14	-0.18	-0.14	1.3832	0.0365	9.9	1.2	9.9	1.4
Reading	6	9048	A.2.4.1	1	3983	0.71	0.71	0.11	0.06	0.12	0.00	0.43	0.43	-0.33	-0.33	-0.34	-0.4249	0.0385	-4.3	0.9	-4.8	0.9
Reading	6	9132	A.2.4.1	1	3983	0.76	0.05	0.76	0.12	0.07	0.00	0.37	-0.27	0.37	-0.27	-0.26	-0.7273	0.0405	-1.5	1.0	-1.6	0.9
Reading	6	9562	A.2.3.2	2	3983	0.51	0.28	0.51	0.12	0.09	0.01	0.28	-0.25	0.28	-0.32	-0.15	0.6164	0.0353	6.8	1.1	5.7	1.1
Reading	6	9648	A.2.2.2	2	3983	0.73	0.12	0.73	0.06	0.08	0.00	0.48	-0.41	0.48	-0.34	-0.31	-0.5373	0.0392	-7.1	0.9	-6.9	0.8
Reading	6	9653	A.1.1.2	1	3983	0.72	0.17	0.05	0.05	0.72	0.00	0.35	-0.21	-0.34	-0.30	0.35	-0.4643	0.0387	0.0	1.0	1.9	1.1
Reading	7	0514	B.3.3.4	1	3974	0.43	0.22	0.17	0.18	0.43	0.00	0.18	-0.12	-0.22	-0.19	0.18	0.7347	0.0349	9.9	1.1	9.2	1.2
Reading	7	0845	A.1.3.1	2	3974	0.57	0.15	0.20	0.57	0.07	0.00	0.38	-0.33	-0.32	0.38	-0.26	0.0663	0.0349	-3.4	1.0	-4.2	0.9
Reading	7	1012	B.1.1.1	2	3974	0.72	0.06	0.13	0.72	0.09	0.00	0.50	-0.39	-0.38	0.50	-0.36	-0.7013	0.0381	-9.9	0.8	-9.9	0.7
Reading	7	1132	B.1.1.1	2	3974	0.59	0.17	0.14	0.09	0.59	0.00	0.27	-0.16	-0.25	-0.26	0.27	-0.0285	0.0352	4.2	1.1	3.6	1.1
Reading	7	1616	B.3.3.1	2	3974	0.77	0.77	0.09	0.07	0.07	0.00	0.44	0.44	-0.30	-0.37	-0.28	-0.9841	0.0402	-7.0	0.9	-7.1	0.8
Reading	7	2002	A.2.4.1	2	3974	0.62	0.09	0.62	0.18	0.11	0.00	0.35	-0.34	0.35	-0.26	-0.24	-0.1653	0.0355	-1.5	1.0	-2.1	1.0
Reading	7	2250	A.2.4.1	1	3974	0.52	0.11	0.08	0.52	0.29	0.00	0.35	-0.29	-0.39	0.35	-0.27	0.3109	0.0346	-1.4	1.0	-0.8	1.0
Reading	7	2464	B.1.1.1	2	3974	0.79	0.09	0.79	0.08	0.04	0.00	0.40	-0.29	0.40	-0.28	-0.27	-1.0886	0.0411	-5.2	0.9	-5.5	0.8
Reading	7	2618	A.1.3.1	2	3974	0.38	0.31	0.13	0.18	0.38	0.01	0.31	-0.28	-0.41	-0.27	0.31	0.9833	0.0355	-0.6	1.0	0.8	1.0
Reading	7	3389	B.1.1.1	2	3974	0.39	0.12	0.32	0.16	0.39	0.01	0.30	-0.40	-0.21	-0.34	0.30	0.9157	0.0353	0.9	1.0	1.9	1.0
Reading	7	3490	B.2.1.1	1	3974	0.58	0.58	0.12	0.15	0.14	0.00	0.27	0.27	-0.25	-0.19	-0.22	0.0160	0.0350	4.4	1.1	4.4	1.1
Reading	7	3794	B.2.2.1	2	3974	0.59	0.59	0.12	0.07	0.22	0.00	0.24	0.24	-0.35	-0.27	-0.06	0.0040	0.0351	5.9	1.1	5.8	1.1
Reading	7	4530	A.1.3.1	2	3974	0.56	0.23	0.56	0.14	0.07	0.00	0.39	-0.35	0.39	-0.32	-0.30	0.1365	0.0348	-4.1	1.0	-4.1	0.9
Reading	7	4891	B.1.1.1	2	3974	0.45	0.45	0.25	0.23	0.06	0.00	0.27	0.27	-0.25	-0.20	-0.36	0.6279	0.0347	4.1	1.1	4.0	1.1
Reading	7	5018	A.2.6.1	2	3974	0.38	0.36	0.15	0.11	0.38	0.00	0.29	-0.23	-0.29	-0.40	0.29	0.9858	0.0355	1.1	1.0	2.4	1.1
Reading	7	5241	A.1.1.2	1	3974	0.75	0.12	0.06	0.06	0.75	0.00	0.41	-0.33	-0.27	-0.28	0.41	-0.8675	0.0393	-5.8	0.9	-5.1	0.9
Reading	7	5560	A.2.1.1	2	3974	0.41	0.21	0.17	0.41	0.21	0.00	0.28	-0.22	-0.24	0.28	-0.33	0.8222	0.0351	2.2	1.0	3.8	1.1
Reading	7	6050	B.2.1.2	2	3974	0.75	0.75	0.07	0.12	0.05	0.00	0.45	0.45	-0.36	-0.32	-0.28	-0.8796	0.0394	-7.2	0.9	-7.9	0.8
Reading	7	6197	B.2.1.1	2	3974	0.63	0.13	0.63	0.11	0.12	0.01	0.39	-0.36	0.39	-0.35	-0.21	-0.2149	0.0357	-4.2	0.9	-4.0	0.9
Reading	7	6779	A.2.4.1	2	3974	0.61	0.10	0.19	0.10	0.61	0.00	0.40	-0.38	-0.23	-0.42	0.40	-0.0990	0.0353	-4.9	0.9	-4.4	0.9
Reading	7	6850	A.1.5.1	2	3974	0.37	0.37	0.32	0.16	0.15	0.01	0.22	0.22	-0.18	-0.26	-0.22	1.0406	0.0357	5.2	1.1	8.0	1.2
Reading	7	7065	A.2.4.1	1	3974	0.55	0.20	0.21	0.04	0.55	0.00	0.32	-0.19	-0.33	-0.31	0.32	0.1862	0.0348	1.1	1.0	0.9	1.0
Reading	7	7532	A.1.3.1	2	3974	0.56	0.18	0.12	0.14	0.56	0.01	0.42	-0.33	-0.31	-0.43	0.42	0.1401	0.0348	-6.4	0.9	-6.5	0.9
Reading	7	7623	A.1.4.1	2	3974	0.91	0.05	0.91	0.02	0.02	0.00	0.38	-0.27	0.38	-0.23	-0.23	-2.1356	0.0555	-4.8	0.8	-7.8	0.6

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Reading	7	7934	B.1.1.1	2	3974	0.57	0.15	0.18	0.57	0.09	0.01	0.37	-0.31	-0.31	0.37	-0.28	0.0663	0.0349	-2.5	1.0	-2.3	1.0
Reading	7	7965	B.1.1.1	2	3974	0.66	0.17	0.66	0.10	0.06	0.00	0.29	-0.22	0.29	-0.20	-0.27	-0.3839	0.0364	1.5	1.0	1.0	1.0
Reading	7	8001	B.1.1.1	2	3974	0.48	0.12	0.07	0.48	0.32	0.00	0.21	-0.33	-0.37	0.21	-0.06	0.4973	0.0346	8.5	1.1	8.4	1.2
Reading	7	8462	A.2.2.1	1	3974	0.66	0.66	0.17	0.09	0.08	0.00	0.39	0.39	-0.36	-0.34	-0.18	-0.3645	0.0363	-4.4	0.9	-3.4	0.9
Reading	7	9042	B.3.3.3	2	3974	0.65	0.65	0.18	0.04	0.12	0.00	0.18	0.18	-0.11	-0.24	-0.11	-0.3222	0.0361	8.2	1.1	8.4	1.2
Reading	7	9865	A.2.4.1	1	3974	0.44	0.09	0.43	0.44	0.05	0.00	0.20	-0.33	-0.12	0.20	-0.30	0.7014	0.0348	8.5	1.1	8.5	1.2
Reading	8	0071	B.2.2.2	2	3647	0.65	0.15	0.65	0.12	0.08	0.00	0.46	-0.36	0.46	-0.39	-0.34	-0.2368	0.0376	-8.6	0.9	-8.7	0.8
Reading	8	0278	B.2.1.2	2	3647	0.59	0.18	0.59	0.11	0.11	0.00	0.20	-0.13	0.20	-0.20	-0.14	0.0594	0.0366	8.3	1.1	6.9	1.1
Reading	8	0486	B.3.1.1	2	3647	0.60	0.15	0.60	0.14	0.10	0.01	0.37	-0.26	0.37	-0.34	-0.28	0.0111	0.0367	-2.7	1.0	-3.2	0.9
Reading	8	0977	A.2.1.2	2	3647	0.56	0.22	0.15	0.06	0.56	0.00	0.39	-0.38	-0.29	-0.32	0.39	0.1899	0.0363	-4.7	0.9	-4.1	0.9
Reading	8	1572	B.1.2.1	3	3647	0.37	0.17	0.26	0.21	0.37	0.00	0.23	-0.26	-0.23	-0.19	0.23	1.1036	0.0372	4.1	1.1	5.4	1.1
Reading	8	1652	B.1.2.1	3	3647	0.50	0.18	0.50	0.21	0.11	0.00	0.28	-0.20	0.28	-0.27	-0.29	0.4633	0.0361	2.8	1.0	2.6	1.1
Reading	8	2420	A.2.6.1	2	3647	0.48	0.48	0.32	0.09	0.11	0.00	0.30	0.30	-0.20	-0.41	-0.29	0.5800	0.0361	1.4	1.0	1.6	1.0
Reading	8	2779	A.1.1.2	1	3647	0.68	0.68	0.14	0.09	0.09	0.00	0.28	0.28	-0.28	-0.23	-0.11	-0.3766	0.0382	1.5	1.0	0.8	1.0
Reading	8	2893	A.2.3.2	2	3647	0.51	0.16	0.12	0.21	0.51	0.00	0.36	-0.19	-0.43	-0.34	0.36	0.4127	0.0361	-2.5	1.0	-2.1	1.0
Reading	8	2934	A.2.2.1	1	3647	0.85	0.07	0.85	0.06	0.03	0.00	0.34	-0.22	0.34	-0.24	-0.22	-1.4457	0.0478	-3.5	0.9	-4.5	0.8
Reading	8	3252	A.2.4.1	2	3647	0.57	0.15	0.19	0.57	0.08	0.00	0.07	-0.09	0.02	0.07	-0.17	0.1449	0.0364	9.9	1.2	9.9	1.3
Reading	8	3600	A.1.4.1	2	3647	0.80	0.80	0.08	0.05	0.07	0.00	0.44	0.44	-0.37	-0.28	-0.27	-1.0858	0.0436	-6.6	0.9	-8.2	0.7
Reading	8	3929	B.1.1.1	2	3647	0.61	0.09	0.61	0.09	0.20	0.00	0.37	-0.29	0.37	-0.36	-0.26	-0.0641	0.0369	-2.7	1.0	-3.1	0.9
Reading	8	4777	A.2.1.2	2	3647	0.55	0.08	0.19	0.18	0.55	0.00	0.39	-0.30	-0.45	-0.24	0.39	0.2474	0.0362	-4.5	0.9	-3.7	0.9
Reading	8	4782	A.1.3.1	2	3647	0.76	0.76	0.05	0.13	0.06	0.00	0.30	0.30	-0.27	-0.20	-0.21	-0.8670	0.0415	-1.4	1.0	0.2	1.0
Reading	8	6098	A.1.3.1	2	3647	0.68	0.07	0.07	0.68	0.17	0.00	0.25	-0.23	-0.28	0.25	-0.12	-0.3866	0.0383	2.9	1.1	3.5	1.1
Reading	8	6430	A.2.4.1	1	3647	0.47	0.23	0.47	0.13	0.16	0.00	0.28	-0.22	0.28	-0.26	-0.28	0.6296	0.0361	2.5	1.0	3.0	1.1
Reading	8	6446	A.2.3.1	2	3647	0.61	0.14	0.07	0.61	0.18	0.00	0.35	-0.27	-0.31	0.35	-0.28	-0.0390	0.0368	-1.9	1.0	-2.6	1.0
Reading	8	6457	B.3.3.1	2	3647	0.48	0.11	0.22	0.48	0.18	0.00	0.24	-0.31	-0.13	0.24	-0.25	0.5673	0.0361	5.9	1.1	6.5	1.1
Reading	8	6847	B.1.1.1	3	3647	0.59	0.04	0.09	0.59	0.28	0.00	0.32	-0.36	-0.22	0.32	-0.27	0.0594	0.0366	0.2	1.0	-0.1	1.0
Reading	8	7099	B.2.1.1	1	3647	0.50	0.24	0.50	0.15	0.11	0.00	0.21	-0.17	0.21	-0.17	-0.22	0.4811	0.0361	7.7	1.1	8.1	1.1
Reading	8	7522	A.1.2.2	2	3647	0.72	0.72	0.10	0.10	0.09	0.00	0.41	0.41	-0.27	-0.36	-0.30	-0.5852	0.0394	-5.5	0.9	-5.5	0.9
Reading	8	7573	A.1.4.1	1	3647	0.38	0.43	0.38	0.08	0.11	0.00	0.13	-0.07	0.13	-0.26	-0.18	1.0228	0.0369	9.9	1.2	9.9	1.2
Reading	8	8017	B.3.1.1	2	3647	0.47	0.12	0.26	0.15	0.47	0.00	0.30	-0.36	-0.18	-0.32	0.30	0.6245	0.0361	1.5	1.0	2.1	1.0
Reading	8	8365	A.2.4.1	1	3647	0.84	0.02	0.84	0.07	0.07	0.00	0.40	-0.22	0.40	-0.32	-0.25	-1.3830	0.0470	-5.0	0.9	-6.4	0.7
Reading	8	8398	B.2.1.2	2	3647	0.75	0.07	0.05	0.13	0.75	0.00	0.43	-0.29	-0.24	-0.38	0.43	-0.7903	0.0409	-6.3	0.9	-6.7	0.8
Reading	8	8603	B.1.1.1	3	3647	0.62	0.09	0.09	0.62	0.20	0.00	0.29	-0.31	-0.37	0.29	-0.11	-0.1107	0.0371	1.6	1.0	1.7	1.0
Reading	8	9638	A.2.3.1	2	3647	0.57	0.57	0.08	0.20	0.15	0.00	0.42	0.42	-0.42	-0.29	-0.41	0.1616	0.0364	-6.4	0.9	-5.3	0.9
Reading	8	9681	B.2.1.1	2	3647	0.58	0.19	0.17	0.58	0.06	0.00	0.34	-0.25	-0.32	0.34	-0.28	0.1022	0.0365	-0.9	1.0	-1.0	1.0
Reading	8	9984	B.3.3.4	1	3647	0.49	0.20	0.19	0.11	0.49	0.00	0.36	-0.30	-0.36	-0.30	0.36	0.5102	0.0361	-2.4	1.0	-2.2	1.0
Reading	11	0852	A.2.4.1	1	3919	0.53	0.06	0.53	0.29	0.11	0.00	0.38	-0.43	0.38	-0.29	-0.36	0.6352	0.0350	-3.6	1.0	-1.9	1.0
Reading	11	1153	A.2.4.1	1	3919	0.79	0.02	0.13	0.79	0.06	0.00	0.33	-0.20	-0.29	0.33	-0.19	-0.7750	0.0423	-1.5	1.0	-0.6	1.0
Reading	11	1674	B.3.3.3	2	3919	0.78	0.78	0.10	0.09	0.02	0.01	0.42	0.42	-0.33	-0.31	-0.25	-0.6954	0.0416	-5.0	0.9	-5.6	0.8
Reading	11	1919	B.2.1.2	2	3919	0.77	0.05	0.08	0.77	0.09	0.00	0.38	-0.28	-0.26	0.38	-0.28	-0.6401	0.0411	-2.7	0.9	-4.6	0.9
Reading	11	2067	A.2.6.1	2	3919	0.57	0.57	0.15	0.10	0.18	0.00	0.17	0.17	-0.20	-0.22	-0.03	0.4613	0.0353	9.9	1.2	9.7	1.2
Reading	11	3239	B.2.1.1	2	3919	0.55	0.20	0.18	0.07	0.55	0.00	0.35	-0.27	-0.31	-0.41	0.35	0.5635	0.0351	-1.4	1.0	-2.2	1.0

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Reading	11	3750	B.3.1.1	2	3919	0.73	0.17	0.08	0.73	0.02	0.00	0.34	-0.24	-0.30	0.34	-0.25	-0.3693	0.0390	-1.1	1.0	-2.0	0.9
Reading	11	4083	B.1.1.1	2	3919	0.50	0.26	0.11	0.13	0.50	0.00	0.33	-0.22	-0.37	-0.39	0.33	0.8013	0.0349	-0.2	1.0	-0.5	1.0
Reading	11	4426	B.1.1.1	2	3919	0.57	0.04	0.57	0.06	0.32	0.00	0.25	-0.26	0.25	-0.24	-0.19	0.4770	0.0352	5.8	1.1	5.2	1.1
Reading	11	4573	A.1.3.1	2	3919	0.55	0.05	0.29	0.11	0.55	0.00	0.28	-0.35	-0.16	-0.35	0.28	0.5371	0.0351	3.7	1.1	2.7	1.1
Reading	11	4752	B.3.2.2	2	3919	0.52	0.52	0.25	0.11	0.11	0.00	0.40	0.40	-0.29	-0.46	-0.40	0.6744	0.0349	-6.0	0.9	-4.7	0.9
Reading	11	5040	B.3.3.1	2	3919	0.64	0.64	0.14	0.06	0.15	0.00	0.40	0.40	-0.32	-0.39	-0.30	0.1135	0.0363	-4.3	0.9	-4.3	0.9
Reading	11	5073	B.2.1.1	2	3919	0.50	0.12	0.06	0.50	0.31	0.00	0.20	-0.29	-0.34	0.20	-0.06	0.7693	0.0349	9.1	1.1	8.6	1.2
Reading	11	5209	A.1.4.1	3	3919	0.59	0.14	0.59	0.17	0.09	0.00	0.28	-0.24	0.28	-0.18	-0.29	0.3615	0.0355	3.5	1.1	3.3	1.1
Reading	11	5320	B.3.1.1	2	3919	0.62	0.62	0.13	0.13	0.11	0.00	0.28	0.28	-0.25	-0.21	-0.22	0.1949	0.0360	3.2	1.1	3.5	1.1
Reading	11	5772	A.2.4.1	1	3919	0.65	0.06	0.06	0.23	0.65	0.01	0.30	-0.41	-0.44	-0.10	0.30	0.0696	0.0365	1.7	1.0	2.5	1.1
Reading	11	6038	B.1.2.1	3	3919	0.45	0.26	0.45	0.12	0.16	0.00	0.24	-0.25	0.24	-0.34	-0.10	0.9947	0.0349	5.0	1.1	6.4	1.1
Reading	11	6160	A.2.4.1	2	3919	0.64	0.18	0.07	0.10	0.64	0.00	0.39	-0.35	-0.43	-0.16	0.39	0.0929	0.0364	-3.2	1.0	-3.0	0.9
Reading	11	6653	A.2.3.1	2	3919	0.69	0.07	0.69	0.15	0.09	0.00	0.32	-0.19	0.32	-0.20	-0.36	-0.1383	0.0375	0.3	1.0	0.6	1.0
Reading	11	6838	A.1.3.1	2	3919	0.67	0.04	0.10	0.67	0.18	0.00	0.31	-0.29	-0.29	0.31	-0.18	-0.0570	0.0371	1.3	1.0	0.0	1.0
Reading	11	7391	B.1.2.1	3	3919	0.67	0.12	0.67	0.11	0.09	0.00	0.19	-0.16	0.19	-0.16	-0.11	-0.0624	0.0371	7.4	1.1	7.1	1.2
Reading	11	7430	B.2.1.2	2	3919	0.80	0.05	0.07	0.80	0.09	0.00	0.41	-0.24	-0.31	0.41	-0.30	-0.7839	0.0424	-4.0	0.9	-5.8	0.8
Reading	11	7723	A.2.2.2	2	3919	0.66	0.05	0.21	0.66	0.08	0.00	0.37	-0.30	-0.35	0.37	-0.20	0.0160	0.0367	-2.3	1.0	-2.5	0.9
Reading	11	8326	A.2.2.1	2	3919	0.75	0.10	0.10	0.05	0.75	0.00	0.45	-0.35	-0.34	-0.32	0.45	-0.4768	0.0397	-6.3	0.9	-6.6	0.8
Reading	11	8934	A.2.4.1	2	3919	0.55	0.15	0.16	0.55	0.13	0.00	0.35	-0.32	-0.30	0.35	-0.29	0.5408	0.0351	-1.4	1.0	-0.8	1.0
Reading	11	9068	B.2.1.1	2	3919	0.71	0.71	0.07	0.12	0.10	0.00	0.35	0.35	-0.37	-0.15	-0.32	-0.2374	0.0381	-1.6	1.0	-1.5	1.0
Reading	11	9404	A.1.4.1	2	3919	0.91	0.91	0.05	0.03	0.01	0.00	0.37	0.37	-0.27	-0.25	-0.18	-1.8238	0.0573	-3.9	0.8	-6.1	0.6
Reading	11	9567	A.2.2.1	1	3919	0.86	0.04	0.06	0.04	0.86	0.00	0.47	-0.32	-0.33	-0.32	0.47	-1.2596	0.0480	-6.7	0.8	-8.5	0.6
Reading	11	9868	A.2.2.2	2	3919	0.83	0.03	0.10	0.04	0.83	0.00	0.40	-0.27	-0.29	-0.29	0.40	-1.0325	0.0451	-4.5	0.9	-4.4	0.8
Reading	11	9880	A.1.1.1	2	3919	0.44	0.11	0.44	0.17	0.27	0.01	0.09	-0.13	0.09	-0.12	-0.04	1.0484	0.0350	9.9	1.2	9.9	1.3
Science	8	0116	A.1.2.3	2	3252	0.50	0.15	0.22	0.14	0.50	0.00	0.34	-0.23	-0.32	-0.32	0.34	0.6558	0.0381	-2.2	1.0	-1.8	1.0
Science	8	0130	A.3.2.2	2	3252	0.36	0.21	0.27	0.15	0.36	0.00	0.22	-0.21	-0.22	-0.24	0.22	1.3210	0.0396	3.7	1.1	4.4	1.1
Science	8	0269	A.2.1.5	2	3252	0.64	0.17	0.09	0.64	0.11	0.00	0.36	-0.28	-0.27	0.36	-0.28	0.0058	0.0393	-3.1	1.0	-3.6	0.9
Science	8	0463	C.2.2.1	1	3252	0.82	0.04	0.10	0.82	0.04	0.00	0.30	-0.18	-0.23	0.30	-0.18	-1.0426	0.0476	-1.8	1.0	-2.1	0.9
Science	8	0498	A.1.1.2	2	3252	0.69	0.06	0.06	0.19	0.69	0.00	0.26	-0.20	-0.24	-0.18	0.26	-0.2823	0.0408	1.2	1.0	2.1	1.1
Science	8	1026	A.1.3.4	2	3252	0.76	0.11	0.76	0.08	0.06	0.00	0.44	-0.31	0.44	-0.35	-0.30	-0.6556	0.0435	-6.6	0.9	-7.1	0.8
Science	8	1070	D.1.3.2	2	3252	0.59	0.14	0.59	0.20	0.07	0.00	0.34	-0.34	0.34	-0.26	-0.21	0.2384	0.0385	-2.1	1.0	-1.7	1.0
Science	8	1215	D.3.1.1	2	3252	0.54	0.54	0.11	0.27	0.08	0.00	0.29	0.29	-0.28	-0.22	-0.25	0.4384	0.0382	1.5	1.0	1.4	1.0
Science	8	1240	D.3.1.2	1	3252	0.66	0.66	0.16	0.10	0.08	0.00	0.17	0.17	-0.03	-0.22	-0.18	-0.1060	0.0398	6.4	1.1	5.2	1.1
Science	8	2853	A.3.1.2	2	3252	0.73	0.73	0.06	0.16	0.06	0.00	0.28	0.28	-0.25	-0.18	-0.23	-0.4600	0.0419	-0.4	1.0	-0.2	1.0
Science	8	3233	B.1.1.2	2	3252	0.89	0.05	0.89	0.04	0.02	0.00	0.27	-0.16	0.27	-0.20	-0.17	-1.6581	0.0571	-2.0	0.9	-2.8	0.8
Science	8	3394	B.2.1.4	2	3252	0.62	0.62	0.15	0.14	0.09	0.00	0.35	0.35	-0.26	-0.30	-0.28	0.0824	0.0390	-2.8	1.0	-3.3	0.9
Science	8	3833	A.3.1.3	1	3252	0.57	0.18	0.13	0.57	0.12	0.00	0.23	-0.21	-0.19	0.23	-0.17	0.3252	0.0384	4.8	1.1	3.2	1.1
Science	8	4049	B.1.1.1	2	3252	0.70	0.70	0.09	0.19	0.01	0.00	0.20	0.20	-0.14	-0.17	-0.12	-0.3068	0.0409	3.9	1.1	3.4	1.1
Science	8	4141	A.2.2.1	2	3252	0.92	0.02	0.92	0.02	0.04	0.00	0.30	-0.14	0.30	-0.17	-0.23	-2.0057	0.0644	-2.5	0.9	-4.0	0.7
Science	8	4333	A.3.2.3	3	3252	0.47	0.18	0.21	0.47	0.14	0.00	0.24	-0.28	-0.19	0.24	-0.18	0.7981	0.0381	4.4	1.1	4.9	1.1
Science	8	4444	B.1.1.2	2	3252	0.63	0.11	0.63	0.20	0.06	0.00	0.33	-0.23	0.33	-0.27	-0.29	0.0285	0.0392	-1.6	1.0	-2.2	1.0
Science	8	4733	C.2.1.3	2	3252	0.55	0.07	0.18	0.20	0.55	0.00	0.34	-0.29	-0.34	-0.25	0.34	0.4198	0.0382	-2.2	1.0	-2.1	1.0

Appendix I: Item Statistics Multiple Choice

Item Information					Classical												Rasch		Infit		Outfit	
Cont	Grade	PubID	Std	DOK	N	PVal	P(A)	P(B)	P(C)	P(D)	P(-)	PtBis	PT(A)	PT(B)	PT(C)	PT(D)	Meas	MeasSE	t	MS	t	MS
Science	8	5084	C.1.1.2	2	3252	0.43	0.43	0.43	0.08	0.06	0.00	0.22	0.22	-0.18	-0.25	-0.24	0.9547	0.0384	5.4	1.1	5.4	1.1
Science	8	5709	B.2.2.2	1	3252	0.52	0.14	0.52	0.25	0.09	0.00	0.32	-0.25	0.32	-0.29	-0.28	0.5366	0.0381	-0.8	1.0	-0.8	1.0
Science	8	5800	A.1.1.2	2	3252	0.67	0.07	0.67	0.16	0.10	0.01	0.33	-0.19	0.33	-0.33	-0.22	-0.1465	0.0400	-2.0	1.0	-1.7	1.0
Science	8	5950	A.2.2.3	2	3252	0.87	0.02	0.07	0.04	0.87	0.00	0.32	-0.20	-0.18	-0.28	0.32	-1.4939	0.0541	-3.0	0.9	-2.8	0.8
Science	8	6439	A.1.1.1	2	3252	0.76	0.76	0.08	0.10	0.06	0.00	0.38	0.38	-0.21	-0.34	-0.25	-0.6630	0.0436	-4.4	0.9	-4.1	0.9
Science	8	6726	C.2.1.2	2	3252	0.37	0.30	0.23	0.37	0.09	0.00	0.24	-0.26	-0.12	0.24	-0.38	1.2445	0.0392	2.6	1.0	5.1	1.1
Science	8	7225	D.1.2.1	1	3252	0.72	0.06	0.13	0.09	0.72	0.00	0.35	-0.25	-0.27	-0.24	0.35	-0.4497	0.0419	-2.9	1.0	-3.9	0.9
Science	8	7711	C.3.1.3	2	3252	0.49	0.22	0.17	0.12	0.49	0.00	0.39	-0.32	-0.40	-0.36	0.39	0.7084	0.0381	-5.9	0.9	-4.9	0.9
Science	8	8212	A.2.2.1	2	3252	0.62	0.18	0.12	0.08	0.62	0.00	0.30	-0.20	-0.25	-0.28	0.30	0.0779	0.0390	0.2	1.0	-0.4	1.0
Science	8	8847	A.3.1.3	1	3252	0.43	0.43	0.36	0.12	0.08	0.00	0.34	0.34	-0.27	-0.37	-0.40	0.9561	0.0384	-2.8	1.0	-1.4	1.0
Science	8	9015	A.3.1.2	2	3252	0.48	0.14	0.16	0.48	0.22	0.01	0.17	-0.19	-0.19	0.17	-0.10	0.7511	0.0381	8.9	1.1	9.0	1.2
Science	8	9719	D.1.3.1	2	3252	0.69	0.69	0.12	0.10	0.09	0.00	0.37	0.37	-0.25	-0.32	-0.29	-0.2725	0.0407	-4.1	0.9	-4.0	0.9
Science	11	0951	A.1.1.5	2	3540	0.62	0.12	0.15	0.62	0.10	0.00	0.32	-0.23	-0.28	0.32	-0.21	-0.3869	0.0373	-1.5	1.0	-2.1	1.0
Science	11	0987	A.3.2.3	2	3540	0.37	0.16	0.37	0.26	0.20	0.00	0.26	-0.19	0.26	-0.30	-0.24	0.7990	0.0375	1.0	1.0	3.2	1.1
Science	11	1355	B.3.3.3	3	3540	0.29	0.29	0.29	0.25	0.17	0.00	0.22	0.22	-0.15	-0.28	-0.28	1.1909	0.0395	1.5	1.0	4.6	1.1
Science	11	1573	A.2.2.1	2	3540	0.66	0.07	0.11	0.15	0.66	0.00	0.44	-0.31	-0.38	-0.32	0.44	-0.5841	0.0381	-9.1	0.9	-8.4	0.8
Science	11	1613	D.3.1.2	2	3540	0.35	0.35	0.18	0.31	0.16	0.01	0.23	0.23	-0.27	-0.21	-0.22	0.9057	0.0379	2.9	1.1	3.6	1.1
Science	11	2292	B.3.2.3	2	3540	0.73	0.05	0.12	0.09	0.73	0.00	0.43	-0.29	-0.35	-0.30	0.43	-0.9467	0.0402	-8.2	0.9	-8.3	0.8
Science	11	2740	A.1.2.1	2	3540	0.54	0.54	0.07	0.08	0.31	0.00	0.32	0.32	-0.31	-0.32	-0.23	0.0126	0.0364	-1.2	1.0	-1.5	1.0
Science	11	2903	A.2.2.1	2	3540	0.48	0.32	0.48	0.17	0.03	0.00	0.25	-0.22	0.25	-0.20	-0.25	0.2775	0.0363	3.7	1.1	4.0	1.1
Science	11	3043	D.1.1.2	2	3540	0.50	0.17	0.50	0.11	0.22	0.00	0.30	-0.19	0.30	-0.37	-0.24	0.1956	0.0363	0.1	1.0	0.2	1.0
Science	11	3070	D.1.3.2	2	3540	0.33	0.33	0.31	0.21	0.15	0.00	0.21	0.21	-0.17	-0.26	-0.23	0.9952	0.0384	3.3	1.1	4.2	1.1
Science	11	3315	C.3.1.2	2	3540	0.67	0.05	0.67	0.12	0.15	0.00	0.18	-0.22	0.18	-0.22	-0.02	-0.6310	0.0383	4.4	1.1	7.0	1.2
Science	11	3604	A.3.1.1	3	3540	0.74	0.08	0.09	0.09	0.74	0.00	0.37	-0.21	-0.31	-0.26	0.37	-0.9900	0.0406	-5.3	0.9	-4.7	0.9
Science	11	3771	C.1.1.1	2	3540	0.67	0.05	0.11	0.67	0.16	0.00	0.37	-0.26	-0.29	0.37	-0.28	-0.6325	0.0383	-5.0	0.9	-4.9	0.9
Science	11	3814	A.1.3.2	2	3540	0.52	0.12	0.10	0.26	0.52	0.00	0.38	-0.36	-0.42	-0.26	0.38	0.1074	0.0363	-5.1	0.9	-4.0	0.9
Science	11	4097	A.1.1.2	2	3540	0.58	0.09	0.19	0.13	0.58	0.00	0.34	-0.22	-0.27	-0.32	0.34	-0.1774	0.0367	-2.6	1.0	-2.5	1.0
Science	11	4127	B.3.1.4	2	3540	0.59	0.20	0.59	0.12	0.10	0.00	0.11	-0.07	0.11	-0.08	-0.12	-0.2036	0.0367	9.9	1.2	9.9	1.2
Science	11	4734	D.3.1.2	1	3540	0.41	0.26	0.12	0.20	0.41	0.00	0.15	-0.12	-0.25	-0.10	0.15	0.6147	0.0369	9.1	1.1	8.1	1.2
Science	11	4926	A.1.3.4	2	3540	0.73	0.12	0.10	0.73	0.05	0.00	0.37	-0.23	-0.33	0.37	-0.24	-0.9230	0.0401	-5.3	0.9	-4.8	0.9
Science	11	5093	B.3.1.3	2	3540	0.49	0.49	0.05	0.16	0.29	0.00	0.32	0.32	-0.34	-0.24	-0.28	0.2314	0.0363	-1.2	1.0	-1.3	1.0
Science	11	5811	A.1.3.1	3	3540	0.50	0.18	0.13	0.19	0.50	0.00	0.37	-0.26	-0.42	-0.32	0.37	0.1790	0.0363	-4.6	0.9	-4.2	0.9
Science	11	6750	A.1.1.3	2	3540	0.77	0.06	0.11	0.06	0.77	0.00	0.21	-0.19	-0.10	-0.17	0.21	-1.1469	0.0419	0.3	1.0	3.1	1.1
Science	11	6803	B.3.1.2	2	3540	0.49	0.49	0.31	0.16	0.04	0.00	0.27	0.27	-0.33	-0.13	-0.07	0.2276	0.0363	2.5	1.0	2.6	1.0
Science	11	6996	A.1.3.3	2	3540	0.59	0.16	0.59	0.15	0.10	0.00	0.33	-0.30	0.33	-0.16	-0.35	-0.2115	0.0368	-2.2	1.0	-1.3	1.0
Science	11	7191	A.3.1.2	2	3540	0.59	0.09	0.25	0.59	0.07	0.00	0.41	-0.27	-0.41	0.41	-0.20	-0.2141	0.0368	-7.3	0.9	-5.8	0.9
Science	11	8280	B.3.2.2	2	3540	0.46	0.46	0.23	0.12	0.19	0.00	0.28	0.28	-0.25	-0.36	-0.17	0.3699	0.0364	1.6	1.0	1.7	1.0
Science	11	8493	D.1.1.3	1	3540	0.39	0.17	0.22	0.39	0.22	0.00	0.16	-0.15	-0.11	0.16	-0.21	0.7054	0.0372	7.5	1.1	7.6	1.2
Science	11	8623	A.2.1.4	2	3540	0.53	0.07	0.53	0.21	0.18	0.00	0.30	-0.24	0.30	-0.29	-0.21	0.0485	0.0364	0.4	1.0	-0.4	1.0
Science	11	8804	A.3.1.3	2	3540	0.60	0.08	0.60	0.23	0.09	0.00	0.29	-0.17	0.29	-0.28	-0.21	-0.2575	0.0369	0.0	1.0	-0.6	1.0
Science	11	9100	C.1.1.5	2	3540	0.50	0.21	0.18	0.50	0.11	0.00	0.19	-0.13	-0.21	0.19	-0.13	0.1815	0.0363	7.8	1.1	6.8	1.1
Science	11	9652	C.2.2.3	1	3540	0.48	0.48	0.14	0.14	0.23	0.00	0.42	0.42	-0.42	-0.41	-0.31	0.2634	0.0363	-8.3	0.9	-8.0	0.9

Appendix I: Item Statistics Open Ended

Item Information					Classical														Rasch		Infit		Outfit	
Cont	Grade	PubID	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	1503	C.3	3	2375	1.90	0.16	0.27	0.20	0.28	0.10	0.00	0.50	-0.38	-0.14	0.00	0.26	0.28	1.0722	0.0221	3.1	1.1	2.8	1.1
Math	4	2328	A.3	2	2375	2.42	0.15	0.12	0.23	0.17	0.33	0.00	0.56	-0.42	-0.18	-0.06	0.08	0.43	0.5110	0.0202	0.6	1.0	1.6	1.1
Math	5	5925	E.1		3366	2.63	0.07	0.15	0.18	0.24	0.35	0.00	0.53	-0.34	-0.27	-0.09	0.08	0.39	0.0065	0.0181	1.0	1.0	1.8	1.1
Math	5	9033	A.2	3	3366	1.36	0.40	0.11	0.26	0.17	0.05	0.00	0.56	-0.50	-0.02	0.16	0.35	0.22	1.4861	0.0180	-1.2	1.0	-2.6	0.9
Math	6	2373	D.2		3600	1.37	0.17	0.43	0.29	0.09	0.02	0.01	0.53	-0.40	-0.10	0.21	0.27	0.21	1.0689	0.0221	-2.9	0.9	-3.1	0.9
Math	6	2587	B.2	3	3600	1.32	0.16	0.50	0.25	0.06	0.04	0.00	0.51	-0.34	-0.17	0.26	0.23	0.24	0.9351	0.0223	-1.8	1.0	-1.8	1.0
Math	7	0482	B.1	2	3972	1.46	0.18	0.44	0.21	0.08	0.09	0.00	0.48	-0.27	-0.21	0.18	0.20	0.30	0.7567	0.0177	1.0	1.0	-0.3	1.0
Math	7	2793	C.3	1	3972	2.04	0.13	0.23	0.24	0.24	0.15	0.00	0.52	-0.36	-0.20	0.00	0.21	0.33	0.3100	0.0163	-0.9	1.0	-0.5	1.0
Math	8	3482	B.2	2	4114	0.77	0.59	0.14	0.22	0.02	0.03	0.00	0.52	-0.49	0.10	0.35	0.13	0.24	1.6099	0.0190	-1.3	1.0	-2.2	0.9
Math	8	6548	C.3	2	4114	1.89	0.14	0.22	0.31	0.27	0.06	0.00	0.54	-0.34	-0.22	0.01	0.31	0.30	0.5375	0.0177	-2.6	1.0	-2.2	1.0
Math	11	7025	E.1		4269	1.48	0.18	0.41	0.23	0.13	0.06	0.03	0.56	-0.33	-0.22	0.14	0.31	0.31	0.6721	0.0184	-0.7	1.0	-1.6	1.0
Math	11	9341	C.3	3	4269	0.98	0.43	0.29	0.16	0.12	0.00	0.03	0.65	-0.56	0.06	0.29	0.43	0.08	2.0839	0.0190	-9.9	0.8	-9.9	0.8
Reading	4	7474	A.1.4.1	2	3388	1.78	0.20	0.21	0.19	0.39		0.01	0.55	-0.46	-0.15	0.08	0.44		-0.1970	0.0193	-2.8	0.9	-1.6	1.0
Reading	4	7797	B.1.1.1	2	3388	1.44	0.23	0.26	0.33	0.17		0.01	0.50	-0.42	-0.09	0.21	0.31		0.2923	0.0211	0.3	1.0	-0.1	1.0
Reading	5	0672	A.2.5.1	2	3947	1.41	0.18	0.36	0.31	0.14		0.01	0.53	-0.43	-0.10	0.23	0.32		0.5681	0.0216	1.1	1.0	1.1	1.0
Reading	5	4479	B.1.1.1	3	3947	1.61	0.10	0.33	0.42	0.15		0.01	0.53	-0.39	-0.23	0.26	0.29		0.2296	0.0231	-1.0	1.0	-1.1	1.0
Reading	6	0976	B.1.1.1	3	3983	1.55	0.14	0.35	0.34	0.17		0.01	0.47	-0.35	-0.17	0.19	0.29		0.5523	0.0216	5.0	1.1	4.9	1.1
Reading	6	9782	B.1.1.1	3	3983	1.43	0.12	0.41	0.37	0.09		0.01	0.48	-0.36	-0.18	0.28	0.24		0.7928	0.0237	1.1	1.0	1.0	1.0
Reading	7	1687	B.1.1.1	2	3974	1.38	0.25	0.29	0.30	0.16		0.01	0.49	-0.38	-0.12	0.23	0.32		0.6063	0.0195	0.7	1.0	-0.1	1.0
Reading	7	2591	A.2.5.1	3	3974	1.42	0.12	0.41	0.39	0.08		0.01	0.40	-0.29	-0.16	0.26	0.19		0.5923	0.0236	2.5	1.1	2.4	1.1
Reading	8	7404	A.2.3.1	3	3647	1.92	0.06	0.24	0.41	0.29		0.01	0.47	-0.30	-0.28	0.08	0.33		-0.2806	0.0229	-1.0	1.0	-0.6	1.0
Reading	8	9658	B.1.1.1	3	3647	1.54	0.13	0.35	0.37	0.15		0.01	0.51	-0.37	-0.21	0.24	0.31		0.4080	0.0224	-2.9	0.9	-2.9	0.9
Reading	11	8002	A.2.3.1	2	3919	1.62	0.07	0.43	0.31	0.19		0.02	0.45	-0.29	-0.25	0.17	0.30		0.3821	0.0222	-0.1	1.0	-0.2	1.0
Reading	11	8793	B.1.1.1	3	3919	1.73	0.08	0.35	0.35	0.23		0.02	0.49	-0.33	-0.27	0.15	0.34		0.2756	0.0217	-2.2	1.0	-2.4	1.0
Science	8	4607	A.1.3.2	2	3252	1.08	0.26	0.39	0.35			0.00	0.26	-0.22	-0.01	0.22			0.4584	0.0264	9.1	1.2	8.8	1.2
Science	8	8858	B.2.2.1	2	3252	1.13	0.23	0.42	0.36			0.01	0.46	-0.39	-0.04	0.38			0.3514	0.0270	-3.3	0.9	-3.2	0.9
Science	11	0157	C.2.2.2	3	3540	0.69	0.51	0.29	0.20			0.03	0.49	-0.46	0.17	0.38			0.8191	0.0250	-6.2	0.9	-5.3	0.9
Science	11	3115	A.3.1.2	3	3540	0.99	0.19	0.63	0.18			0.01	0.33	-0.26	0.00	0.26			0.2279	0.0306	0.8	1.0	0.9	1.0

Appendix J:

Linking Item Statistics

Column Heading	Definition
Type	Item type
Form	Form
Seq	Sequence
Prev Form	Previous form
Prev Seq	Previous sequence
Prev P-Val	Previous P-Value
P-Val	P-Value
Prev Meas	Previous Rasch item measure
Meas	Rasch item measure

Appendix J: Linking Item Statistics

Mathematics Grade 4

ID	Type	Form	Seq	Prev	Prev	Prev	P-Val	P-Val	Prev	Meas
				Form	Seq	P-Val			Meas	
560699	MC	0	2	0	2	0.68	0.73	0.1173	-0.2203	
561509	MC	0	3	0	4	0.46	0.35	1.2120	1.6139	
561478	MC	0	4	0	5	0.49	0.47	1.0536	1.0519	
561515	MC	0	6	0	7	0.60	0.56	0.5382	0.6300	
561495	MC	0	7	0	8	0.48	0.46	1.1137	1.1078	
561517	MC	0	8	0	9	0.44	0.40	1.3200	1.3879	
560696	MC	0	9	0	11	0.72	0.73	-0.0594	-0.2519	
560700	MC	0	18	0	15	0.56	0.54	0.7319	0.7160	
561875	MC	0	20	0	21	0.82	0.83	-0.7483	-0.9540	
561756	MC	0	21	0	24	0.67	0.67	0.1626	0.0638	
561693	MC	0	22	0	25	0.40	0.35	1.4888	1.6465	
561805	MC	0	24	0	26	0.83	0.81	-0.8069	-0.7799	
561848	MC	0	26	0	29	0.88	0.87	-1.3053	-1.2888	
561878	MC	0	28	0	30	0.88	0.87	-1.3155	-1.2201	
Mean						0.64	0.62	0.25	0.25	

Appendix J: Linking Item Statistics

Mathematics Grade 5

ID	Type	Form	Seq	Prev	Prev	Prev	P-Val	P-Val	Prev	Meas
				Form	Seq	P-Val			Meas	
561692	MC	0	3	0	3	0.63	0.59	0.2170	0.2524	
561757	MC	0	5	0	6	0.79	0.75	-0.6598	-0.5833	
561879	MC	0	6	0	7	0.54	0.52	0.6908	0.6211	
561883	MC	0	11	0	12	0.50	0.48	0.8437	0.7821	
560692	MC	0	12	0	13	0.30	0.26	1.8285	1.8823	
561920	MC	0	13	0	15	0.55	0.53	0.5934	0.5544	
560715	MC	0	19	0	20	0.82	0.78	-0.8497	-0.7518	
561913	MC	0	21	0	22	0.61	0.60	0.3519	0.2224	
561858	MC	0	22	0	24	0.67	0.65	0.0374	-0.0250	
560705	MC	0	24	0	25	0.52	0.54	0.7338	0.4862	
561841	MC	0	28	0	26	0.83	0.79	-0.9803	-0.8301	
560704	MC	0	29	0	27	0.66	0.60	0.0641	0.2266	
561914	MC	0	30	0	29	0.38	0.34	1.4489	1.4588	
560697	MC	0	31	0	32	0.59	0.55	0.4402	0.4638	
Mean						0.60	0.57	0.34	0.34	

Appendix J: Linking Item Statistics

Mathematics Grade 6

ID	Type	Form	Seq	Prev	Prev	Prev	Prev	Meas	Meas
				Form	Seq	P-Val	P-Val		
560709	MC	0	1	0	1	0.41	0.39	0.6270	0.7194
562078	MC	0	3	0	3	0.66	0.63	-0.5513	-0.4186
561905	MC	0	8	0	6	0.45	0.49	0.4502	0.2075
561838	MC	0	11	0	8	0.62	0.64	-0.3168	-0.4580
561906	MC	0	12	0	9	0.36	0.34	0.8922	0.9634
561857	MC	0	13	0	10	0.50	0.49	0.2417	0.2434
561936	MC	0	14	0	18	0.36	0.36	0.8995	0.8661
561927	MC	0	17	0	19	0.49	0.49	0.2945	0.2204
561842	MC	0	18	0	20	0.30	0.28	1.1907	1.2824
561853	MC	0	21	0	23	0.46	0.46	0.4193	0.3552
561818	MC	0	23	0	25	0.46	0.42	0.4347	0.5332
561696	MC	0	24	0	27	0.78	0.82	-1.2258	-1.5299
561642	MC	0	29	0	31	0.36	0.32	0.8959	1.0545
561769	MC	0	31	0	32	0.53	0.47	0.0933	0.3063
Mean						0.48	0.47	0.31	0.31

Appendix J: Linking Item Statistics

Mathematics Grade 7

ID	Type	Form	Seq	Prev	Prev	Prev	P-Val	P-Val	Prev	Meas
				Form	Seq	P-Val			Meas	
561341	MC	0	2	0	2	0.53	0.55	0.1868	0.1534	
560755	MC	0	8	0	5	0.51	0.57	0.2706	0.0341	
561385	MC	0	10	0	8	0.44	0.42	0.5767	0.7305	
561410	MC	0	12	0	11	0.52	0.49	0.2119	0.3890	
560759	MC	0	15	0	15	0.58	0.60	-0.0824	-0.0878	
561901	MC	0	17	0	22	0.56	0.60	0.0385	-0.0831	
561691	MC	0	22	0	25	0.55	0.54	0.0656	0.1672	
561916	MC	0	23	0	26	0.54	0.53	0.1583	0.2218	
561651	MC	0	27	0	27	0.46	0.49	0.4753	0.3981	
561670	MC	0	28	0	28	0.75	0.76	-0.9248	-0.9376	
561682	MC	0	29	0	29	0.70	0.73	-0.6902	-0.7424	
561941	MC	0	30	0	31	0.43	0.43	0.6483	0.6920	
Mean						0.55	0.56	0.08	0.08	

Appendix J: Linking Item Statistics

Mathematics Grade 8

ID	Type	Form	Seq	Prev	Prev	Prev	Prev	Meas	Meas
				Form	Seq	P-Val	P-Val		
561346	MC	0	1	0	1	0.63	0.62	-0.3466	-0.2943
561358	MC	0	3	0	3	0.64	0.66	-0.3830	-0.4904
561409	MC	0	7	0	4	0.30	0.30	1.2843	1.2694
561412	MC	0	8	0	6	0.46	0.46	0.4485	0.4651
561366	MC	0	9	0	8	0.48	0.44	0.3538	0.5501
560742	MC	0	10	0	10	0.54	0.53	0.0984	0.1232
561637	MC	0	11	0	12	0.52	0.53	0.1728	0.1456
561647	MC	0	13	0	14	0.41	0.39	0.7200	0.7647
561582	MC	0	17	0	21	0.41	0.43	0.6720	0.6002
560746	MC	0	20	0	22	0.58	0.57	-0.1145	-0.0344
561554	MC	0	23	0	24	0.48	0.49	0.3802	0.2977
561546	MC	0	24	0	25	0.51	0.51	0.2316	0.2217
560748	MC	0	27	0	28	0.70	0.69	-0.7010	-0.6541
561561	MC	0	28	0	30	0.51	0.51	0.2347	0.2161
561401	MC	0	30	0	31	0.40	0.42	0.7635	0.6335
Mean						0.50	0.50	0.25	0.25

Appendix J: Linking Item Statistics

Mathematics Grade 11

ID	Type	Form	Seq	Prev	Prev	Prev	Prev	Meas	Meas
				Form	Seq	P-Val	P-Val		
561124	MC	0	2	0	2	0.66	0.58	-0.6527	-0.2787
561129	MC	0	3	0	4	0.48	0.48	0.1884	0.1938
561138	MC	0	6	0	6	0.72	0.71	-0.9624	-0.9316
561166	MC	0	11	0	8	0.39	0.43	0.6302	0.4185
561118	MC	0	12	0	9	0.53	0.54	-0.0190	-0.0727
561229	MC	0	15	0	15	0.49	0.48	0.1646	0.1959
561133	MC	0	17	0	18	0.45	0.46	0.3517	0.2853
561139	MC	0	19	0	19	0.40	0.39	0.5626	0.6235
561414	MC	0	21	0	20	0.44	0.44	0.3932	0.3566
561143	MC	0	22	0	21	0.38	0.37	0.6636	0.7079
561144	MC	0	23	0	22	0.43	0.45	0.4106	0.3445
561150	MC	0	26	0	24	0.52	0.55	0.0021	-0.1182
561430	MC	0	29	0	26	0.37	0.38	0.7029	0.6441
561156	MC	0	30	0	27	0.45	0.44	0.3157	0.3522
561158	MC	0	31	0	29	0.29	0.29	1.1182	1.1488
Mean						0.47	0.47	0.26	0.26

Appendix K:

Reliabilities

Column Heading	Definition
Strand	Strand (Tot.=total)
Group	Subgroup
Pts.	Points possible
Len.	Length
N	N
Mean	Mean
<i>SD</i>	Standard deviation
r	Reliability coefficient
<i>SEM</i>	Standard error of measurement
Items	Item types present

Appendix K: Reliabilities

Mathematics Grade 4

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	2375	22.7	6.95	0.82	2.9	MC/OE
	A	All	16	13	2375	9.0	3.57	0.68	2.0	MC/OE
	B	All	5	5	2375	2.5	1.33	0.39	1.0	MC
	C	All	6	3	2375	3.5	1.52	0.26	1.3	MC/OE
	D	All	5	5	2375	2.7	1.27	0.42	1.0	MC
	E	All	6	6	2375	4.9	1.35	0.63	0.8	MC

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	38	32	1415	22.7	7.12	0.83	2.9	MC/OE
		Female	38	32	956	22.6	6.70	0.81	2.9	MC/OE
	A	Male	16	13	1415	9.0	3.64	0.69	2.0	MC/OE
		Female	16	13	956	9.1	3.48	0.67	2.0	MC/OE
	B	Male	5	5	1415	2.7	1.34	0.41	1.0	MC
		Female	5	5	956	2.4	1.29	0.34	1.1	MC
	C	Male	6	3	1415	3.4	1.53	0.26	1.3	MC/OE
		Female	6	3	956	3.5	1.49	0.27	1.3	MC/OE
	D	Male	5	5	1415	2.8	1.27	0.42	1.0	MC
		Female	5	5	956	2.6	1.28	0.42	1.0	MC
	E	Male	6	6	1415	4.8	1.41	0.65	0.8	MC
		Female	6	6	956	5.0	1.25	0.60	0.8	MC

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	38	32	1478	23.1	6.82	0.82	2.9	MC/OE
		Af. Amer.	38	32	564	21.7	7.22	0.83	2.9	MC/OE
		Hispanic	38	32	263	21.8	6.88	0.82	2.9	MC/OE
		Asian	38	32	19	23.2	7.27	0.84	2.9	MC/OE
		Am. Indian	38	32	6	25.5	2.43	-0.17	2.6	MC/OE
		Multi	38	32	42	24.0	7.04	0.83	2.9	MC/OE
	A	White	16	13	1478	9.2	3.56	0.68	2.0	MC/OE
		Af. Amer.	16	13	564	8.7	3.65	0.69	2.0	MC/OE
		Hispanic	16	13	263	8.8	3.46	0.66	2.0	MC/OE
		Asian	16	13	19	9.4	3.52	0.67	2.0	MC/OE
		Am. Indian	16	13	6	10.3	2.34	0.36	1.9	MC/OE
		Multi	16	13	42	10.0	3.74	0.72	2.0	MC/OE
	B	White	5	5	1478	2.6	1.32	0.38	1.0	MC
		Af. Amer.	5	5	564	2.4	1.34	0.39	1.0	MC
		Hispanic	5	5	263	2.3	1.32	0.38	1.0	MC
		Asian	5	5	19	3.0	1.25	0.33	1.0	MC
		Am. Indian	5	5	6	3.0	1.10	0.42	0.8	MC
		Multi	5	5	42	2.6	1.35	0.39	1.1	MC
	C	White	6	3	1478	3.6	1.49	0.26	1.3	MC/OE
		Af. Amer.	6	3	564	3.2	1.52	0.27	1.3	MC/OE
		Hispanic	6	3	263	3.4	1.57	0.31	1.3	MC/OE
		Asian	6	3	19	3.3	1.49	0.18	1.4	MC/OE
		Am. Indian	6	3	6	3.3	1.37	0.37	1.1	MC/OE
		Multi	6	3	42	3.6	1.58	0.30	1.3	MC/OE
	D	White	5	5	1478	2.7	1.27	0.42	1.0	MC
		Af. Amer.	5	5	564	2.7	1.29	0.43	1.0	MC
		Hispanic	5	5	263	2.6	1.27	0.43	1.0	MC
		Asian	5	5	19	2.6	1.42	0.58	0.9	MC
		Am. Indian	5	5	6	3.0	0.63	-1.46	1.0	MC
Multi		5	5	42	2.6	1.21	0.45	0.9	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	6	6	1478	5.0	1.30	0.64	0.8	MC
		Af. Amer.	6	6	564	4.7	1.43	0.62	0.9	MC
		Hispanic	6	6	263	4.8	1.38	0.61	0.9	MC
		Asian	6	6	19	4.8	1.62	0.76	0.8	MC
		Am. Indian	6	6	6	5.8	0.41	0.00	0.4	MC
		Multi	6	6	42	5.1	1.15	0.55	0.8	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	122	22.2	7.20	0.84	2.9	MC/OE
	A	All	16	13	122	9.1	3.62	0.69	2.0	MC/OE
	B	All	5	5	122	2.3	1.22	0.23	1.1	MC
	C	All	6	3	122	3.5	1.57	0.36	1.2	MC/OE
	D	All	5	5	122	2.7	1.24	0.40	1.0	MC
	E	All	6	6	122	4.7	1.48	0.65	0.9	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	1510	22.2	6.95	0.82	2.9	MC/OE
	A	All	16	13	1510	8.9	3.56	0.68	2.0	MC/OE
	B	All	5	5	1510	2.5	1.32	0.38	1.0	MC
	C	All	6	3	1510	3.4	1.53	0.27	1.3	MC/OE
	D	All	5	5	1510	2.6	1.28	0.42	1.0	MC
	E	All	6	6	1510	4.8	1.37	0.62	0.8	MC

Appendix K: Reliabilities

Mathematics Grade 5

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	3366	21.2	7.12	0.83	2.9	MC/OE
	A	All	16	13	3366	7.7	3.40	0.67	2.0	MC/OE
	B	All	6	6	3366	3.5	1.43	0.47	1.0	MC
	C	All	5	5	3366	2.9	1.42	0.51	1.0	MC
	D	All	5	5	3366	3.5	1.32	0.52	0.9	MC
	E	All	6	3	3366	3.5	1.66	0.33	1.4	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	Male		38	32	1923	21.4	7.20	0.83	2.9	MC/OE
		Female		38	32	1440	20.9	7.01	0.82	2.9	MC/OE
	A	Male		16	13	1923	7.9	3.41	0.68	1.9	MC/OE
		Female		16	13	1440	7.4	3.36	0.66	2.0	MC/OE
	B	Male		6	6	1923	3.5	1.42	0.47	1.0	MC
		Female		6	6	1440	3.4	1.44	0.47	1.0	MC
	C	Male		5	5	1923	2.9	1.43	0.52	1.0	MC
		Female		5	5	1440	3.0	1.41	0.50	1.0	MC
	D	Male		5	5	1923	3.6	1.33	0.56	0.9	MC
		Female		5	5	1440	3.5	1.29	0.48	0.9	MC
	E	Male		6	3	1923	3.4	1.69	0.33	1.4	MC/OE
		Female		6	3	1440	3.7	1.61	0.34	1.3	MC/OE

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	White		38	32	2158	21.5	7.01	0.83	2.9	MC/OE
		Af. Amer.		38	32	775	20.1	7.30	0.83	3.0	MC/OE
		Hispanic		38	32	352	21.9	7.30	0.84	2.9	MC/OE
		Asian		38	32	38	21.1	6.56	0.79	3.0	MC/OE
		Am. Indian		38	32	8	18.1	6.27	0.79	2.9	MC/OE
		Multi		38	32	31	20.7	6.25	0.77	3.0	MC/OE
		A	White		16	13	2158	7.7	3.37	0.67	1.9
	Af. Amer.			16	13	775	7.4	3.43	0.68	1.9	MC/OE
	Hispanic			16	13	352	8.0	3.54	0.69	2.0	MC/OE
	Asian			16	13	38	7.8	3.14	0.58	2.0	MC/OE
	Am. Indian			16	13	8	5.8	3.62	0.80	1.6	MC/OE
	Multi			16	13	31	7.4	3.06	0.58	2.0	MC/OE
	B		White		6	6	2158	3.5	1.43	0.47	1.0
		Af. Amer.		6	6	775	3.3	1.45	0.47	1.1	MC
		Hispanic		6	6	352	3.7	1.40	0.43	1.1	MC
		Asian		6	6	38	3.4	1.29	0.25	1.1	MC
		Am. Indian		6	6	8	3.4	1.92	0.78	0.9	MC
		Multi		6	6	31	3.5	1.15	0.16	1.1	MC
		C	White		5	5	2158	3.0	1.43	0.52	1.0
	Af. Amer.			5	5	775	2.8	1.41	0.48	1.0	MC
	Hispanic			5	5	352	3.0	1.40	0.50	1.0	MC
	Asian			5	5	38	2.8	1.61	0.64	1.0	MC
	Am. Indian			5	5	8	2.5	1.60	0.66	0.9	MC
	Multi			5	5	31	2.8	1.24	0.23	1.1	MC
	D		White		5	5	2158	3.6	1.29	0.51	0.9
		Af. Amer.		5	5	775	3.4	1.39	0.55	0.9	MC
		Hispanic		5	5	352	3.6	1.25	0.47	0.9	MC
		Asian		5	5	38	3.4	1.40	0.54	0.9	MC
		Am. Indian		5	5	8	3.1	1.73	0.73	0.9	MC
Multi			5	5	31	3.4	1.34	0.48	1.0	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	6	3	2158	3.6	1.64	0.35	1.3	MC/OE
		Af. Amer.	6	3	775	3.2	1.70	0.30	1.4	MC/OE
		Hispanic	6	3	352	3.6	1.65	0.34	1.3	MC/OE
		Asian	6	3	38	3.7	1.58	0.38	1.2	MC/OE
		Am. Indian	6	3	8	3.4	1.77	0.51	1.2	MC/OE
		Multi	6	3	31	3.5	1.29	-0.33	1.5	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	145	20.6	7.35	0.84	2.9	MC/OE
	A	All	16	13	145	7.3	3.47	0.69	1.9	MC/OE
	B	All	6	6	145	3.4	1.33	0.33	1.1	MC
	C	All	5	5	145	2.8	1.51	0.58	1.0	MC
	D	All	5	5	145	3.5	1.29	0.49	0.9	MC
	E	All	6	3	145	3.6	1.69	0.37	1.3	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	2132	21.0	7.21	0.83	3.0	MC/OE
	A	All	16	13	2132	7.6	3.44	0.68	2.0	MC/OE
	B	All	6	6	2132	3.5	1.42	0.46	1.0	MC
	C	All	5	5	2132	2.9	1.42	0.51	1.0	MC
	D	All	5	5	2132	3.5	1.33	0.52	0.9	MC
	E	All	6	3	2132	3.5	1.68	0.33	1.4	MC/OE

Appendix K: Reliabilities

Mathematics Grade 6

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	3600	17.6	6.54	0.83	2.7	MC/OE
	A	All	11	11	3600	5.0	2.78	0.74	1.4	MC
	B	All	6	3	3600	2.1	1.28	0.27	1.1	MC/OE
	C	All	6	6	3600	3.2	1.34	0.33	1.1	MC
	D	All	8	5	3600	3.5	1.73	0.52	1.2	MC/OE
	E	All	7	7	3600	3.7	1.57	0.43	1.2	MC

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	Male		38	32	2116	17.8	6.69	0.83	2.7	MC/OE
		Female		38	32	1482	17.2	6.31	0.82	2.7	MC/OE
	A	Male		11	11	2116	5.1	2.86	0.75	1.4	MC
		Female		11	11	1482	4.8	2.65	0.71	1.4	MC
	B	Male		6	3	2116	2.2	1.30	0.26	1.1	MC/OE
		Female		6	3	1482	2.0	1.24	0.29	1.1	MC/OE
	C	Male		6	6	2116	3.3	1.34	0.33	1.1	MC
		Female		6	6	1482	3.2	1.33	0.33	1.1	MC
	D	Male		8	5	2116	3.6	1.74	0.52	1.2	MC/OE
		Female		8	5	1482	3.4	1.72	0.51	1.2	MC/OE
	E	Male		7	7	2116	3.7	1.59	0.45	1.2	MC
		Female		7	7	1482	3.7	1.55	0.42	1.2	MC

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	White		38	32	2401	17.7	6.51	0.83	2.7	MC/OE
		Af. Amer.		38	32	745	16.8	6.64	0.83	2.7	MC/OE
		Hispanic		38	32	367	18.2	6.63	0.83	2.7	MC/OE
		Asian		38	32	29	17.3	5.82	0.77	2.8	MC/OE
		Am. Indian		38	32	4	22.0	8.41	0.90	2.7	MC/OE
		Multi		38	32	50	18.2	5.60	0.76	2.7	MC/OE
	A	White		11	11	2401	4.9	2.78	0.74	1.4	MC
		Af. Amer.		11	11	745	5.0	2.77	0.74	1.4	MC
		Hispanic		11	11	367	5.5	2.73	0.72	1.4	MC
		Asian		11	11	29	4.6	2.43	0.64	1.4	MC
		Am. Indian		11	11	4	7.3	3.86	0.90	1.2	MC
		Multi		11	11	50	5.1	2.77	0.74	1.4	MC
	B	White		6	3	2401	2.2	1.27	0.26	1.1	MC/OE
		Af. Amer.		6	3	745	2.0	1.30	0.30	1.1	MC/OE
		Hispanic		6	3	367	2.1	1.29	0.23	1.1	MC/OE
		Asian		6	3	29	1.9	1.52	0.54	1.0	MC/OE
		Am. Indian		6	3	4	3.3	2.06	0.59	1.3	MC/OE
		Multi		6	3	50	2.1	1.31	0.36	1.0	MC/OE
	C	White		6	6	2401	3.3	1.34	0.34	1.1	MC
		Af. Amer.		6	6	745	3.1	1.33	0.30	1.1	MC
		Hispanic		6	6	367	3.2	1.34	0.34	1.1	MC
		Asian		6	6	29	3.1	1.13	-0.11	1.2	MC
		Am. Indian		6	6	4	3.5	1.29	0.48	0.9	MC
		Multi		6	6	50	3.4	1.09	-0.07	1.1	MC
	D	White		8	5	2401	3.5	1.74	0.52	1.2	MC/OE
		Af. Amer.		8	5	745	3.3	1.67	0.50	1.2	MC/OE
		Hispanic		8	5	367	3.6	1.80	0.57	1.2	MC/OE
		Asian		8	5	29	4.1	1.62	0.32	1.3	MC/OE
		Am. Indian		8	5	4	4.0	0.82	-0.62	1.0	MC/OE
Multi			8	5	50	3.8	1.68	0.56	1.1	MC/OE	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	7	7	2401	3.8	1.55	0.42	1.2	MC
		Af. Amer.	7	7	745	3.4	1.58	0.43	1.2	MC
		Hispanic	7	7	367	3.8	1.65	0.50	1.2	MC
		Asian	7	7	29	3.6	1.57	0.45	1.2	MC
		Am. Indian	7	7	4	4.0	2.45	0.81	1.1	MC
		Multi	7	7	50	3.9	1.57	0.45	1.2	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	148	17.9	6.05	0.80	2.7	MC/OE
	A	All	11	11	148	5.4	2.57	0.68	1.5	MC
	B	All	6	3	148	2.1	1.17	0.25	1.0	MC/OE
	C	All	6	6	148	3.2	1.25	0.25	1.1	MC
	D	All	8	5	148	3.6	1.76	0.53	1.2	MC/OE
	E	All	7	7	148	3.7	1.54	0.39	1.2	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	2198	17.3	6.48	0.82	2.7	MC/OE
	A	All	11	11	2198	4.9	2.76	0.73	1.4	MC
	B	All	6	3	2198	2.1	1.26	0.26	1.1	MC/OE
	C	All	6	6	2198	3.2	1.34	0.32	1.1	MC
	D	All	8	5	2198	3.4	1.72	0.53	1.2	MC/OE
	E	All	7	7	2198	3.6	1.57	0.43	1.2	MC

Appendix K: Reliabilities

Mathematics Grade 7

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	3972	19.9	6.74	0.80	3.0	MC/OE
	A	All	9	9	3972	4.9	2.15	0.59	1.4	MC
	B	All	6	3	3972	2.3	1.50	0.29	1.3	MC/OE
	C	All	8	5	3972	4.3	2.06	0.49	1.5	MC/OE
	D	All	9	9	3972	4.8	1.99	0.50	1.4	MC
	E	All	6	6	3972	3.6	1.39	0.42	1.1	MC

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	38	32	2384	19.7	6.87	0.81	3.0	MC/OE
		Female	38	32	1580	20.1	6.54	0.80	3.0	MC/OE
	A	Male	9	9	2384	4.9	2.17	0.60	1.4	MC
		Female	9	9	1580	4.9	2.13	0.58	1.4	MC
	B	Male	6	3	2384	2.4	1.52	0.28	1.3	MC/OE
		Female	6	3	1580	2.2	1.47	0.30	1.2	MC/OE
	C	Male	8	5	2384	4.2	2.09	0.50	1.5	MC/OE
		Female	8	5	1580	4.5	2.00	0.46	1.5	MC/OE
	D	Male	9	9	2384	4.7	2.03	0.52	1.4	MC
		Female	9	9	1580	4.8	1.93	0.47	1.4	MC
	E	Male	6	6	2384	3.6	1.41	0.43	1.1	MC
		Female	6	6	1580	3.6	1.36	0.40	1.1	MC

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	38	32	2656	20.2	6.61	0.80	3.0	MC/OE
		Af. Amer.	38	32	861	18.7	6.91	0.82	3.0	MC/OE
		Hispanic	38	32	339	20.5	7.04	0.82	3.0	MC/OE
		Asian	38	32	30	22.5	5.91	0.74	3.0	MC/OE
		Am. Indian	38	32	11	19.3	6.07	0.75	3.0	MC/OE
		Multi	38	32	65	18.8	6.94	0.81	3.0	MC/OE
		A	White	9	9	2656	4.9	2.11	0.58	1.4
	Af. Amer.		9	9	861	4.7	2.24	0.63	1.4	MC
	Hispanic		9	9	339	5.2	2.18	0.61	1.4	MC
	Asian		9	9	30	5.4	1.73	0.30	1.4	MC
	Am. Indian		9	9	11	4.5	1.97	0.46	1.4	MC
	Multi		9	9	65	4.8	2.39	0.69	1.3	MC
	B		White	6	3	2656	2.4	1.50	0.28	1.3
		Af. Amer.	6	3	861	2.1	1.47	0.29	1.2	MC/OE
		Hispanic	6	3	339	2.3	1.55	0.31	1.3	MC/OE
		Asian	6	3	30	2.7	1.63	0.41	1.3	MC/OE
		Am. Indian	6	3	11	2.5	1.21	0.30	1.0	MC/OE
		Multi	6	3	65	2.3	1.52	0.35	1.2	MC/OE
		C	White	8	5	2656	4.4	2.02	0.47	1.5
	Af. Amer.		8	5	861	3.9	2.08	0.52	1.4	MC/OE
	Hispanic		8	5	339	4.4	2.12	0.54	1.4	MC/OE
	Asian		8	5	30	5.1	1.94	0.42	1.5	MC/OE
	Am. Indian		8	5	11	3.5	2.07	0.39	1.6	MC/OE
	Multi		8	5	65	3.9	2.12	0.50	1.5	MC/OE
	D		White	9	9	2656	4.8	1.98	0.49	1.4
		Af. Amer.	9	9	861	4.5	1.98	0.50	1.4	MC
		Hispanic	9	9	339	5.0	2.12	0.58	1.4	MC
		Asian	9	9	30	5.6	1.71	0.35	1.4	MC
		Am. Indian	9	9	11	4.9	1.81	0.32	1.5	MC
Multi		9	9	65	4.4	1.69	0.27	1.5	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	6	6	2656	3.7	1.36	0.40	1.1	MC
		Af. Amer.	6	6	861	3.4	1.45	0.44	1.1	MC
		Hispanic	6	6	339	3.6	1.44	0.48	1.0	MC
		Asian	6	6	30	3.7	1.37	0.46	1.0	MC
		Am. Indian	6	6	11	3.9	0.94	-0.44	1.1	MC
		Multi	6	6	65	3.4	1.46	0.42	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	114	20.2	6.46	0.79	3.0	MC/OE
	A	All	9	9	114	5.3	2.10	0.60	1.3	MC
	B	All	6	3	114	2.1	1.59	0.35	1.3	MC/OE
	C	All	8	5	114	4.5	1.95	0.47	1.4	MC/OE
	D	All	9	9	114	4.9	1.90	0.46	1.4	MC
	E	All	6	6	114	3.5	1.48	0.49	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	2331	19.7	6.78	0.81	3.0	MC/OE
	A	All	9	9	2331	4.9	2.16	0.60	1.4	MC
	B	All	6	3	2331	2.3	1.50	0.30	1.3	MC/OE
	C	All	8	5	2331	4.2	2.07	0.50	1.5	MC/OE
	D	All	9	9	2331	4.7	1.98	0.49	1.4	MC
	E	All	6	6	2331	3.6	1.42	0.43	1.1	MC

Appendix K: Reliabilities

Mathematics Grade 8

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	4114	18.5	6.87	0.83	2.9	MC/OE
	A	All	7	7	4114	3.7	1.58	0.41	1.2	MC
	B	All	6	3	4114	1.7	1.46	0.35	1.2	MC/OE
	C	All	9	6	4114	4.6	2.04	0.49	1.5	MC/OE
	D	All	10	10	4114	5.3	2.52	0.70	1.4	MC
	E	All	6	6	4114	3.2	1.46	0.47	1.1	MC

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	38	32	2431	18.4	6.98	0.83	2.9	MC/OE
		Female	38	32	1675	18.6	6.72	0.82	2.8	MC/OE
	A	Male	7	7	2431	3.8	1.58	0.42	1.2	MC
		Female	7	7	1675	3.6	1.56	0.40	1.2	MC
	B	Male	6	3	2431	1.7	1.46	0.35	1.2	MC/OE
		Female	6	3	1675	1.7	1.47	0.36	1.2	MC/OE
	C	Male	9	6	2431	4.6	2.08	0.49	1.5	MC/OE
		Female	9	6	1675	4.7	1.98	0.48	1.4	MC/OE
	D	Male	10	10	2431	5.1	2.54	0.70	1.4	MC
		Female	10	10	1675	5.5	2.48	0.69	1.4	MC
	E	Male	6	6	2431	3.3	1.47	0.47	1.1	MC
		Female	6	6	1675	3.1	1.44	0.47	1.0	MC

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	38	32	2760	18.6	6.77	0.82	2.8	MC/OE
		Af. Amer.	38	32	863	17.8	7.13	0.84	2.9	MC/OE
		Hispanic	38	32	394	19.2	6.97	0.83	2.9	MC/OE
		Asian	38	32	40	17.9	6.69	0.81	2.9	MC/OE
		Am. Indian	38	32	2	22.5	4.95	0.67	2.8	MC/OE
		Multi	38	32	47	17.3	6.72	0.82	2.8	MC/OE
		A	White	7	7	2760	3.7	1.56	0.40	1.2
	Af. Amer.		7	7	863	3.7	1.60	0.43	1.2	MC
	Hispanic		7	7	394	3.8	1.60	0.45	1.2	MC
	Asian		7	7	40	3.6	1.41	0.23	1.2	MC
	Am. Indian		7	7	2	2.5	0.71	-2.33	1.3	MC
	Multi		7	7	47	3.8	1.73	0.54	1.2	MC
	B		White	6	3	2760	1.7	1.45	0.35	1.2
		Af. Amer.	6	3	863	1.6	1.46	0.34	1.2	MC/OE
		Hispanic	6	3	394	1.9	1.57	0.35	1.3	MC/OE
		Asian	6	3	40	1.6	1.48	0.49	1.1	MC/OE
		Am. Indian	6	3	2	2.0	1.41	0.00	1.4	MC/OE
		Multi	6	3	47	1.5	1.40	0.41	1.1	MC/OE
		C	White	9	6	2760	4.7	2.01	0.49	1.4
	Af. Amer.		9	6	863	4.3	2.11	0.49	1.5	MC/OE
	Hispanic		9	6	394	4.9	2.05	0.46	1.5	MC/OE
	Asian		9	6	40	4.3	1.95	0.36	1.6	MC/OE
	Am. Indian		9	6	2	7.0	0.00			MC/OE
	Multi		9	6	47	4.1	1.97	0.47	1.4	MC/OE
	D		White	10	10	2760	5.3	2.51	0.69	1.4
		Af. Amer.	10	10	863	5.2	2.57	0.70	1.4	MC
		Hispanic	10	10	394	5.5	2.54	0.70	1.4	MC
		Asian	10	10	40	5.2	2.45	0.66	1.4	MC
		Am. Indian	10	10	2	7.0	2.83	0.83	1.2	MC
Multi		10	10	47	4.9	2.71	0.74	1.4	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	6	6	2760	3.3	1.46	0.47	1.1	MC
		Af. Amer.	6	6	863	3.1	1.47	0.47	1.1	MC
		Hispanic	6	6	394	3.1	1.46	0.46	1.1	MC
		Asian	6	6	40	3.2	1.44	0.45	1.1	MC
		Am. Indian	6	6	2	4.0	1.41	0.00	1.4	MC
		Multi	6	6	47	3.0	1.57	0.52	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	132	18.4	7.06	0.83	2.9	MC/OE
	A	All	7	7	132	3.7	1.66	0.48	1.2	MC
	B	All	6	3	132	1.9	1.60	0.36	1.3	MC/OE
	C	All	9	6	132	4.6	1.98	0.41	1.5	MC/OE
	D	All	10	10	132	5.2	2.55	0.70	1.4	MC
	E	All	6	6	132	2.9	1.58	0.54	1.1	MC

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	2368	18.4	6.95	0.83	2.9	MC/OE
	A	All	7	7	2368	3.7	1.61	0.44	1.2	MC
	B	All	6	3	2368	1.7	1.48	0.36	1.2	MC/OE
	C	All	9	6	2368	4.6	2.04	0.49	1.5	MC/OE
	D	All	10	10	2368	5.3	2.52	0.69	1.4	MC
	E	All	6	6	2368	3.2	1.46	0.46	1.1	MC

Appendix K: Reliabilities

Mathematics Grade 11

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	38	32	4269	16.4	7.28	0.85	2.8	MC/OE
	A	All	5	5	4269	2.2	1.44	0.53	1.0	MC
	B	All	5	5	4269	2.9	1.36	0.49	1.0	MC
	C	All	6	3	4269	1.9	1.45	0.37	1.2	MC/OE
	D	All	16	16	4269	7.1	3.38	0.70	1.8	MC
	E	All	6	3	4269	2.2	1.52	0.43	1.2	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	38	32	2573	16.4	7.29	0.85	2.8	MC/OE
		Female	38	32	1690	16.5	7.25	0.85	2.8	MC/OE
	A	Male	5	5	2573	2.3	1.46	0.55	1.0	MC
		Female	5	5	1690	2.2	1.41	0.50	1.0	MC
	B	Male	5	5	2573	2.9	1.37	0.50	1.0	MC
		Female	5	5	1690	2.9	1.36	0.49	1.0	MC
	C	Male	6	3	2573	1.9	1.45	0.38	1.1	MC/OE
		Female	6	3	1690	2.0	1.47	0.36	1.2	MC/OE
	D	Male	16	16	2573	7.0	3.35	0.70	1.8	MC
		Female	16	16	1690	7.2	3.43	0.71	1.8	MC
	E	Male	6	3	2573	2.2	1.57	0.44	1.2	MC/OE
		Female	6	3	1690	2.2	1.46	0.40	1.1	MC/OE

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	38	32	2993	17.1	7.29	0.85	2.9	MC/OE
		Af. Amer.	38	32	862	14.4	6.93	0.84	2.8	MC/OE
		Hispanic	38	32	326	15.6	7.00	0.84	2.8	MC/OE
		Asian	38	32	32	16.9	7.63	0.87	2.8	MC/OE
		Am. Indian	38	32	7	14.1	9.10	0.90	2.9	MC/OE
		Multi	38	32	40	17.0	6.20	0.78	2.9	MC/OE
		A	White	5	5	2993	2.3	1.44	0.52	1.0
	Af. Amer.		5	5	862	1.9	1.39	0.52	1.0	MC
	Hispanic		5	5	326	2.2	1.46	0.54	1.0	MC
	Asian		5	5	32	2.1	1.18	0.15	1.1	MC
	Am. Indian		5	5	7	2.0	2.08	0.87	0.8	MC
	Multi		5	5	40	2.2	1.25	0.30	1.0	MC
	B		White	5	5	2993	3.0	1.38	0.51	1.0
		Af. Amer.	5	5	862	2.7	1.31	0.44	1.0	MC
		Hispanic	5	5	326	2.8	1.23	0.33	1.0	MC
		Asian	5	5	32	3.1	1.48	0.65	0.9	MC
		Am. Indian	5	5	7	2.6	1.62	0.61	1.0	MC
		Multi	5	5	40	3.0	1.34	0.49	1.0	MC
		C	White	6	3	2993	2.1	1.47	0.37	1.2
	Af. Amer.		6	3	862	1.6	1.35	0.35	1.1	MC/OE
	Hispanic		6	3	326	1.9	1.42	0.35	1.1	MC/OE
	Asian		6	3	32	2.0	1.45	0.40	1.1	MC/OE
	Am. Indian		6	3	7	1.7	1.50	0.73	0.8	MC/OE
	Multi		6	3	40	2.2	1.49	0.29	1.3	MC/OE
	D		White	16	16	2993	7.3	3.37	0.70	1.8
		Af. Amer.	16	16	862	6.4	3.31	0.70	1.8	MC
		Hispanic	16	16	326	6.8	3.31	0.69	1.8	MC
		Asian	16	16	32	7.7	3.86	0.78	1.8	MC
		Am. Indian	16	16	7	5.7	4.46	0.87	1.6	MC
Multi		16	16	40	7.5	3.27	0.67	1.9	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
E		White	6	3	2993	2.3	1.56	0.44	1.2	MC/OE
		Af. Amer.	6	3	862	1.9	1.39	0.34	1.1	MC/OE
		Hispanic	6	3	326	1.9	1.38	0.41	1.1	MC/OE
		Asian	6	3	32	2.1	1.43	0.37	1.1	MC/OE
		Am. Indian	6	3	7	2.1	1.86	0.08	1.8	MC/OE
		Multi	6	3	40	2.3	1.50	0.57	1.0	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	38	32	58	14.0	5.90	0.78	2.8	MC/OE
	A	All	5	5	58	2.0	1.35	0.45	1.0	MC
	B	All	5	5	58	2.6	1.30	0.47	0.9	MC
	C	All	6	3	58	1.5	1.27	0.30	1.1	MC/OE
	D	All	16	16	58	6.2	2.68	0.51	1.9	MC
	E	All	6	3	58	1.7	1.31	0.37	1.0	MC/OE

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Eco. Disadv.	Tot.	All	38	32	2129	15.5	7.10	0.84	2.8	MC/OE
	A	All	5	5	2129	2.1	1.44	0.53	1.0	MC
	B	All	5	5	2129	2.8	1.34	0.47	1.0	MC
	C	All	6	3	2129	1.8	1.41	0.38	1.1	MC/OE
	D	All	16	16	2129	6.7	3.30	0.69	1.8	MC
	E	All	6	3	2129	2.1	1.48	0.42	1.1	MC/OE

Appendix K: Reliabilities

Reading Grade 4

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	3388	19.2	6.87	0.83	2.9
A	All		25	23	3388	13.9	5.10	0.78	2.4	MC/OE
B	All		11	9	3388	5.4	2.30	0.52	1.6	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		Male	36	32	2195	19.1	6.99	0.83	2.8
Female			36	32	1191	19.4	6.65	0.82	2.8	MC/OE
A		Male	25	23	2195	13.8	5.20	0.79	2.4	MC/OE
		Female	25	23	1191	14.0	4.89	0.76	2.4	MC/OE
B		Male	11	9	2195	5.3	2.31	0.52	1.6	MC/OE
		Female	11	9	1191	5.5	2.27	0.52	1.6	MC/OE

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		White	36	32	2295	19.8	6.74	0.82	2.8
Af. Amer.			36	32	674	17.7	6.93	0.83	2.9	MC/OE
Hispanic			36	32	334	18.1	7.09	0.84	2.9	MC/OE
Asian			36	32	29	18.3	7.35	0.85	2.8	MC/OE
Am. Indian			36	32	7	24.9	5.46	0.79	2.5	MC/OE
Multi			36	32	47	20.9	6.40	0.81	2.8	MC/OE
A		White	25	23	2295	14.3	5.03	0.78	2.4	MC/OE
		Af. Amer.	25	23	674	12.7	5.12	0.78	2.4	MC/OE
		Hispanic	25	23	334	13.1	5.07	0.78	2.4	MC/OE
		Asian	25	23	29	13.5	5.41	0.81	2.3	MC/OE
		Am. Indian	25	23	7	18.4	3.55	0.67	2.1	MC/OE
		Multi	25	23	47	15.3	4.86	0.78	2.3	MC/OE
B		White	11	9	2295	5.5	2.24	0.50	1.6	MC/OE
		Af. Amer.	11	9	674	5.0	2.32	0.53	1.6	MC/OE
		Hispanic	11	9	334	5.0	2.47	0.59	1.6	MC/OE
		Asian	11	9	29	4.8	2.65	0.67	1.5	MC/OE
		Am. Indian	11	9	7	6.4	2.07	0.52	1.4	MC/OE
		Multi	11	9	47	5.6	1.98	0.34	1.6	MC/OE

ELL	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	163	17.8	7.36	0.85	2.9
A	All		25	23	163	12.9	5.27	0.79	2.4	MC/OE
B	All		11	9	163	5.0	2.61	0.64	1.6	MC/OE

Eco. Dis.	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	2050	18.5	6.93	0.83	2.9
A	All		25	23	2050	13.3	5.15	0.79	2.4	MC/OE
B	All		11	9	2050	5.2	2.31	0.52	1.6	MC/OE

Appendix K: Reliabilities

Reading Grade 5

			Pts.	Len.	N	Mean	SD	r	SEM	Items
	Strand	Group								
Overall	Tot.	All	36	32	3947	20.5	7.19	0.86	2.7	MC/OE
	A	All	25	23	3947	14.5	5.24	0.82	2.2	MC/OE
	B	All	11	9	3947	6.0	2.43	0.63	1.5	MC/OE
Gender	Tot.	Male	36	32	2423	20.2	7.33	0.87	2.7	MC/OE
		Female	36	32	1518	21.0	6.90	0.85	2.6	MC/OE
	A	Male	25	23	2423	14.3	5.36	0.83	2.2	MC/OE
		Female	25	23	1518	14.8	5.00	0.81	2.2	MC/OE
	B	Male	11	9	2423	5.9	2.45	0.63	1.5	MC/OE
		Female	11	9	1518	6.2	2.40	0.63	1.5	MC/OE
Ethnicity	Tot.	White	36	32	2636	21.2	7.00	0.86	2.6	MC/OE
		Af. Amer.	36	32	827	18.9	7.21	0.86	2.7	MC/OE
		Hispanic	36	32	375	19.2	7.67	0.88	2.7	MC/OE
		Asian	36	32	47	20.2	6.99	0.86	2.6	MC/OE
		Am. Indian	36	32	10	23.6	4.86	0.72	2.6	MC/OE
		Multi	36	32	46	19.4	6.90	0.85	2.7	MC/OE
	A	White	25	23	2636	15.1	5.10	0.81	2.2	MC/OE
		Af. Amer.	25	23	827	13.3	5.24	0.82	2.2	MC/OE
		Hispanic	25	23	375	13.5	5.58	0.84	2.2	MC/OE
		Asian	25	23	47	14.3	5.01	0.80	2.2	MC/OE
		Am. Indian	25	23	10	16.8	3.68	0.68	2.1	MC/OE
		Multi	25	23	46	13.6	5.16	0.82	2.2	MC/OE
	B	White	11	9	2636	6.2	2.38	0.62	1.5	MC/OE
		Af. Amer.	11	9	827	5.6	2.49	0.65	1.5	MC/OE
		Hispanic	11	9	375	5.7	2.55	0.66	1.5	MC/OE
		Asian	11	9	47	5.8	2.38	0.66	1.4	MC/OE
		Am. Indian	11	9	10	6.8	1.87	0.34	1.5	MC/OE
		Multi	11	9	46	5.7	2.27	0.50	1.6	MC/OE
ELL	Tot.	All	36	32	158	18.2	7.56	0.88	2.7	MC/OE
	A	All	25	23	158	12.8	5.41	0.83	2.2	MC/OE
	B	All	11	9	158	5.5	2.58	0.67	1.5	MC/OE
Eco. Dis.	Tot.	All	36	32	2390	19.8	7.29	0.87	2.7	MC/OE
	A	All	25	23	2390	14.0	5.33	0.83	2.2	MC/OE
	B	All	11	9	2390	5.8	2.45	0.63	1.5	MC/OE

Appendix K: Reliabilities

Reading Grade 6

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	3983	21.8	6.98	0.86	2.6	MC/OE
	A	All	18	18	3983	12.0	3.81	0.79	1.8	MC
	B	All	18	14	3983	9.8	3.67	0.73	1.9	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	36	32	2501	21.4	7.17	0.87	2.6	MC/OE
		Female	36	32	1480	22.5	6.60	0.85	2.6	MC/OE
	A	Male	18	18	2501	11.9	3.95	0.80	1.8	MC
		Female	18	18	1480	12.3	3.54	0.75	1.8	MC
	B	Male	18	14	2501	9.5	3.71	0.73	1.9	MC/OE
Female		18	14	1480	10.2	3.57	0.72	1.9	MC/OE	

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	36	32	2742	22.4	6.86	0.86	2.6	MC/OE
		Af. Amer.	36	32	757	20.2	7.04	0.86	2.7	MC/OE
		Hispanic	36	32	393	20.6	7.08	0.86	2.7	MC/OE
		Asian	36	32	29	21.1	7.34	0.87	2.6	MC/OE
		Am. Indian	36	32	5	23.4	7.54	0.90	2.4	MC/OE
		Multi	36	32	53	22.0	7.23	0.87	2.6	MC/OE
	A	White	18	18	2742	12.4	3.71	0.78	1.7	MC
		Af. Amer.	18	18	757	11.2	3.93	0.79	1.8	MC
		Hispanic	18	18	393	11.2	3.91	0.79	1.8	MC
		Asian	18	18	29	11.5	3.48	0.72	1.8	MC
		Am. Indian	18	18	5	12.6	3.36	0.72	1.8	MC
		Multi	18	18	53	12.4	3.64	0.77	1.8	MC
	B	White	18	14	2742	10.0	3.65	0.73	1.9	MC/OE
		Af. Amer.	18	14	757	9.1	3.57	0.70	2.0	MC/OE
		Hispanic	18	14	393	9.4	3.68	0.72	1.9	MC/OE
		Asian	18	14	29	9.7	4.34	0.83	1.8	MC/OE
		Am. Indian	18	14	5	10.8	4.38	0.86	1.6	MC/OE
Multi		18	14	53	9.6	3.95	0.77	1.9	MC/OE	

ELL	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	165	19.3	6.89	0.85	2.7	MC/OE
	A	All	18	18	165	10.5	3.90	0.77	1.9	MC
	B	All	18	14	165	8.8	3.49	0.69	2.0	MC/OE

Eco. Dis.	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	2374	20.9	7.02	0.86	2.6	MC/OE
	A	All	18	18	2374	11.5	3.91	0.79	1.8	MC
	B	All	18	14	2374	9.4	3.62	0.72	1.9	MC/OE

Appendix K: Reliabilities

Reading Grade 7

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	3974	20.1	6.54	0.83	2.7	MC/OE
	A	All	18	16	3974	9.7	3.52	0.72	1.9	MC/OE
	B	All	18	16	3974	10.4	3.56	0.69	2.0	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	Male		36	32	2480	19.6	6.75	0.84	2.7	MC/OE
		Female		36	32	1489	21.0	6.07	0.80	2.7	MC/OE
	A	Male		18	16	2480	9.5	3.60	0.73	1.9	MC/OE
		Female		18	16	1489	10.0	3.36	0.69	1.9	MC/OE
	B	Male		18	16	2480	10.1	3.66	0.71	2.0	MC/OE
Female			18	16	1489	11.0	3.31	0.65	1.9	MC/OE	

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items	
	Tot.	White		36	32	2710	20.8	6.39	0.82	2.7	MC/OE
		Af. Amer.		36	32	769	18.4	6.72	0.84	2.7	MC/OE
		Hispanic		36	32	373	19.0	6.31	0.81	2.8	MC/OE
		Asian		36	32	45	21.2	7.35	0.87	2.6	MC/OE
		Am. Indian		36	32	11	21.8	4.47	0.68	2.5	MC/OE
		Multi		36	32	60	20.0	6.51	0.82	2.7	MC/OE
	A	White		18	16	2710	10.1	3.49	0.72	1.9	MC/OE
		Af. Amer.		18	16	769	8.7	3.53	0.72	1.9	MC/OE
		Hispanic		18	16	373	9.1	3.31	0.67	1.9	MC/OE
		Asian		18	16	45	10.0	3.34	0.68	1.9	MC/OE
		Am. Indian		18	16	11	9.8	2.93	0.63	1.8	MC/OE
		Multi		18	16	60	9.5	3.18	0.64	1.9	MC/OE
	B	White		18	16	2710	10.7	3.45	0.68	2.0	MC/OE
		Af. Amer.		18	16	769	9.7	3.73	0.72	2.0	MC/OE
		Hispanic		18	16	373	9.9	3.53	0.68	2.0	MC/OE
		Asian		18	16	45	11.2	4.42	0.82	1.9	MC/OE
		Am. Indian		18	16	11	12.0	2.32	0.39	1.8	MC/OE
Multi			18	16	60	10.5	3.73	0.72	2.0	MC/OE	

ELL	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	135	17.2	5.93	0.78	2.8	MC/OE
	A	All	18	16	135	8.3	3.03	0.59	2.0	MC/OE
	B	All	18	16	135	8.9	3.38	0.65	2.0	MC/OE

Eco. Dis.	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	36	32	2314	19.6	6.47	0.82	2.7	MC/OE
	A	All	18	16	2314	9.4	3.47	0.71	1.9	MC/OE
	B	All	18	16	2314	10.2	3.55	0.69	2.0	MC/OE

Appendix K: Reliabilities

Reading Grade 8

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	3647	21.3	6.46	0.83	2.7
A	All		19	17	3647	11.9	3.58	0.71	1.9	MC/OE
B	All		17	15	3647	9.3	3.41	0.69	1.9	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		Male	36	32	2273	21.0	6.57	0.83	2.7
Female			36	32	1372	21.8	6.23	0.81	2.7	MC/OE
A		Male	19	17	2273	11.8	3.66	0.72	1.9	MC/OE
		Female	19	17	1372	12.2	3.44	0.69	1.9	MC/OE
B		Male	17	15	2273	9.1	3.43	0.69	1.9	MC/OE
		Female	17	15	1372	9.7	3.35	0.68	1.9	MC/OE

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		White	36	32	2457	22.0	6.28	0.82	2.7
Af. Amer.			36	32	746	19.8	6.62	0.83	2.7	MC/OE
Hispanic			36	32	360	19.8	6.48	0.82	2.7	MC/OE
Asian			36	32	42	18.5	6.11	0.79	2.8	MC/OE
Am. Indian			36	32	3	22.3	12.50	0.96	2.4	MC/OE
Multi			36	32	37	19.7	6.84	0.84	2.7	MC/OE
A		White	19	17	2457	12.3	3.45	0.70	1.9	MC/OE
		Af. Amer.	19	17	746	11.2	3.71	0.72	2.0	MC/OE
		Hispanic	19	17	360	11.0	3.68	0.72	2.0	MC/OE
		Asian	19	17	42	10.0	3.55	0.67	2.0	MC/OE
		Am. Indian	19	17	3	12.0	7.00	0.94	1.7	MC/OE
		Multi	19	17	37	11.6	3.44	0.67	2.0	MC/OE
B		White	17	15	2457	9.7	3.35	0.68	1.9	MC/OE
		Af. Amer.	17	15	746	8.6	3.44	0.69	1.9	MC/OE
		Hispanic	17	15	360	8.8	3.36	0.67	1.9	MC/OE
		Asian	17	15	42	8.5	3.21	0.64	1.9	MC/OE
		Am. Indian	17	15	3	10.3	5.51	0.89	1.8	MC/OE
		Multi	17	15	37	8.1	3.94	0.77	1.9	MC/OE

ELL	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	121	18.6	6.44	0.82	2.8
A	All		19	17	121	10.4	3.82	0.73	2.0	MC/OE
B	All		17	15	121	8.2	3.12	0.61	1.9	MC/OE

Eco. Dis.	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	2153	20.5	6.48	0.82	2.7
A	All		19	17	2153	11.5	3.65	0.72	1.9	MC/OE
B	All		17	15	2153	9.0	3.37	0.68	1.9	MC/OE

Appendix K: Reliabilities

Reading Grade 11

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	3919	22.9	6.34	0.83	2.6
A	All		19	17	3919	12.3	3.50	0.71	1.9	MC/OE
B	All		17	15	3919	10.6	3.35	0.70	1.8	MC/OE

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		Male	36	32	2484	22.7	6.54	0.84	2.6
Female			36	32	1420	23.1	5.95	0.81	2.6	MC/OE
A		Male	19	17	2484	12.3	3.62	0.73	1.9	MC/OE
		Female	19	17	1420	12.3	3.28	0.67	1.9	MC/OE
B		Male	17	15	2484	10.4	3.41	0.71	1.8	MC/OE
		Female	17	15	1420	10.8	3.22	0.68	1.8	MC/OE

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.		White	36	32	2693	23.6	6.08	0.82	2.6
Af. Amer.			36	32	823	21.1	6.66	0.84	2.7	MC/OE
Hispanic			36	32	315	21.9	6.23	0.82	2.6	MC/OE
Asian			36	32	32	21.0	7.15	0.86	2.7	MC/OE
Am. Indian			36	32	7	27.0	5.23	0.78	2.4	MC/OE
Multi			36	32	33	22.3	6.32	0.82	2.7	MC/OE
A		White	19	17	2693	12.7	3.35	0.69	1.9	MC/OE
		Af. Amer.	19	17	823	11.4	3.77	0.74	1.9	MC/OE
		Hispanic	19	17	315	11.9	3.48	0.70	1.9	MC/OE
		Asian	19	17	32	11.7	4.14	0.79	1.9	MC/OE
		Am. Indian	19	17	7	14.3	1.98	0.19	1.8	MC/OE
		Multi	19	17	33	12.2	3.47	0.71	1.9	MC/OE
B		White	17	15	2693	10.9	3.26	0.69	1.8	MC/OE
		Af. Amer.	17	15	823	9.6	3.39	0.69	1.9	MC/OE
		Hispanic	17	15	315	10.0	3.26	0.68	1.8	MC/OE
		Asian	17	15	32	9.4	3.54	0.72	1.9	MC/OE
		Am. Indian	17	15	7	12.7	3.35	0.74	1.7	MC/OE
		Multi	17	15	33	10.1	3.42	0.68	1.9	MC/OE

ELL	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	56	17.9	6.82	0.85	2.7
A	All		19	17	56	9.9	4.04	0.78	1.9	MC/OE
B	All		17	15	56	8.1	3.17	0.64	1.9	MC/OE

Eco. Dis.	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All		36	32	2016	21.9	6.48	0.83	2.7
A	All		19	17	2016	11.9	3.60	0.72	1.9	MC/OE
B	All		17	15	2016	10.0	3.38	0.70	1.9	MC/OE

Appendix K: Reliabilities

Science Grade 8

		Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
Overall	Tot.	All		34	32	3252	20.9	5.82	0.80	2.6	MC/OE
	A	All		17	16	3252	10.5	3.11	0.66	1.8	MC/OE
	B	All		7	6	3252	4.5	1.64	0.48	1.2	MC/OE
	C	All		5	5	3252	2.7	1.29	0.40	1.0	MC
	D	All		5	5	3252	3.2	1.31	0.42	1.0	MC
Gender	Tot.	Male		34	32	2005	21.3	6.03	0.82	2.6	MC/OE
		Female		34	32	1242	20.3	5.37	0.76	2.6	MC/OE
	A	Male		17	16	2005	10.6	3.21	0.68	1.8	MC/OE
		Female		17	16	1242	10.4	2.94	0.61	1.8	MC/OE
	B	Male		7	6	2005	4.5	1.62	0.49	1.2	MC/OE
		Female		7	6	1242	4.4	1.67	0.48	1.2	MC/OE
	C	Male		5	5	2005	2.8	1.34	0.46	1.0	MC
		Female		5	5	1242	2.5	1.17	0.24	1.0	MC
	D	Male		5	5	2005	3.4	1.33	0.47	1.0	MC
		Female		5	5	1242	3.0	1.23	0.29	1.0	MC
Ethnicity	Tot.	White		34	32	2209	21.8	5.53	0.79	2.6	MC/OE
		Af. Amer.		34	32	620	18.3	5.78	0.79	2.7	MC/OE
		Hispanic		34	32	339	19.8	5.94	0.80	2.6	MC/OE
		Asian		34	32	41	18.3	6.19	0.82	2.6	MC/OE
		Am. Indian		34	32	4	21.0	6.78	0.87	2.5	MC/OE
		Multi		34	32	34	19.8	5.72	0.78	2.7	MC/OE
	A	White		17	16	2209	10.9	2.96	0.63	1.8	MC/OE
		Af. Amer.		17	16	620	9.4	3.25	0.67	1.9	MC/OE
		Hispanic		17	16	339	10.1	3.25	0.68	1.8	MC/OE
		Asian		17	16	41	9.6	3.11	0.65	1.8	MC/OE
		Am. Indian		17	16	4	10.3	2.75	0.56	1.8	MC/OE
		Multi		17	16	34	10.0	3.13	0.64	1.9	MC/OE
	B	White		7	6	2209	4.7	1.57	0.47	1.1	MC/OE
		Af. Amer.		7	6	620	3.9	1.60	0.43	1.2	MC/OE
		Hispanic		7	6	339	4.2	1.74	0.51	1.2	MC/OE
		Asian		7	6	41	4.0	1.82	0.58	1.2	MC/OE
		Am. Indian		7	6	4	5.3	1.71	0.69	1.0	MC/OE
		Multi		7	6	34	4.3	1.70	0.47	1.2	MC/OE
	C	White		5	5	2209	2.8	1.27	0.39	1.0	MC
		Af. Amer.		5	5	620	2.3	1.24	0.35	1.0	MC
		Hispanic		5	5	339	2.5	1.26	0.34	1.0	MC
		Asian		5	5	41	2.2	1.47	0.60	0.9	MC
		Am. Indian		5	5	4	3.0	1.15	0.16	1.1	MC
		Multi		5	5	34	2.4	1.30	0.43	1.0	MC
D	White		5	5	2209	3.4	1.27	0.42	1.0	MC	
	Af. Amer.		5	5	620	2.7	1.30	0.35	1.1	MC	
	Hispanic		5	5	339	3.1	1.30	0.38	1.0	MC	
	Asian		5	5	41	2.5	1.38	0.44	1.0	MC	
	Am. Indian		5	5	4	2.5	1.29	0.25	1.1	MC	
	Multi		5	5	34	3.1	1.10	0.03	1.1	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	34	32	115	18.8	5.84	0.80	2.6	MC/OE
	A	All	17	16	115	9.6	3.25	0.68	1.8	MC/OE
	B	All	7	6	115	3.9	1.69	0.48	1.2	MC/OE
	C	All	5	5	115	2.3	1.20	0.27	1.0	MC
	D	All	5	5	115	2.9	1.40	0.50	1.0	MC
Eco. Disadv.	Tot.	All	34	32	1857	20.1	5.86	0.80	2.6	MC/OE
	A	All	17	16	1857	10.2	3.14	0.65	1.8	MC/OE
	B	All	7	6	1857	4.3	1.66	0.49	1.2	MC/OE
	C	All	5	5	1857	2.5	1.27	0.38	1.0	MC
	D	All	5	5	1857	3.1	1.31	0.40	1.0	MC

Appendix K: Reliabilities

Science Grade 11

Overall	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	All	34	32	3540	17.8	5.99	0.80	2.7	MC/OE
	A	All	17	16	3540	9.8	3.46	0.72	1.8	MC/OE
	B	All	6	6	3540	3.1	1.44	0.39	1.1	MC
	C	All	6	5	3540	3.0	1.55	0.44	1.2	MC/OE
	D	All	5	5	3540	2.0	1.21	0.25	1.0	MC

Gender	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	Male	34	32	2233	18.4	6.25	0.82	2.7	MC/OE
		Female	34	32	1300	16.9	5.40	0.76	2.6	MC/OE
	A	Male	17	16	2233	10.0	3.61	0.75	1.8	MC/OE
		Female	17	16	1300	9.5	3.17	0.67	1.8	MC/OE
	B	Male	6	6	2233	3.1	1.45	0.40	1.1	MC
		Female	6	6	1300	2.9	1.40	0.36	1.1	MC
	C	Male	6	5	2233	3.2	1.59	0.47	1.2	MC/OE
		Female	6	5	1300	2.7	1.44	0.38	1.1	MC/OE
	D	Male	5	5	2233	2.1	1.24	0.28	1.1	MC
		Female	5	5	1300	1.8	1.15	0.18	1.0	MC

Ethnicity	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
	Tot.	White	34	32	2466	18.9	5.85	0.79	2.7	MC/OE
		Af. Amer.	34	32	696	15.1	5.47	0.76	2.7	MC/OE
		Hispanic	34	32	293	16.0	5.66	0.78	2.7	MC/OE
		Asian	34	32	28	16.4	6.01	0.80	2.7	MC/OE
		Am. Indian	34	32	10	22.4	5.72	0.80	2.5	MC/OE
		Multi	34	32	36	15.9	6.36	0.83	2.6	MC/OE
	A	White	17	16	2466	10.4	3.35	0.71	1.8	MC/OE
		Af. Amer.	17	16	696	8.3	3.29	0.68	1.9	MC/OE
		Hispanic	17	16	293	8.9	3.37	0.71	1.8	MC/OE
		Asian	17	16	28	8.7	3.27	0.66	1.9	MC/OE
		Am. Indian	17	16	10	11.5	3.95	0.83	1.6	MC/OE
		Multi	17	16	36	8.6	3.71	0.76	1.8	MC/OE
	B	White	6	6	2466	3.2	1.43	0.39	1.1	MC
		Af. Amer.	6	6	696	2.6	1.38	0.34	1.1	MC
		Hispanic	6	6	293	2.8	1.38	0.33	1.1	MC
		Asian	6	6	28	3.0	1.60	0.50	1.1	MC
		Am. Indian	6	6	10	3.8	1.03	-0.02	1.0	MC
		Multi	6	6	36	2.6	1.20	0.09	1.1	MC
	C	White	6	5	2466	3.2	1.53	0.42	1.2	MC/OE
		Af. Amer.	6	5	696	2.5	1.47	0.42	1.1	MC/OE
		Hispanic	6	5	293	2.6	1.53	0.42	1.2	MC/OE
		Asian	6	5	28	2.9	1.50	0.42	1.1	MC/OE
		Am. Indian	6	5	10	4.1	1.10	-0.23	1.2	MC/OE
		Multi	6	5	36	2.9	1.82	0.63	1.1	MC/OE
	D	White	5	5	2466	2.1	1.24	0.28	1.1	MC
		Af. Amer.	5	5	696	1.8	1.10	0.09	1.0	MC
		Hispanic	5	5	293	1.7	1.16	0.21	1.0	MC
		Asian	5	5	28	1.9	1.17	0.18	1.1	MC
		Am. Indian	5	5	10	3.0	1.41	0.46	1.0	MC
Multi		5	5	36	1.8	1.14	0.13	1.1	MC	

Appendix K: Reliabilities

	Strand	Group	Pts.	Len.	N	Mean	SD	r	SEM	Items
ELL	Tot.	All	34	32	56	13.6	4.77	0.67	2.7	MC/OE
	A	All	17	16	56	7.3	2.97	0.59	1.9	MC/OE
	B	All	6	6	56	2.4	1.32	0.22	1.2	MC
	C	All	6	5	56	2.3	1.35	0.26	1.2	MC/OE
	D	All	5	5	56	1.6	0.97	-0.22	1.1	MC
Eco. Disadv.	Tot.	All	34	32	1802	16.9	5.93	0.80	2.7	MC/OE
	A	All	17	16	1802	9.3	3.53	0.73	1.8	MC/OE
	B	All	6	6	1802	2.9	1.39	0.34	1.1	MC
	C	All	6	5	1802	2.8	1.53	0.43	1.2	MC/OE
	D	All	5	5	1802	1.9	1.17	0.19	1.0	MC

Appendix L:
Cut Scores and Scale Transformations

Column Heading	Definition
LOSS	Lowest Obtainable Scaled Score

Appendix L: Cut Scores and Scale Transformations

	Grade	Scaling	LOSS	True Scaled Score Cuts			True 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
Reading	4	106.00X + 1286.67	1075	1150	1275	1363	-1.2893	-0.1101	0.7217
	5	103.97X + 1255.27	1075	1150	1275	1391	-1.0125	0.1898	1.3018
	6	95.72X + 1216.58	1075	1150	1275	1381	-0.6956	0.6103	1.7201
	7	107.79X + 1223.17	1075	1150	1275	1385	-0.6788	0.4809	1.5026
	8	109.97X + 1203.43	1050	1150	1275	1399	-0.4859	0.6508	1.7754
	11	117.92X + 1155.25	1000	1150	1275	1433	-0.0445	1.0155	2.3512
Sci.	8	116.16X + 1188.19	1050	1150	1275	1419	-0.3288	0.7473	1.9833
	11	90.16X + 1245.16	1100	1150	1275	1401	-1.0555	0.3310	1.7325

Appendix M:
PSSA-M Historical Statistics

Appendix M: PSSA-M Historical Statistics

		2010	2011	2010	2011	2010	2011
Raw Score	Mean	22.89	22.68	21.89	21.18	18.27	17.56
	SD	6.55	6.95	6.43	7.12	6.01	6.54
	Max	38	38	38	38	38	38
Scaled Score	Mean	1286.2	1277.8	1275.5	1261.2	1264.6	1260.6
	SD	83.8	83.9	84.7	85.5	83.5	89.5
	Max	1655	1666	1712	1691	1772	1770
Raw Cuts	Bel. Basic/Basic	12	11	12	12	10	10
	Basic/Prof.	22	23	22	23	19	19
	Prof./Adv.	29	30	30	31	27	27
Theta Cuts	Bel. Basic/Basic	-0.5891	-0.5898	-0.5352	-0.4198	-0.9606	-0.8620
	Basic/Prof.	0.8935	0.8911	0.8640	0.8636	0.3441	0.3944
	Prof./Adv.	1.8540	1.8538	1.9734	2.0164	1.4543	1.4781
Impact %	Bel. Basic	4.8	5.8	5.5	10.4	7.3	11.4
	Basic	35.8	40.5	43.4	44.4	44.6	45.8
	Proficient	38.3	36.2	38.0	35.7	38.9	33.2
	Advanced	21.2	17.5	13.0	9.5	9.2	9.7
	Prof. + Adv.	59.4	53.7	51.1	45.2	48.1	42.8
Demographics	N Count	2169	2375	2552	3366	2700	3600
	% City	6.4	11.2	6.0	9.0	6.9	7.7
	% White	68.8	62.2	69.9	64.1	69.4	66.7
	% Black	18.9	23.7	18.5	23.0	18.9	20.7
	% Hispanic	9.1	11.1	8.5	10.5	8.6	10.2

		2010	2011	2010	2011	2010	2011
Raw Score	Mean	19.05	19.88	18.91	18.49	16.95	16.41
	SD	7.00	6.74	6.41	6.87	6.27	7.28
	Max	38	38	38	38	38	38
Scaled Score	Mean	1252.0	1256.0	1250.3	1250.8	1228.3	1227.8
	SD	79.2	74.5	83.5	87.0	99.6	107.2
	Max	1663	1662	1730	1722	1890	2041
Raw Cuts	Bel. Basic/Basic	10	10	11	11	12	11
	Basic/Prof.	21	22	21	21	20	20
	Prof./Adv.	30	30	30	30	28	29
Theta Cuts	Bel. Basic/Basic	-0.8442	-0.8393	-0.7858	-0.6649	-0.5492	-0.5395
	Basic/Prof.	0.5878	0.5948	0.5407	0.5752	0.5317	0.5465
	Prof./Adv.	1.6086	1.5970	1.8139	1.8116	1.6389	1.6829
Impact %	Bel. Basic	8.3	5.5	10.0	13.0	21.6	24.4
	Basic	50.4	53.5	49.2	48.7	45.3	43.1
	Proficient	33.2	32.5	35.2	31.7	27.6	25.4
	Advanced	8.1	8.5	5.6	6.7	5.6	7.0
	Prof. + Adv.	41.2	41.0	40.8	38.3	33.2	32.4
Demographics	N Count	2818	3972	3019	4114	3539	4269
	% City	7.1	8.8	6.1	9.2	6.9	7.8
	% White	70.5	66.9	70.9	67.1	70.7	70.1
	% Black	18.1	21.7	18.4	21.0	19.9	20.2
	% Hispanic	8.4	8.5	8.3	9.6	6.4	7.6

Appendix M: PSSA-M Historical Statistics

		2011	2011	2011	2011	2011	2011
Raw Score	Mean	19.21	20.50	21.80	20.14	21.27	22.87
	SD	6.87	7.19	6.98	6.54	6.46	6.34
	Max	36	36	36	36	36	36
Scaled Score	Mean	1305.6	1301.8	1281.9	1268.3	1257.4	1249.1
	SD	98.2	110.8	101.5	99.3	101.5	110.4
	Max	1808	1791	1724	1788	1756	1753
Raw Cuts	Bel. Basic/Basic	8	10	12	12	14	17
	Basic/Prof.	17	19	22	21	23	25
	Prof./Adv.	24	27	29	28	30	32
Theta Cuts	Bel. Basic/Basic	-1.2893	-1.0125	-0.6956	-0.6788	-0.4859	-0.0445
	Basic/Prof.	-0.1101	0.1898	0.6103	0.4809	0.6508	1.0155
	Prof./Adv.	0.7217	1.3018	1.7201	1.5026	1.7754	2.3512
Impact %	Bel. Basic	3.8	7.9	10.1	11.7	13.9	17.4
	Basic	32.3	30.7	33.6	37.7	40.0	37.1
	Proficient	33.7	37.1	37.8	36.9	36.3	40.2
	Advanced	30.2	24.3	18.4	13.7	9.8	5.3
	Prof. + Adv.	63.9	61.4	56.2	50.6	46.1	45.5
Demographics	N Count	3388	3947	3983	3974	3647	3919
	% City	9.9	9.6	7.9	8.6	9.0	8.5
	% White	67.7	66.8	68.8	68.2	67.4	68.7
	% Black	19.9	21.0	19.0	19.4	20.5	21.0
	% Hispanic	9.9	9.5	9.9	9.4	9.9	8.0

Appendix M: PSSA-M Historical Statistics

		2011	2011
Raw Score	Mean	20.89	17.84
	SD	5.82	5.99
	Max	34	34
Scaled Score	Mean	1266.6	1263.5
	SD	107.4	79.2
	Max	1769	1694
Raw Cuts	Bel. Basic/Basic	14	9
	Basic/Prof.	22	19
	Prof./Adv.	29	28
Theta Cuts	Bel. Basic/Basic	-0.3288	-1.0555
	Basic/Prof.	0.7473	0.3310
	Prof./Adv.	1.9833	1.7325
Impact %	Bel. Basic	12.0	5.8
	Basic	40.1	47.7
	Proficient	38.2	40.7
	Advanced	9.7	5.8
	Prof. + Adv.	47.8	46.6
Demographic	N Count	3252	3540
	% City	7.4	7.4
	% White	67.9	69.7
	% Black	19.1	19.7
	% Hispanic	10.4	8.3

Science-M Grade 8

Science-M Grade 11

Appendix N:

Raw-to-Scaled Score Conversion Tables

Column Heading	Definition
Raw	Raw score
Meas	Rasch measure
MeasSE	Rasch measure standard error
SS	Scaled score
SSSE	Scaled score standard error
Freq	Frequency
Freq%	Frequency percent
Cum	Cumulative frequency
Cum%	Cumulative frequency percent
Pct	Percentile

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 4

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.8193	1.8415	1075	155	0	0.0	0	0.0	0
1	-3.5749	1.0286	1075	87	0	0.0	0	0.0	0
2	-2.8234	0.7473	1075	63	1	0.0	1	0.0	1
3	-2.3593	0.6262	1075	53	1	0.0	2	0.1	1
4	-2.0126	0.5560	1075	47	6	0.3	8	0.3	1
5	-1.7303	0.5091	1075	43	6	0.3	14	0.6	1
6	-1.4888	0.4752	1075	40	11	0.5	25	1.1	1
7	-1.2756	0.4493	1092	38	12	0.5	37	1.6	1
8	-1.0832	0.4287	1108	36	28	1.2	65	2.7	2
9	-0.9068	0.4118	1123	35	32	1.3	97	4.1	3
10	-0.7432	0.3977	1137	34	41	1.7	138	5.8	5
11	-0.5898	0.3858	1150	33	33	1.4	171	7.2	7
12	-0.4449	0.3756	1162	32	47	2.0	218	9.2	8
13	-0.3072	0.3669	1174	31	43	1.8	261	11.0	10
14	-0.1754	0.3595	1185	30	58	2.4	319	13.4	12
15	-0.0484	0.3533	1196	30	73	3.1	392	16.5	15
16	0.0746	0.3483	1206	29	66	2.8	458	19.3	18
17	0.1945	0.3443	1216	29	83	3.5	541	22.8	21
18	0.3120	0.3415	1226	29	92	3.9	633	26.7	25
19	0.4279	0.3397	1236	29	111	4.7	744	31.3	29
20	0.5430	0.3390	1245	29	121	5.1	865	36.4	34
21	0.6581	0.3395	1255	29	126	5.3	991	41.7	39
22	0.7738	0.3412	1265	29	109	4.6	1100	46.3	44
23	0.8911	0.3441	1275	29	120	5.1	1220	51.4	49
24	1.0109	0.3484	1285	29	134	5.6	1354	57.0	54
25	1.1343	0.3542	1295	30	148	6.2	1502	63.2	60
26	1.2622	0.3616	1306	30	128	5.4	1630	68.6	66
27	1.3963	0.3710	1317	31	133	5.6	1763	74.2	71
28	1.5381	0.3825	1329	32	109	4.6	1872	78.8	77
29	1.6897	0.3967	1342	33	87	3.7	1959	82.5	81
30	1.8538	0.4141	1356	35	87	3.7	2046	86.1	84
31	2.0341	0.4359	1371	37	88	3.7	2134	89.9	88
32	2.2358	0.4634	1388	39	73	3.1	2207	92.9	91
33	2.4668	0.4993	1408	42	54	2.3	2261	95.2	94
34	2.7400	0.5486	1431	46	40	1.7	2301	96.9	96
35	3.0797	0.6215	1459	52	36	1.5	2337	98.4	98
36	3.5393	0.7453	1498	63	21	0.9	2358	99.3	99
37	4.2895	1.0289	1561	87	12	0.5	2370	99.8	99
38	5.5352	1.8424	1666	155	5	0.2	2375	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 5

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.6817	1.8395	1075	164	0	0.0	0	0.0	0
1	-3.4425	1.0248	1075	92	0	0.0	0	0.0	0
2	-2.6995	0.7414	1075	66	0	0.0	0	0.0	0
3	-2.2446	0.6185	1075	55	5	0.1	5	0.1	1
4	-1.9078	0.5468	1075	49	8	0.2	13	0.4	1
5	-1.6359	0.4987	1075	45	16	0.5	29	0.9	1
6	-1.4049	0.4639	1075	41	21	0.6	50	1.5	1
7	-1.2022	0.4374	1090	39	40	1.2	90	2.7	2
8	-1.0202	0.4165	1107	37	48	1.4	138	4.1	3
9	-0.8539	0.3995	1122	36	58	1.7	196	5.8	5
10	-0.7000	0.3856	1135	34	72	2.1	268	8.0	7
11	-0.5559	0.3740	1148	33	82	2.4	350	10.4	9
12	-0.4198	0.3643	1160	33	107	3.2	457	13.6	12
13	-0.2900	0.3562	1172	32	80	2.4	537	16.0	15
14	-0.1656	0.3496	1183	31	121	3.6	658	19.5	18
15	-0.0454	0.3442	1194	31	121	3.6	779	23.1	21
16	0.0716	0.3400	1204	30	135	4.0	914	27.2	25
17	0.1861	0.3369	1214	30	136	4.0	1050	31.2	29
18	0.2989	0.3349	1225	30	147	4.4	1197	35.6	33
19	0.4107	0.3341	1235	30	161	4.8	1358	40.3	38
20	0.5223	0.3343	1244	30	159	4.7	1517	45.1	43
21	0.6345	0.3357	1254	30	156	4.6	1673	49.7	47
22	0.7480	0.3383	1265	30	173	5.1	1846	54.8	52
23	0.8636	0.3421	1275	31	183	5.4	2029	60.3	58
24	0.9824	0.3474	1286	31	187	5.6	2216	65.8	63
25	1.1053	0.3541	1297	32	157	4.7	2373	70.5	68
26	1.2336	0.3625	1308	32	140	4.2	2513	74.7	73
27	1.3686	0.3727	1320	33	140	4.2	2653	78.8	77
28	1.5121	0.3851	1333	34	144	4.3	2797	83.1	81
29	1.6660	0.3999	1347	36	124	3.7	2921	86.8	85
30	1.8329	0.4178	1362	37	125	3.7	3046	90.5	89
31	2.0164	0.4396	1378	39	74	2.2	3120	92.7	92
32	2.2214	0.4669	1396	42	93	2.8	3213	95.5	94
33	2.4555	0.5022	1417	45	54	1.6	3267	97.1	96
34	2.7313	0.5505	1442	49	36	1.1	3303	98.1	98
35	3.0724	0.6221	1472	56	30	0.9	3333	99.0	99
36	3.5318	0.7444	1513	67	24	0.7	3357	99.7	99
37	4.2793	1.0268	1580	92	8	0.2	3365	100.0	99
38	5.5214	1.8406	1691	164	1	0.0	3366	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 6

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9295	1.8406	1075	176	0	0.0	0	0.0	0
1	-3.6873	1.0271	1075	98	0	0.0	0	0.0	0
2	-2.9389	0.7454	1075	71	8	0.2	8	0.2	1
3	-2.4776	0.6241	1075	60	8	0.2	16	0.4	1
4	-2.1333	0.5540	1075	53	18	0.5	34	0.9	1
5	-1.8530	0.5075	1075	49	26	0.7	60	1.7	1
6	-1.6127	0.4743	1088	45	47	1.3	107	3.0	2
7	-1.3998	0.4494	1108	43	70	1.9	177	4.9	4
8	-1.2068	0.4300	1126	41	100	2.8	277	7.7	6
9	-1.0286	0.4147	1143	40	132	3.7	409	11.4	10
10	-0.8620	0.4023	1159	39	131	3.6	540	15.0	13
11	-0.7043	0.3922	1175	38	157	4.4	697	19.4	17
12	-0.5538	0.3840	1189	37	176	4.9	873	24.2	22
13	-0.4090	0.3773	1203	36	180	5.0	1053	29.2	27
14	-0.2687	0.3719	1216	36	185	5.1	1238	34.4	32
15	-0.1321	0.3677	1229	35	202	5.6	1440	40.0	37
16	0.0019	0.3645	1242	35	207	5.8	1647	45.8	43
17	0.1339	0.3622	1255	35	213	5.9	1860	51.7	49
18	0.2645	0.3608	1267	35	198	5.5	2058	57.2	54
19	0.3944	0.3602	1280	35	192	5.3	2250	62.5	60
20	0.5242	0.3604	1292	35	179	5.0	2429	67.5	65
21	0.6544	0.3614	1305	35	167	4.6	2596	72.1	70
22	0.7855	0.3631	1317	35	149	4.1	2745	76.3	74
23	0.9182	0.3657	1330	35	178	4.9	2923	81.2	79
24	1.0532	0.3691	1343	35	117	3.3	3040	84.4	83
25	1.1909	0.3734	1356	36	116	3.2	3156	87.7	86
26	1.3322	0.3787	1370	36	96	2.7	3252	90.3	89
27	1.4781	0.3854	1384	37	79	2.2	3331	92.5	91
28	1.6297	0.3936	1398	38	67	1.9	3398	94.4	93
29	1.7885	0.4038	1413	39	55	1.5	3453	95.9	95
30	1.9567	0.4169	1430	40	34	0.9	3487	96.9	96
31	2.1373	0.4339	1447	42	35	1.0	3522	97.8	97
32	2.3351	0.4566	1466	44	29	0.8	3551	98.6	98
33	2.5574	0.4880	1487	47	19	0.5	3570	99.2	99
34	2.8169	0.5333	1512	51	14	0.4	3584	99.6	99
35	3.1371	0.6034	1543	58	11	0.3	3595	99.9	99
36	3.5715	0.7260	1584	70	3	0.1	3598	99.9	99
37	4.2898	1.0116	1653	97	2	0.1	3600	100.0	99
38	5.5090	1.8312	1770	175	0	0.0	3600	100.0	100

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 7

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.7508	1.8367	1075	160	0	0.0	0	0.0	0
1	-3.5189	1.0198	1075	89	0	0.0	0	0.0	0
2	-2.7856	0.7350	1075	64	0	0.0	0	0.0	0
3	-2.3397	0.6114	1075	53	0	0.0	0	0.0	0
4	-2.0113	0.5394	1075	47	7	0.2	7	0.2	1
5	-1.7471	0.4913	1075	43	11	0.3	18	0.5	1
6	-1.5231	0.4567	1091	40	24	0.6	42	1.1	1
7	-1.3268	0.4305	1108	38	47	1.2	89	2.2	2
8	-1.1504	0.4101	1123	36	59	1.5	148	3.7	3
9	-0.9890	0.3938	1137	34	71	1.8	219	5.5	5
10	-0.8393	0.3806	1150	33	113	2.8	332	8.4	7
11	-0.6986	0.3698	1163	32	133	3.3	465	11.7	10
12	-0.5653	0.3609	1174	32	140	3.5	605	15.2	13
13	-0.4377	0.3537	1185	31	172	4.3	777	19.6	17
14	-0.3148	0.3478	1196	30	188	4.7	965	24.3	22
15	-0.1955	0.3431	1207	30	167	4.2	1132	28.5	26
16	-0.0790	0.3394	1217	30	192	4.8	1324	33.3	31
17	0.0352	0.3368	1227	29	212	5.3	1536	38.7	36
18	0.1480	0.3349	1237	29	194	4.9	1730	43.6	41
19	0.2597	0.3339	1246	29	194	4.9	1924	48.4	46
20	0.3712	0.3337	1256	29	217	5.5	2141	53.9	51
21	0.4826	0.3343	1266	29	203	5.1	2344	59.0	56
22	0.5948	0.3356	1276	29	180	4.5	2524	63.5	61
23	0.7080	0.3377	1285	29	203	5.1	2727	68.7	66
24	0.8231	0.3406	1296	30	202	5.1	2929	73.7	71
25	0.9403	0.3445	1306	30	181	4.6	3110	78.3	76
26	1.0607	0.3496	1316	31	136	3.4	3246	81.7	80
27	1.1851	0.3560	1327	31	161	4.1	3407	85.8	84
28	1.3147	0.3643	1338	32	122	3.1	3529	88.8	87
29	1.4512	0.3751	1350	33	107	2.7	3636	91.5	90
30	1.5970	0.3892	1364	34	90	2.3	3726	93.8	93
31	1.7554	0.4077	1377	36	77	1.9	3803	95.7	95
32	1.9315	0.4326	1392	38	42	1.1	3845	96.8	96
33	2.1331	0.4669	1410	41	47	1.2	3892	98.0	97
34	2.3732	0.5158	1431	45	31	0.8	3923	98.8	98
35	2.6763	0.5901	1457	52	23	0.6	3946	99.3	99
36	3.0964	0.7173	1494	63	18	0.5	3964	99.8	99
37	3.8043	1.0076	1556	88	7	0.2	3971	100.0	99
38	5.0193	1.8302	1662	160	1	0.0	3972	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 8

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.7659	1.8387	1075	173	0	0.0	0	0.0	0
1	-3.5290	1.0231	1075	96	1	0.0	1	0.0	1
2	-2.7896	0.7389	1075	70	1	0.0	2	0.0	1
3	-2.3384	0.6155	1075	58	1	0.0	3	0.1	1
4	-2.0053	0.5435	1075	51	15	0.4	18	0.4	1
5	-1.7368	0.4954	1075	47	23	0.6	41	1.0	1
6	-1.5089	0.4609	1082	43	60	1.5	101	2.5	2
7	-1.3087	0.4349	1101	41	82	2.0	183	4.4	3
8	-1.1286	0.4147	1118	39	93	2.3	276	6.7	6
9	-0.9634	0.3987	1133	38	118	2.9	394	9.6	8
10	-0.8097	0.3859	1148	36	141	3.4	535	13.0	11
11	-0.6649	0.3755	1161	35	178	4.3	713	17.3	15
12	-0.5271	0.3672	1174	35	176	4.3	889	21.6	19
13	-0.3948	0.3604	1187	34	200	4.9	1089	26.5	24
14	-0.2669	0.3551	1199	33	207	5.0	1296	31.5	29
15	-0.1423	0.3510	1211	33	206	5.0	1502	36.5	34
16	-0.0203	0.3479	1222	33	218	5.3	1720	41.8	39
17	0.0999	0.3457	1233	33	194	4.7	1914	46.5	44
18	0.2190	0.3445	1245	32	204	5.0	2118	51.5	49
19	0.3375	0.3441	1256	32	221	5.4	2339	56.9	54
20	0.4560	0.3446	1267	32	198	4.8	2537	61.7	59
21	0.5752	0.3460	1278	33	189	4.6	2726	66.3	64
22	0.6957	0.3484	1290	33	189	4.6	2915	70.9	69
23	0.8182	0.3517	1301	33	182	4.4	3097	75.3	73
24	0.9435	0.3562	1313	34	142	3.5	3239	78.7	77
25	1.0723	0.3619	1325	34	175	4.3	3414	83.0	81
26	1.2058	0.3690	1338	35	126	3.1	3540	86.0	85
27	1.3450	0.3775	1351	36	127	3.1	3667	89.1	88
28	1.4913	0.3877	1365	37	96	2.3	3763	91.5	90
29	1.6462	0.3997	1379	38	77	1.9	3840	93.3	92
30	1.8116	0.4140	1395	39	80	1.9	3920	95.3	94
31	1.9900	0.4312	1412	41	68	1.7	3988	96.9	96
32	2.1851	0.4530	1430	43	45	1.1	4033	98.0	97
33	2.4029	0.4820	1450	45	30	0.7	4063	98.8	98
34	2.6547	0.5241	1474	49	21	0.5	4084	99.3	99
35	2.9627	0.5909	1503	56	12	0.3	4096	99.6	99
36	3.3794	0.7119	1542	67	10	0.2	4106	99.8	99
37	4.0758	1.0002	1608	94	6	0.1	4112	100.0	99
38	5.2798	1.8259	1722	172	2	0.0	4114	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Mathematics Grade 11

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.6435	1.8387	1075	213	0	0.0	0	0.0	0
1	-3.4065	1.0233	1075	118	1	0.0	1	0.0	1
2	-2.6664	0.7394	1075	86	5	0.1	6	0.1	1
3	-2.2143	0.6163	1075	71	16	0.4	22	0.5	1
4	-1.8802	0.5444	1075	63	27	0.6	49	1.1	1
5	-1.6107	0.4963	1075	57	72	1.7	121	2.8	2
6	-1.3821	0.4616	1075	53	115	2.7	236	5.5	4
7	-1.1814	0.4353	1077	50	158	3.7	394	9.2	7
8	-1.0011	0.4147	1098	48	199	4.7	593	13.9	12
9	-0.8361	0.3981	1117	46	227	5.3	820	19.2	17
10	-0.6832	0.3847	1135	44	222	5.2	1042	24.4	22
11	-0.5395	0.3736	1151	43	247	5.8	1289	30.2	27
12	-0.4035	0.3644	1167	42	243	5.7	1532	35.9	33
13	-0.2735	0.3569	1182	41	211	4.9	1743	40.8	38
14	-0.1484	0.3507	1196	41	227	5.3	1970	46.1	43
15	-0.0272	0.3457	1210	40	205	4.8	2175	50.9	49
16	0.0908	0.3417	1224	40	189	4.4	2364	55.4	53
17	0.2066	0.3388	1237	39	185	4.3	2549	59.7	58
18	0.3207	0.3369	1251	39	177	4.1	2726	63.9	62
19	0.4338	0.3359	1264	39	158	3.7	2884	67.6	66
20	0.5465	0.3358	1277	39	159	3.7	3043	71.3	69
21	0.6595	0.3367	1290	39	131	3.1	3174	74.3	73
22	0.7735	0.3388	1303	39	138	3.2	3312	77.6	76
23	0.8893	0.3419	1316	40	132	3.1	3444	80.7	79
24	1.0076	0.3464	1330	40	130	3.0	3574	83.7	82
25	1.1296	0.3524	1344	41	107	2.5	3681	86.2	85
26	1.2565	0.3601	1359	42	106	2.5	3787	88.7	87
27	1.3896	0.3699	1374	43	89	2.1	3876	90.8	90
28	1.5309	0.3823	1391	44	94	2.2	3970	93.0	92
29	1.6829	0.3980	1408	46	72	1.7	4042	94.7	94
30	1.8490	0.4179	1427	48	55	1.3	4097	96.0	95
31	2.0341	0.4436	1449	51	55	1.3	4152	97.3	97
32	2.2458	0.4778	1473	55	46	1.1	4198	98.3	98
33	2.4960	0.5250	1502	61	38	0.9	4236	99.2	99
34	2.8070	0.5946	1538	69	18	0.4	4254	99.6	99
35	3.2258	0.7084	1587	82	11	0.3	4265	99.9	99
36	3.8782	0.9306	1662	108	4	0.1	4269	100.0	99
37	5.2178	1.4149	1817	164	0	0.0	4269	100.0	100
38	7.1532	2.0682	2041	239	0	0.0	4269	100.0	100

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 4

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9153	1.8388	1075	195	0	0.0	0	0.0	0
1	-3.6782	1.0231	1075	108	0	0.0	0	0.0	0
2	-2.9388	0.7387	1075	78	1	0.0	1	0.0	1
3	-2.4882	0.6149	1075	65	2	0.1	3	0.1	1
4	-2.1561	0.5423	1075	57	9	0.3	12	0.4	1
5	-1.8892	0.4936	1086	52	20	0.6	32	0.9	1
6	-1.6634	0.4582	1110	49	36	1.1	68	2.0	1
7	-1.4660	0.4313	1131	46	60	1.8	128	3.8	3
8	-1.2893	0.4102	1150	43	89	2.6	217	6.4	5
9	-1.1281	0.3932	1167	42	98	2.9	315	9.3	8
10	-0.9790	0.3795	1183	40	123	3.6	438	12.9	11
11	-0.8393	0.3683	1198	39	121	3.6	559	16.5	15
12	-0.7072	0.3592	1212	38	121	3.6	680	20.1	18
13	-0.5808	0.3519	1225	37	133	3.9	813	24.0	22
14	-0.4591	0.3463	1238	37	127	3.7	940	27.7	26
15	-0.3407	0.3421	1251	36	144	4.3	1084	32.0	30
16	-0.2247	0.3393	1263	36	138	4.1	1222	36.1	34
17	-0.1101	0.3378	1275	36	135	4.0	1357	40.1	38
18	0.0038	0.3375	1287	36	173	5.1	1530	45.2	43
19	0.1179	0.3384	1299	36	169	5.0	1699	50.1	48
20	0.2331	0.3406	1311	36	180	5.3	1879	55.5	53
21	0.3502	0.3439	1324	36	161	4.8	2040	60.2	58
22	0.4700	0.3485	1336	37	169	5.0	2209	65.2	63
23	0.5935	0.3545	1350	38	155	4.6	2364	69.8	67
24	0.7217	0.3619	1363	38	161	4.8	2525	74.5	72
25	0.8559	0.3710	1377	39	168	5.0	2693	79.5	77
26	0.9975	0.3820	1392	40	147	4.3	2840	83.8	82
27	1.1484	0.3954	1408	42	106	3.1	2946	87.0	85
28	1.3111	0.4118	1426	44	118	3.5	3064	90.4	89
29	1.4888	0.4323	1444	46	102	3.0	3166	93.4	92
30	1.6867	0.4584	1465	49	89	2.6	3255	96.1	95
31	1.9122	0.4929	1489	52	63	1.9	3318	97.9	97
32	2.1781	0.5408	1518	57	34	1.0	3352	98.9	98
33	2.5079	0.6126	1553	65	17	0.5	3369	99.4	99
34	2.9551	0.7358	1600	78	10	0.3	3379	99.7	99
35	3.6894	1.0203	1678	108	6	0.2	3385	99.9	99
36	4.9219	1.8368	1808	195	3	0.1	3388	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 5

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.0005	1.8385	1075	191	0	0.0	0	0.0	0
1	-3.7636	1.0233	1075	106	0	0.0	0	0.0	0
2	-3.0231	0.7399	1075	77	1	0.0	1	0.0	1
3	-2.5699	0.6175	1075	64	11	0.3	12	0.3	1
4	-2.2341	0.5463	1075	57	11	0.3	23	0.6	1
5	-1.9621	0.4991	1075	52	23	0.6	46	1.2	1
6	-1.7303	0.4653	1075	48	43	1.1	89	2.3	2
7	-1.5258	0.4400	1097	46	63	1.6	152	3.9	3
8	-1.3411	0.4203	1116	44	66	1.7	218	5.5	5
9	-1.1712	0.4048	1134	42	95	2.4	313	7.9	7
10	-1.0125	0.3924	1150	41	106	2.7	419	10.6	9
11	-0.8625	0.3824	1166	40	109	2.8	528	13.4	12
12	-0.7194	0.3744	1180	39	124	3.1	652	16.5	15
13	-0.5816	0.3681	1195	38	136	3.4	788	20.0	18
14	-0.4480	0.3631	1209	38	145	3.7	933	23.6	22
15	-0.3176	0.3595	1222	37	137	3.5	1070	27.1	25
16	-0.1893	0.3569	1236	37	147	3.7	1217	30.8	29
17	-0.0625	0.3555	1249	37	132	3.3	1349	34.2	33
18	0.0636	0.3550	1262	37	174	4.4	1523	38.6	36
19	0.1898	0.3556	1275	37	169	4.3	1692	42.9	41
20	0.3167	0.3571	1288	37	151	3.8	1843	46.7	45
21	0.4451	0.3598	1302	37	188	4.8	2031	51.5	49
22	0.5759	0.3636	1315	38	189	4.8	2220	56.2	54
23	0.7098	0.3686	1329	38	202	5.1	2422	61.4	59
24	0.8480	0.3751	1343	39	211	5.3	2633	66.7	64
25	0.9917	0.3832	1358	40	170	4.3	2803	71.0	69
26	1.1423	0.3933	1374	41	184	4.7	2987	75.7	73
27	1.3018	0.4059	1391	42	207	5.2	3194	80.9	78
28	1.4727	0.4215	1408	44	187	4.7	3381	85.7	83
29	1.6585	0.4413	1428	46	169	4.3	3550	89.9	88
30	1.8643	0.4669	1449	49	137	3.5	3687	93.4	92
31	2.0976	0.5008	1473	52	85	2.2	3772	95.6	94
32	2.3714	0.5482	1502	57	76	1.9	3848	97.5	97
33	2.7095	0.6194	1537	64	54	1.4	3902	98.9	98
34	3.1654	0.7419	1584	77	28	0.7	3930	99.6	99
35	3.9091	1.0251	1662	107	13	0.3	3943	99.9	99
36	5.1487	1.8396	1791	191	4	0.1	3947	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 6

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.9590	1.8379	1075	176	0	0.0	0	0.0	0
1	-3.7237	1.0222	1075	98	0	0.0	0	0.0	0
2	-2.9855	0.7384	1075	71	0	0.0	0	0.0	0
3	-2.5346	0.6157	1075	59	2	0.1	2	0.1	1
4	-2.2007	0.5445	1075	52	10	0.3	12	0.3	1
5	-1.9307	0.4973	1075	48	16	0.4	28	0.7	1
6	-1.7007	0.4635	1075	44	29	0.7	57	1.4	1
7	-1.4978	0.4383	1075	42	42	1.1	99	2.5	2
8	-1.3144	0.4189	1091	40	53	1.3	152	3.8	3
9	-1.1455	0.4036	1107	39	62	1.6	214	5.4	5
10	-0.9877	0.3915	1122	37	93	2.3	307	7.7	7
11	-0.8383	0.3818	1136	37	97	2.4	404	10.1	9
12	-0.6956	0.3741	1150	36	103	2.6	507	12.7	11
13	-0.5579	0.3681	1163	35	104	2.6	611	15.3	14
14	-0.4242	0.3635	1176	35	110	2.8	721	18.1	17
15	-0.2933	0.3602	1189	34	104	2.6	825	20.7	19
16	-0.1644	0.3580	1201	34	133	3.3	958	24.1	22
17	-0.0367	0.3569	1213	34	132	3.3	1090	27.4	26
18	0.0907	0.3568	1225	34	158	4.0	1248	31.3	29
19	0.2183	0.3578	1237	34	161	4.0	1409	35.4	33
20	0.3469	0.3597	1250	34	145	3.6	1554	39.0	37
21	0.4773	0.3627	1262	35	190	4.8	1744	43.8	41
22	0.6103	0.3669	1275	35	194	4.9	1938	48.7	46
23	0.7468	0.3723	1288	36	201	5.0	2139	53.7	51
24	0.8879	0.3792	1302	36	204	5.1	2343	58.8	56
25	1.0349	0.3877	1316	37	214	5.4	2557	64.2	62
26	1.1892	0.3983	1330	38	217	5.4	2774	69.6	67
27	1.3529	0.4114	1346	39	244	6.1	3018	75.8	73
28	1.5286	0.4277	1363	41	232	5.8	3250	81.6	79
29	1.7201	0.4482	1381	43	204	5.1	3454	86.7	84
30	1.9324	0.4745	1402	45	177	4.4	3631	91.2	89
31	2.1737	0.5094	1425	49	120	3.0	3751	94.2	93
32	2.4572	0.5579	1452	53	115	2.9	3866	97.1	96
33	2.8073	0.6302	1485	60	69	1.7	3935	98.8	98
34	3.2784	0.7535	1530	72	29	0.7	3964	99.5	99
35	4.0418	1.0359	1603	99	16	0.4	3980	99.9	99
36	5.2987	1.8470	1724	177	3	0.1	3983	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 7

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.0233	1.8426	1075	199	0	0.0	0	0.0	0
1	-3.7768	1.0295	1075	111	0	0.0	0	0.0	0
2	-3.0248	0.7469	1075	81	0	0.0	0	0.0	0
3	-2.5623	0.6242	1075	67	3	0.1	3	0.1	1
4	-2.2189	0.5525	1075	60	9	0.2	12	0.3	1
5	-1.9408	0.5046	1075	54	25	0.6	37	0.9	1
6	-1.7040	0.4701	1075	51	30	0.8	67	1.7	1
7	-1.4956	0.4441	1075	48	42	1.1	109	2.7	2
8	-1.3076	0.4238	1082	46	54	1.4	163	4.1	3
9	-1.1350	0.4076	1101	44	83	2.1	246	6.2	5
10	-0.9744	0.3946	1118	43	114	2.9	360	9.1	8
11	-0.8229	0.3840	1134	41	106	2.7	466	11.7	10
12	-0.6788	0.3755	1150	40	124	3.1	590	14.8	13
13	-0.5405	0.3686	1165	40	131	3.3	721	18.1	16
14	-0.4067	0.3632	1179	39	145	3.6	866	21.8	20
15	-0.2764	0.3591	1193	39	149	3.7	1015	25.5	24
16	-0.1485	0.3562	1207	38	172	4.3	1187	29.9	28
17	-0.0224	0.3544	1221	38	172	4.3	1359	34.2	32
18	0.1029	0.3536	1234	38	191	4.8	1550	39.0	37
19	0.2280	0.3540	1248	38	206	5.2	1756	44.2	42
20	0.3537	0.3554	1261	38	208	5.2	1964	49.4	47
21	0.4809	0.3580	1275	39	210	5.3	2174	54.7	52
22	0.6103	0.3618	1289	39	225	5.7	2399	60.4	58
23	0.7430	0.3669	1303	40	202	5.1	2601	65.5	63
24	0.8800	0.3736	1318	40	245	6.2	2846	71.6	69
25	1.0226	0.3820	1333	41	206	5.2	3052	76.8	74
26	1.1724	0.3926	1350	42	194	4.9	3246	81.7	79
27	1.3316	0.4057	1367	44	184	4.6	3430	86.3	84
28	1.5026	0.4221	1385	45	145	3.6	3575	90.0	88
29	1.6893	0.4428	1405	48	127	3.2	3702	93.2	92
30	1.8968	0.4694	1428	51	110	2.8	3812	95.9	95
31	2.1333	0.5047	1453	54	73	1.8	3885	97.8	97
32	2.4120	0.5536	1483	60	46	1.2	3931	98.9	98
33	2.7574	0.6266	1520	68	31	0.8	3962	99.7	99
34	3.2241	0.7505	1571	81	8	0.2	3970	99.9	99
35	3.9833	1.0340	1653	111	4	0.1	3974	100.0	99
36	5.2374	1.8460	1788	199	0	0.0	3974	100.0	100

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 8

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.0257	1.8383	1050	202	0	0.0	0	0.0	0
1	-3.7895	1.0228	1050	112	0	0.0	0	0.0	0
2	-3.0499	0.7394	1050	81	1	0.0	1	0.0	1
3	-2.5974	0.6169	1050	68	3	0.1	4	0.1	1
4	-2.2622	0.5457	1050	60	2	0.1	6	0.2	1
5	-1.9909	0.4984	1050	55	9	0.2	15	0.4	1
6	-1.7598	0.4646	1050	51	11	0.3	26	0.7	1
7	-1.5561	0.4391	1050	48	30	0.8	56	1.5	1
8	-1.3722	0.4194	1053	46	41	1.1	97	2.7	2
9	-1.2030	0.4037	1071	44	63	1.7	160	4.4	4
10	-1.0452	0.3911	1088	43	69	1.9	229	6.3	5
11	-0.8963	0.3810	1105	42	88	2.4	317	8.7	7
12	-0.7545	0.3728	1120	41	84	2.3	401	11.0	10
13	-0.6180	0.3662	1135	40	106	2.9	507	13.9	12
14	-0.4859	0.3610	1150	40	110	3.0	617	16.9	15
15	-0.3570	0.3571	1164	39	120	3.3	737	20.2	19
16	-0.2306	0.3543	1178	39	156	4.3	893	24.5	22
17	-0.1057	0.3526	1192	39	149	4.1	1042	28.6	27
18	0.0183	0.3519	1205	39	185	5.1	1227	33.6	31
19	0.1421	0.3521	1219	39	162	4.4	1389	38.1	36
20	0.2665	0.3534	1233	39	155	4.3	1544	42.3	40
21	0.3922	0.3558	1247	39	238	6.5	1782	48.9	46
22	0.5200	0.3593	1261	40	184	5.0	1966	53.9	51
23	0.6508	0.3641	1275	40	222	6.1	2188	60.0	57
24	0.7855	0.3703	1290	41	195	5.3	2383	65.3	63
25	0.9255	0.3782	1305	42	202	5.5	2585	70.9	68
26	1.0723	0.3881	1321	43	174	4.8	2759	75.7	73
27	1.2275	0.4004	1338	44	193	5.3	2952	80.9	78
28	1.3939	0.4159	1357	46	170	4.7	3122	85.6	83
29	1.5748	0.4356	1377	48	168	4.6	3290	90.2	88
30	1.7754	0.4610	1399	51	128	3.5	3418	93.7	92
31	2.0031	0.4950	1424	54	107	2.9	3525	96.7	95
32	2.2709	0.5424	1453	60	57	1.6	3582	98.2	97
33	2.6024	0.6139	1490	68	41	1.1	3623	99.3	99
34	3.0511	0.7369	1539	81	16	0.4	3639	99.8	99
35	3.7871	1.0211	1620	112	6	0.2	3645	99.9	99
36	5.0207	1.8373	1756	202	2	0.1	3647	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Reading Grade 11

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.0885	1.8410	1000	217	0	0.0	0	0.0	0
1	-3.8453	1.0277	1000	121	0	0.0	0	0.0	0
2	-3.0960	0.7458	1000	88	0	0.0	0	0.0	0
3	-2.6343	0.6242	1000	74	2	0.1	2	0.1	1
4	-2.2901	0.5537	1000	65	6	0.2	8	0.2	1
5	-2.0103	0.5068	1000	60	8	0.2	16	0.4	1
6	-1.7709	0.4732	1000	56	12	0.3	28	0.7	1
7	-1.5593	0.4479	1000	53	26	0.7	54	1.4	1
8	-1.3677	0.4282	1000	50	31	0.8	85	2.2	2
9	-1.1912	0.4125	1015	49	46	1.2	131	3.3	3
10	-1.0264	0.3999	1034	47	56	1.4	187	4.8	4
11	-0.8706	0.3896	1053	46	52	1.3	239	6.1	5
12	-0.7222	0.3813	1070	45	67	1.7	306	7.8	7
13	-0.5795	0.3745	1087	44	83	2.1	389	9.9	9
14	-0.4412	0.3691	1103	44	82	2.1	471	12.0	11
15	-0.3066	0.3649	1119	43	95	2.4	566	14.4	13
16	-0.1746	0.3618	1135	43	116	3.0	682	17.4	16
17	-0.0445	0.3597	1150	42	122	3.1	804	20.5	19
18	0.0845	0.3586	1165	42	135	3.4	939	24.0	22
19	0.2129	0.3584	1180	42	140	3.6	1079	27.5	26
20	0.3416	0.3592	1196	42	167	4.3	1246	31.8	30
21	0.4712	0.3610	1211	43	176	4.5	1422	36.3	34
22	0.6025	0.3639	1226	43	222	5.7	1644	41.9	39
23	0.7363	0.3679	1242	43	229	5.8	1873	47.8	45
24	0.8736	0.3733	1258	44	262	6.7	2135	54.5	51
25	1.0155	0.3804	1275	45	232	5.9	2367	60.4	57
26	1.1634	0.3893	1292	46	258	6.6	2625	67.0	64
27	1.3193	0.4007	1311	47	247	6.3	2872	73.3	70
28	1.4854	0.4151	1330	49	258	6.6	3130	79.9	77
29	1.6652	0.4337	1352	51	221	5.6	3351	85.5	83
30	1.8635	0.4581	1375	54	205	5.2	3556	90.7	88
31	2.0880	0.4910	1401	58	156	4.0	3712	94.7	93
32	2.3512	0.5376	1433	63	100	2.6	3812	97.3	96
33	2.6769	0.6085	1471	72	78	2.0	3890	99.3	98
34	3.1181	0.7312	1523	86	17	0.4	3907	99.7	99
35	3.8450	1.0162	1609	120	12	0.3	3919	100.0	99
36	5.0711	1.8342	1753	216	0	0.0	3919	100.0	100

Appendix N: Raw-to-Scaled Score Conversion Tables

Science Grade 8

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-5.0178	1.8435	1050	214	0	0.0	0	0.0	0
1	-3.7687	1.0315	1050	120	0	0.0	0	0.0	0
2	-3.0122	0.7502	1050	87	0	0.0	0	0.0	0
3	-2.5445	0.6285	1050	73	0	0.0	0	0.0	0
4	-2.1955	0.5576	1050	65	3	0.1	3	0.1	1
5	-1.9118	0.5102	1050	59	6	0.2	9	0.3	1
6	-1.6694	0.4760	1050	55	13	0.4	22	0.7	1
7	-1.4554	0.4502	1050	52	21	0.6	43	1.3	1
8	-1.2619	0.4301	1050	50	23	0.7	66	2.0	2
9	-1.0840	0.4141	1062	48	40	1.2	106	3.3	3
10	-0.9180	0.4012	1082	47	46	1.4	152	4.7	4
11	-0.7613	0.3909	1100	45	68	2.1	220	6.8	6
12	-0.6119	0.3825	1117	44	75	2.3	295	9.1	8
13	-0.4682	0.3760	1134	44	96	3.0	391	12.0	11
14	-0.3288	0.3709	1150	43	84	2.6	475	14.6	13
15	-0.1927	0.3672	1166	43	135	4.2	610	18.8	17
16	-0.0588	0.3647	1181	42	119	3.7	729	22.4	21
17	0.0736	0.3635	1197	42	159	4.9	888	27.3	25
18	0.2057	0.3634	1212	42	190	5.8	1078	33.1	30
19	0.3381	0.3646	1227	42	198	6.1	1276	39.2	36
20	0.4719	0.3670	1243	43	205	6.3	1481	45.5	42
21	0.6079	0.3708	1259	43	215	6.6	1696	52.2	49
22	0.7473	0.3761	1275	44	208	6.4	1904	58.5	55
23	0.8913	0.3832	1292	45	197	6.1	2101	64.6	62
24	1.0415	0.3923	1309	46	205	6.3	2306	70.9	68
25	1.1998	0.4039	1328	47	217	6.7	2523	77.6	74
26	1.3688	0.4188	1347	49	168	5.2	2691	82.7	80
27	1.5519	0.4378	1368	51	123	3.8	2814	86.5	85
28	1.7542	0.4627	1392	54	123	3.8	2937	90.3	88
29	1.9833	0.4961	1419	58	104	3.2	3041	93.5	92
30	2.2520	0.5431	1450	63	81	2.5	3122	96.0	95
31	2.5840	0.6142	1488	71	71	2.2	3193	98.2	97
32	3.0328	0.7368	1540	86	36	1.1	3229	99.3	99
33	3.7683	1.0207	1626	119	16	0.5	3245	99.8	99
34	5.0014	1.8370	1769	213	7	0.2	3252	100.0	99

Appendix N: Raw-to-Scaled Score Conversion Tables

Science Grade 11

Raw	Meas	MeasSE	SS	SSSE	Freq	Freq%	Cum	Cum%	Pct
0	-4.8807	1.8380	1100	166	0	0.0	0	0.0	0
1	-3.6453	1.0223	1100	92	0	0.0	0	0.0	0
2	-2.9065	0.7389	1100	67	1	0.0	1	0.0	1
3	-2.4547	0.6166	1100	56	4	0.1	5	0.1	1
4	-2.1196	0.5458	1100	49	11	0.3	16	0.5	1
5	-1.8481	0.4989	1100	45	20	0.6	36	1.0	1
6	-1.6164	0.4654	1100	42	40	1.1	76	2.1	2
7	-1.4116	0.4405	1118	40	46	1.3	122	3.4	3
8	-1.2264	0.4212	1135	38	82	2.3	204	5.8	5
9	-1.0555	0.4061	1150	37	113	3.2	317	9.0	7
10	-0.8956	0.3941	1164	36	103	2.9	420	11.9	10
11	-0.7442	0.3845	1178	35	153	4.3	573	16.2	14
12	-0.5993	0.3769	1191	34	169	4.8	742	21.0	19
13	-0.4596	0.3710	1204	33	183	5.2	925	26.1	24
14	-0.3237	0.3665	1216	33	180	5.1	1105	31.2	29
15	-0.1906	0.3633	1228	33	193	5.5	1298	36.7	34
16	-0.0595	0.3612	1240	33	193	5.5	1491	42.1	39
17	0.0707	0.3604	1252	32	194	5.5	1685	47.6	45
18	0.2005	0.3606	1263	33	207	5.8	1892	53.4	51
19	0.3310	0.3620	1275	33	230	6.5	2122	59.9	57
20	0.4629	0.3646	1287	33	217	6.1	2339	66.1	63
21	0.5971	0.3685	1299	33	186	5.3	2525	71.3	69
22	0.7348	0.3739	1311	34	188	5.3	2713	76.6	74
23	0.8771	0.3810	1324	34	159	4.5	2872	81.1	79
24	1.0257	0.3902	1338	35	128	3.6	3000	84.7	83
25	1.1824	0.4020	1352	36	130	3.7	3130	88.4	87
26	1.3498	0.4170	1367	38	112	3.2	3242	91.6	90
27	1.5315	0.4362	1383	39	92	2.6	3334	94.2	93
28	1.7325	0.4614	1401	42	91	2.6	3425	96.8	95
29	1.9605	0.4952	1422	45	50	1.4	3475	98.2	97
30	2.2284	0.5426	1446	49	35	1.0	3510	99.2	99
31	2.5602	0.6142	1476	55	16	0.5	3526	99.6	99
32	3.0094	0.7373	1516	66	7	0.2	3533	99.8	99
33	3.7461	1.0215	1583	92	4	0.1	3537	99.9	99
34	4.9806	1.8377	1694	166	3	0.1	3540	100.0	99

Appendix O:

PSSA and PSSA-M Demographics Comparison

PSSA and PSSA-M Demographics Comparison

In response to a recommendation by Pennsylvania's Technical Advisory Committee (TAC), tables were assembled to provide a demographic comparison of students taking the PSSA-Modified (PSSA-M) assessment for the current year with those of the previous year. The recommendation included the advisability of comparisons with students taking the standard assessment (PSSA). The following tables contain the percent of assessed students contributing to state summary statistics for gender and ethnicity categories. Since the PSSA-M Mathematics was first assessed in 2010, there are tables for both 2011 and 2010. Assessment of the PSSA-M Reading and Science was introduced in the spring 2011 assessment, permitting a comparison with current year regular PSSA demographics. Tables providing a gender and ethnicity breakdown by subject, type of test, and year are the following:

- 2011 PSSA-M Mathematics
- 2010 PSSA-M Mathematics
- 2011 PSSA Mathematics
- 2010 PSSA Mathematics
- 2011 PSSA-M Reading
- 2011 PSSA Reading
- 2011 PSSA-M Science
- 2011 PSSA Science

A comparison between demographics for students receiving the PSSA-M Mathematics in 2011 with that of 2010 there is a caveat to keep in mind. In 2010 the PSSA-M statistics were based on data from the Pre-Appeals file while 2011 PSSA-M results were obtained from the Post-Appeals file. Likewise, comparisons between the 2010 PSSA-M and 2010 PSSA have the same issue, which means that there were some differences due to the data file used. In 2011 both the PSSA-M and PSSA data were based on Post-Appeals files, thereby removing this as a potential interpretive issue. In the tables displaying PSSA-M Mathematics results for 2010 and 2011, the following may be observed:

- The number of students participating in the PSSA-M Mathematics increased from 2010 to 2011.
- The demographic composition of those assessed in 2011 appear to be slightly more male and minority (black and Latino/Hispanic) than those assessed in 2010; however, the difference in data set used, as explained above, may partially account for this.

Demographic Comparisons Between PSSA-M and PSSA Across Subject Areas

The percent of students in major demographic categories were averaged across the assessed grade levels (Grades 4-8, 11 for PSSA-M and PSSA Mathematics and Reading and Grades 8 and 11 for PSSA-M and PSSA Science). The consistent finding for each subject, as noted in the table below, is that the demographic composition of the PSSA-M group was clearly more male and minority (black and Latino/Hispanic) than those taking the standard PSSA.

**Average Percent of Students in Selected Demographic Categories
Receiving Scores on the 2011 PSSA-M or the 2011 PSSA**

Subject	Test	Male	Black	Lat/Hisp	White
Mathematics	PSSA-M	59.2	21.7	9.6	66.2
	PSSA	50.9	14.8	7.4	73.3
Reading	PSSA-M	62.9	20.1	9.4	67.9
	PSSA	50.7	14.8	7.7	72.9
Science	PSSA-M	62.4	19.4	9.4	68.8
	PSSA	50.4	13.9	6.8	75.0

Gender and Ethnicity Percent Breakdown for 2011 PSSA-M: Mathematics

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	40.3	42.8	41.2	39.8	40.7	39.6
Male	59.6	57.1	58.8	60.0	59.1	60.3
Race/Ethnicity						
Amer. Indian or Alaskan Native	0.3	0.2	0.1	0.3	0.0	0.2
Asian or Pacific Islander	0.8	1.1	0.8	0.8	1.0	0.7
Black/African American non-Hispanic	23.7	23.0	20.7	21.7	21.0	20.2
Latino/Hispanic	11.1	10.5	10.2	8.5	9.6	7.6
White non-Hispanic	62.2	64.1	66.7	66.9	67.1	70.1
Multi-Racial/Ethnic	1.8	0.9	1.4	1.6	1.1	0.9
Assessed Students in State Summaries	2,375	3,366	3,600	3,972	4,114	4,269

Gender and Ethnicity Percent Breakdown for 2010 PSSA-M: Mathematics

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	42.0	41.9	39.9	41.0	39.6	37.7
Male	57.6	57.6	59.9	58.5	59.7	61.2
Race/Ethnicity						
Amer. Indian or Alaskan Native	0.2	0.2	0.3	0.1	0.3	0.1
Asian or Pacific Islander	1.2	0.7	1.1	1.3	0.7	0.8
Black/African American non-Hispanic	19.1	18.2	18.9	18.1	18.4	19.8
Latino/Hispanic	8.9	8.7	8.6	8.4	8.3	6.6
White non-Hispanic	68.8	69.9	69.4	70.5	70.8	70.6
Multi-Racial/Ethnic	1.3	1.6	1.3	1.0	0.9	1.0
Assessed Students in State Summaries	2,169	2,552	2,700	2,817	3,019	3,536

Gender and Ethnicity Percent Breakdown for 2011 PSSA: Mathematics

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	49.0	49.0	48.8	49.2	48.6	49.6
Male	51.0	51.0	51.1	50.7	51.2	50.3
Race/Ethnicity						
American Indian or Alaskan Native	0.1	0.2	0.1	0.1	0.1	0.2
Asian or Pacific Islander	3.4	3.3	3.0	3.0	2.9	3.0
Black/African American non-Hispanic	15.1	15.0	15.2	15.1	14.7	13.9
Latino/Hispanic	8.3	7.9	7.6	7.4	7.0	6.2
White non-Hispanic	71.5	72.2	72.8	73.3	74.1	75.9
Multi-Racial/Ethnic	1.5	1.4	1.1	1.0	0.9	0.8
Assessed Students in State Summaries	125,604	126,578	126,630	126,993	126,786	127,797

Gender and Ethnicity Percent Breakdown for 2010 PSSA: Mathematics

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	48.9	48.8	48.8	49.2	48.6	49.5
Male	50.9	51.0	51.1	50.7	51.2	50.2
Race/Ethnicity						
American Indian or Alaskan Native	0.2	0.1	0.1	0.1	0.1	0.2
Asian or Pacific Islander	3.2	3.0	3.0	3.0	2.9	2.8
Black/African American non-Hispanic	15.3	15.5	15.2	15.1	14.7	13.7
Latino/Hispanic	7.7	7.8	7.6	7.4	7.0	5.8
White non-Hispanic	72.2	72.3	72.8	73.3	74.1	76.6
Multi-Racial/Ethnic	1.3	1.2	1.1	1.0	0.9	0.6
Assessed Students in State Summaries	126,333	126,419	126,288	127,685	129,983	129,910

Gender and Ethnicity Percent Breakdown for 2011 PSSA-M: Reading

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	35.2	38.5	37.2	37.5	37.6	36.2
Male	64.8	61.4	62.8	62.4	62.3	63.4
Race/Ethnicity						
American Indian or Alaskan Native	0.2	0.3	0.1	0.3	0.1	0.2
Asian or Pacific Islander	0.9	1.2	0.7	1.1	1.2	0.8
Black/African American non-Hispanic	19.9	21.0	19.0	19.4	20.5	21.0
Latino/Hispanic	9.9	9.5	9.9	9.4	9.9	8.0
White non-Hispanic	67.7	66.8	68.8	68.2	67.4	68.7
Multi-Racial/Ethnic	1.4	1.2	1.3	1.5	1.0	0.8
Assessed Students in State Summaries	3,388	3,947	3,983	3,974	3,647	3,919

Gender and Ethnicity Percent Breakdown for 2011 PSSA: Reading

Demographic or Educational Characteristic	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8	Gr. 11
Gender						
Female	49.2	49.2	49.1	49.0	49.4	49.7
Male	50.8	50.8	50.9	51.0	50.6	50.3
Race/Ethnicity						
American Indian or Alaskan Native	0.1	0.2	0.2	0.1	0.1	0.2
Asian or Pacific Islander	3.4	3.3	3.1	3.1	3.0	3.0
Black/African American non-Hispanic	15.2	15.0	15.2	14.9	14.8	13.9
Latino/Hispanic	8.4	7.9	8.0	7.9	7.6	6.2
White non-Hispanic	71.4	72.2	72.2	72.7	73.3	75.9
Multi-Racial/Ethnic	1.5	1.4	1.3	1.3	1.1	0.8
Assessed Students in State Summaries	124,535	125,963	126,170	126,902	127,125	127,997

Gender and Ethnicity Percent Breakdown for 2011 PSSA-M: Science

Demographic or Educational Characteristic	Gr. 8	Gr. 11
Gender		
Female	38.2	36.7
Male	61.7	63.1
Race/Ethnicity		
American Indian or Alaskan Native	0.1	0.3
Asian or Pacific Islander	1.3	0.8
Black/African American non-Hispanic	19.1	19.7
Latino/Hispanic	10.4	8.3
White non-Hispanic	67.9	69.7
Multi-Racial/Ethnic	1.0	1.0
Assessed Students in State Summaries	3,252	3,540

Gender and Ethnicity Percent Breakdown for 2011 PSSA: Science

Demographic or Educational Characteristic	Gr. 8	Gr. 11
Gender		
Female	49.3	49.6
Male	50.6	50.3
Race/Ethnicity		
American Indian or Alaskan Native	0.1	0.2
Asian or Pacific Islander	3.1	3.0
Black/African American non-Hispanic	14.7	13.1
Latino/Hispanic	7.6	6.0
White non-Hispanic	73.2	76.8
Multi-Racial/Ethnic	1.1	0.8
Assessed Students in State Summaries	127,075	125,307