TECHNICAL REPORT

for the
Pennsylvania
System of School Assessment

2010 Modified Grade 12 Fall Retest
Mathematics

Provided by
Data Recognition Corporation

April 2011

Table of Contents

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.

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.)

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).

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 *P*-value and Point-Biserial Correlation). Each test form includes a set of FT items. FT items are not part of any student scores.
Frequency	The number of times that a certain value or range of values (score interval) occurs in a distribution of scores.
Frequency Distribution	A tabulation of scores from low to high or high to low showing the number and/or percent of individuals who obtain each score or who fall within each score interval or category.
Infit/Outfit	Statistical indicators of the agreement of the data and the measurement model (see also Outfit/Infit).
Item Difficulty	For the Rasch model, the dichotomous item difficulty represents the point along the latent trait continuum where an examinee has a 0.50 probability of making a correct response. For a polytomous item, the difficulty is the average of the item's step difficulties (see 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 *p*-value (see also *P*-value). The logit difficulty scale is inversely related to *p*-values. A higher logit value would represent a relatively harder item, while a lower logit value would represent a relatively easier item.
Mean	Also referred to as the arithmetic mean of a set of scores, 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.
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.

Term	Common Definition
Median	The middle point or score in a set of rank-ordered observations that divides the distribution into two equal parts such that each part contains 50 percent of the total data set. More simply put, half of the scores are below the median value and half of the scores are above the median value. As an example, the median for the following ranked set of scores {2, 3, 6, 8, 9} is 6.
Multiple-Choice Item	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).
N-count	Sometimes designated as *N* or *n*, 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}, $n = 6$.
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.
Percent Correct	When referring to an individual item, the percent correct is the item's *p*-value expressed as a percent (instead of a proportion). When referring to a total test score, it is the percentage of the total number of points that a student received. The percent correct score is obtained by dividing the student's raw score by the total number of 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.

Term	Common Definition
Percentile	The score or point in a score distribution at or below which a given percentage of scores fall. It should be emphasized that it is a value on the score scale, not the associated percentage (although sometimes in casual usage this misinterpretation is made). For example, if 72 percent of the students score at or below a Scaled Score of 1500 on a given test, then the Scaled Score of 1500 would be considered the 72nd percentile. As another example, the median is the 50th percentile.
Percentile Rank	The percentage of scores in a specified distribution falling at/below a certain point on a score distribution. Percentile Ranks range in value from 1 to 99, and indicate the status or relative standing of an individual within a specified group, by indicating the percent of individuals in that group who obtained equal or lower scores. An individual's percentile rank can vary depending on which group is used to determine the ranking. As suggested above, 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.
P-value	An index indicating an item's difficulty for some specified group (perhaps grade). It is calculated as the proportion (sometimes percent) of students in the group who answer an item correctly. *P*-values range from 0.0 to 1.0 on the proportion scale. Lower values correspond to more difficult items and higher values correspond to easier items. *P*-values are usually provided for multiple-choice items or other items worth one point. For open-ended items or items worth more than one point, difficulty on a *p*-value-like scale can be estimated by dividing the item mean score by the maximum number of points possible for the item (see also Logit).

Term	Common Definition
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.
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.

Term	Common Definition
Standard Deviation (SD)	A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statistic is always greater than or equal to zero. If all of the scores in a distribution are identical, the standard deviation is equal to zero. The further the scores are away from each other in value, the greater the standard deviation. This statistic is calculated using the information about the deviations (distances) between each score and the distribution's mean. It is equivalent to the square root of the variance statistic. The standard deviation is a commonly used method of examining a distribution's variability since the standard deviation is expressed in the same units as the data.
Standard Error of Measurement (SEM)	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.

PSSA-M: The Modified Pennsylvania System of School Assessment

In 2010, the *Pennsylvania System of School Assessment Modified* (PSSA-M) became operational with a mathematics modified assessment at Grades 4–8 and 11. The modified assessment was created to address the unique needs of students whose learning disabilities are not severe enough to warrant taking an alternate assessment and yet interfere significantly with their ability to respond optimally on the standard state assessment.

By issuing additional regulations permitting states to develop assessments based on modified academic standards for the approximately 2% of students with disabilities, the U.S. Department of Education recognized the need for a modified assessment. The modified assessment is aligned to a set of modified academic achievement standards designed to measure the same grade-level content as the state's general assessment. The modified assessment allows for a closer compliance with the No Child Left Behind intent that all students be included in state assessment and accountability systems.

The broad purpose of the Modified State Assessments is to provide information to teachers and schools to guide the improvement of curricula and instructional strategies for these students to achieve the academic standards. *The Department strongly discourages the use of this testing information for "ranking" schools.*

PSSA-M Grade 12 Fall Retest

Chapter 4 Regulations state that students who score at the *Proficient* or *Advanced* level on the state assessment in mathematics administered in Grade 11 or Grade 12 are eligible to receive Certificates of *Proficiency* and/or Certificates of *Distinction*. The purpose of the PSSA-M Grade 12 Retest is to provide students who did not achieve a *Proficient* level or higher on the Grade 11 modified assessment the opportunity to achieve a proficient or higher level and receive certificates.

A Grade 12 student is ELIGIBLE for the PSSA-M Grade 12 Retest if:

- Student achieved *Basic* or *Below Basic* performance level on that specific modified assessment, **OR**
- Student's PSSA-M performance level is *unknown,* and attempts to determine student's performance level by contacting the student's former school *cannot confirm* that the student achieved *Proficient* or *Advanced* performance level on the PSSA-M.

A Student is NOT ELIGIBLE for the PSSA-M Grade 12 Retest if:

- Student achieved *Proficient* or *Advanced* performance level on that specific modified assessment, **OR**
- Student participated in the PSSA or PASA, **OR**
- Student is not currently in Grade 12, **OR**
- Student did not participate in the PSSA or PSSA-M.

The Grade 12 Retest is not a mandatory assessment, so a student may choose not to participate without parental request for exclusion and school/district officials are not required to authorize student exclusions. The Pennsylvania Department of Education (PDE) recommends schools that do not require student retest participation to encourage eligible students to discuss the retest with parents/guardians. Though the final decision about whether a student should participate in the retest is made by the student and his/her parents/guardians, the district must provide eligible students with the opportunity to participate.

This technical report provides the retest results for the PSSA-M assessments including Item Analysis, Raw-to-Scale Score Conversions, and Performance Levels results.

Item Analysis

Multiple-Choice (MC) Items

The most familiar indices of item performance for MC items are those that reflect item difficulty (i.e., *proportion correct*, generally referred to as a "*p*-value") and those that reflect item discrimination (often represented by the *point-biserial correlation* coefficient). The point-biserial correlation for an item is the Pearson product-moment correlation between students' item scores and their total test scores. It is expected that students who respond to the item correctly should have a higher total test score mean than students who respond incorrectly. An item that performs as expected should have a positive point-biserial correlation coefficient.

The item-level analyses done for the Grade 12 retests' MC items also included statistics for the incorrect responses (i.e., distractors) such as proportion of students selecting each distractor, and the point-biserial correlation for each distractor. The results from distractor analyses provide additional information for understanding the item's behavior. For example, the percent selecting each response is an indicator of which responses are particularly attractive.

Item level statistics for the MC items for mathematics can be found in Appendix A. These statistics include the number of students attempting each item, *p*-values, proportions of students selecting each response, item-total correlations, and point-biserial correlations for each response category. The tabled values indicate that the MC items on the PSSA-M retests performed as expected.

Open-Ended (OE) Items

A first step when evaluating OE item performance is to examine the item's score-point distribution (percentages of students in each scoring category) as this can provide a rough "snap shot" of an item's performance. For example, a four-point OE item with a vast majority of students receiving *ones* and/or *fours* with virtually no other scores occurring would be unusual. Another useful statistic is the correlation between the item scores and total test scores. Similar to the MC item's point-biserial index, this correlation reflects how an OE item discriminates between low scoring and high scoring students. The students with higher test scores are expected to have higher mean score on the item.

Item level statistics for the mathematics OE items for can be found in Appendix C. In the appendix, the "B" code denotes a blank non-response, the "K" code denotes an off-task response, and the "U" code denotes an unreadable response. The score-point distributions and the item-total score correlations indicate that all the OE items performed as expected.

Raw-to-Scaled Score Conversions

A *scaled score*, in the simplest sense, is a transformed raw score. For the PSSA-M retest, this transformation was done in two steps. First, the students attempting the Grade 12 retest were scored using the Rasch scaling model by anchoring the Rasch item difficulties at the values calibrated from the 2010 spring operational data. This scoring transformed student raw scores into Rasch logit scores which typically fall between -5.0 to 5.0. This transformation is non-linear and often referred to as the "Raw-to-Logit conversion." Appendix B presents the anchored Rasch item logit difficulties, their corresponding standard errors, and fit statistics for all the mathematics MC items.

The second step is to convert these logit scores into PSSA-M score scales using linear transformations. Table 1 gives the linear logit-to-scaled score conversion functions for Grade 12 PSSA-M mathematics.

Table 1: Logit-to-Scaled Score Conversions

Content	Transformation
Mathematics	115.64X + 1213.51

Note. X denotes the Rasch logit ability values.

Scaled scores have several interpretive advantages over raw scores, as illustrated in the following example. A raw score of 30, for instance, is almost meaningless unless the reader is also given how many points are possible. The same score has a different meaning if it is based on a thirty-item test or on a sixty-item test. Total points attained are transformed to percent correct scores to remove the effect of test length. In the same way, a score based on sixty difficult items is different from the same score based on sixty easy items. Total points attained are transformed to scaled scores to remove the effects of test length and item difficulty.

In 2010, a lowest obtainable scaled score (LOSS) of 1075 was implemented for the PSSA-M mathematics exams. However, the highest obtainable scale scores for PSSA-M tests are not fixed. They are allowed to float for each subject and grade. The RS-SS conversion tables for mathematics can be found in Appendix D. The students' raw scores were transformed to the scaled scores based on those tables.

Summary of the PSSA-M Grade 12 Retest Results

Scaled Score Results

The performance of students attempting the fall retests was compared with the performance of students attempting 2010 spring operational tests. Table 2 summarizes the spring and fall test results for these two groups of students including the mean, standard deviation (SD), maximum, and minimum scaled scores as well as the reliability of the assessments. The mean scaled scores on the fall retest were lower than the mean scores on the spring test, indicating that the students who took the fall retest did not perform as well as the students who took the previous spring test. These results are expected in a retest situation since the group taking the retest is typically comprised of students who had not performed well on the previous administration.

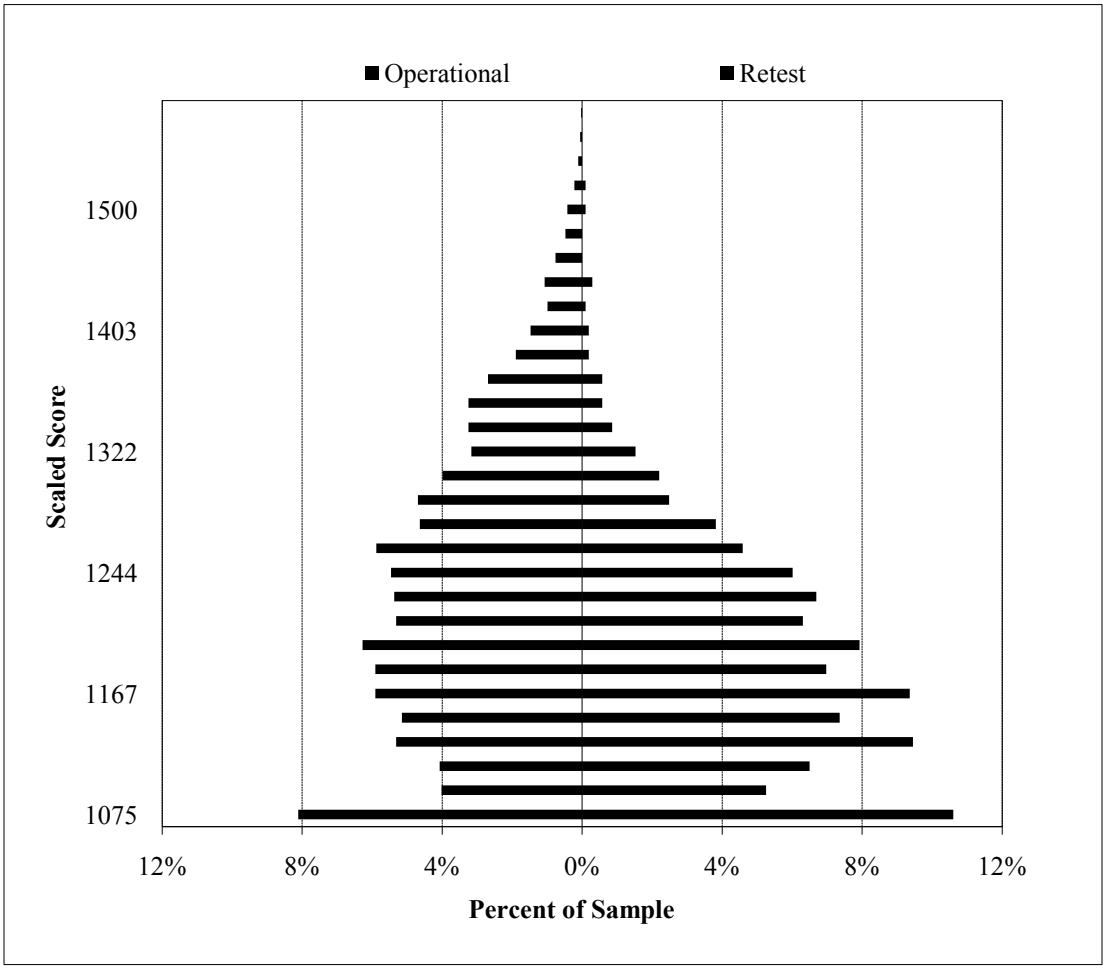
The standard deviations were also lower for the retest group. Smaller standard deviations were the result of a more homogeneous score distribution and an artifact of the aforementioned group of retesters. The relatively lower test reliabilities (based on Coefficient Alpha) for mathematics can also be attributed to the decreased variability in test scores.

Table 2: Operational and Retest Summary Statistics (Scaled Score Metric)

	Mathematics	
	Oper.	**Retest**
N	3536	1047
Mean	1228.4	1184.1
St. Dev.	99.6	73.7
Min	1075	1075
Max	1724	1529
Reli.	0.81	0.68

Figure 1 contrasts the fall retest frequency distributions against the spring operational frequency distributions for mathematics. As seen from Figure 1, the distributions of scaled scores for the fall mathematics retests are more positively-skewed relative to their operational counterparts with lower test scores occurring with much greater frequency than higher scores. In contrast, the spring operational test scores are more negatively distributed.

Figure 1: Mathematics Operational and Retest Scaled Score Frequency Distributions



Performance Level Results

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.

The scores that correspond with each performance level are located in Table 3. The cumulative percentage of students who achieved a *Proficient* or *Advanced* performance level on the mathematics retest was 13.0. Approximately 87% of the students who took the retest still scored *Basic* or *Below Basic* levels.

Table 3: Grade 12 Retest Performance Standards

	Mathematics		
Performance Level	**Scaled Score**	**Frequency**	**Percent**
Advanced-M	1403 and up	8	0.8
Proficient-M	1275-1402	128	12.2
Basic-M	1150-1274	578	55.2
Below Basic-M	1149 and below	333	31.8

Note. Numbers may not add exactly to 100% due to rounding.

Of the students with scores for both the spring operational and the fall retest administrations, 62.1% of the students remained at the same performance level in mathematics, while 23.4% transitioned to a higher level and 14.5% regressed to a lower level.

Appendix A: 2010 Grade 12 Fall Mathematics Retest Multiple-Choice Item Statistics

Seq.	N	P-Value	P(-)	P(*)	Tot. Corr.
1	1047	0.790	0.007	0.000	0.239
2	1047	0.558	0.007	0.000	0.200
3	1047	0.459	0.005	0.000	0.155
4	1047	0.400	0.005	0.000	0.168
5	1047	0.406	0.002	0.000	0.255
6	1047	0.640	0.000	0.000	0.199
7	1047	0.404	0.002	0.000	0.033
8	1047	0.324	0.001	0.000	0.155
9	1047	0.445	0.000	0.000	0.143
10	1047	0.844	0.001	0.000	0.073
11	1047	0.471	0.002	0.000	0.182
12	1047	0.426	0.001	0.000	0.311
13	1047	0.354	0.008	0.000	0.210
14	1047	0.464	0.006	0.000	0.189
15	1047	0.401	0.003	0.000	0.235
17	1047	0.361	0.007	0.000	0.129
18	1047	0.293	0.003	0.000	0.185
19	1047	0.359	0.008	0.000	0.115
20	1047	0.329	0.004	0.000	0.117
21	1047	0.391	0.004	0.001	0.195
22	1047	0.404	0.003	0.000	0.339
23	1047	0.435	0.001	0.000	0.178
24	1047	0.558	0.003	0.000	0.362
25	1047	0.261	0.004	0.000	0.175
26	1047	0.391	0.006	0.000	0.231
27	1047	0.690	0.005	0.000	0.091
28	1047	0.216	0.004	0.000	0.169
29	1047	0.569	0.001	0.001	0.228
30	1047	0.172	0.006	0.000	0.139
31	1047	0.461	0.002	0.000	0.273

Note. "-" denotes omits; "*" denotes multiple marks.

Appendix B: 2010 Grade 12 Fall Mathematics Retest Multiple-Choice Rasch Item Statistics

Seq.	Anchored Measure	Measure SE	InFit MS	InFit ZSTD	OutFit MS	OutFit ZSTD
1	-1.9019	0.0823	1.05	1.1	1.03	0.5
2	-0.6527	0.0657	1.02	0.8	1.02	0.8
3	0.1395	0.0661	1.07	3.4	1.09	3.2
4	0.1884	0.0663	1.03	1.2	1.04	1.5
5	0.2586	0.0668	0.99	-0.3	0.99	-0.4
6	-0.9624	0.0677	1.00	0.1	1.01	0.4
7	0.2679	0.0669	1.14	5.8	1.17	5.7
8	0.6302	0.0703	1.07	2.1	1.07	1.9
9	-0.0190	0.0654	1.04	2.1	1.05	2.0
10	-1.7270	0.0785	0.85	-3.6	0.89	-1.8
11	-0.0866	0.0652	1.02	1.0	1.02	0.9
12	-0.0548	0.0652	0.92	-4.1	0.91	-4.1
13	0.3464	0.0674	0.99	-0.5	0.99	-0.4
14	-0.0641	0.0652	1.02	0.8	1.01	0.6
15	0.1646	0.0662	0.98	-0.9	0.96	-1.3
17	0.3517	0.0675	1.05	1.9	1.07	2.2
18	0.5626	0.0695	0.97	-1.2	0.98	-0.4
19	0.3932	0.0678	1.06	2.5	1.09	2.8
20	0.6636	0.0707	1.11	3.5	1.16	3.8
21	0.4106	0.0680	1.06	2.4	1.07	2.2
22	-0.0376	0.0653	0.90	-5.4	0.88	-5.2
23	0.0021	0.0654	1.02	0.9	1.02	1.0
24	-0.7397	0.0661	0.93	-3.4	0.92	-3.0
25	0.7029	0.0712	0.95	-1.4	0.95	-1.2
26	0.3157	0.0672	1.01	0.3	1.02	0.6
27	-0.8878	0.0671	1.00	-0.1	1.01	0.4
28	1.1182	0.0779	1.00	-0.1	1.04	0.7
29	-0.5620	0.0653	0.97	-1.6	0.99	-0.6
30	1.1325	0.0781	0.88	-3.0	0.93	-1.3
31	0.0471	0.0656	0.98	-1.2	0.98	-0.9

Appendix C: 2010 Grade 12 Fall Mathematics Retest Open-ended Item Statistics

Item Description			Proportions									Correlations					
Seq.	Max	N	Mean	0	1	2	3	4	B	K	U	Tot. Corr.	0	1	2	3	4
16	4	1047	0.382	0.676	0.292	0.017	0.002	0.012	0.061	0.000	0.000	0.334	-0.306	0.233	0.119	0.091	0.159
32	4	1047	0.514	0.534	0.432	0.023	0.010	0.002	0.060	0.000	0.000	0.403	-0.374	0.304	0.112	0.133	0.138

Note. B = blank; K = off task; U = unreadable.

Appendix D: 2010 Grade 12 Fall Mathematics Retest Raw-to-Scaled Score Conversion Table

Raw Score	Measure	Measure SE	Scaled Score	Scaled Score SE	Freq.	Freq. %	Cum. Freq.	Cum. Freq. %	Percentile
0	-4.9393	1.8430	1075	213	0	0.0	0	0.0	0
1	-3.6912	1.0307	1075	119	0	0.0	0	0.0	0
2	-2.9364	0.7490	1075	87	0	0.0	0	0.0	0
3	-2.4705	0.6271	1075	73	0	0.0	0	0.0	0
4	-2.1233	0.5560	1075	64	4	0.4	4	0.4	1
5	-1.8414	0.5085	1075	59	13	1.2	17	1.6	1
6	-1.6006	0.4744	1075	55	15	1.4	32	3.1	2
7	-1.3881	0.4486	1075	52	25	2.4	57	5.4	4
8	-1.1960	0.4287	1075	50	54	5.2	111	10.6	8
9	-1.0192	0.4128	1096	48	55	5.3	166	15.9	13
10	-0.8541	0.4002	1115	46	68	6.5	234	22.3	19
11	-0.6981	0.3901	1133	45	99	9.5	333	31.8	27
12	-0.5492	0.3820	1150	44	77	7.4	410	39.2	35
13	-0.4058	0.3756	1167	43	98	9.4	508	48.5	44
14	-0.2666	0.3708	1183	43	73	7.0	581	55.5	52
15	-0.1305	0.3672	1198	42	83	7.9	664	63.4	59
16	0.0033	0.3647	1214	42	66	6.3	730	69.7	67
17	0.1358	0.3633	1229	42	70	6.7	800	76.4	73
18	0.2676	0.3629	1244	42	63	6.0	863	82.4	79
19	0.3994	0.3632	1260	42	48	4.6	911	87.0	85
20	0.5317	0.3643	1275	42	40	3.8	951	90.8	89
21	0.6649	0.3659	1290	42	26	2.5	977	93.3	92
22	0.7995	0.3679	1306	43	23	2.2	1000	95.5	94
23	0.9357	0.3701	1322	43	16	1.5	1016	97.0	96
24	1.0734	0.3721	1338	43	9	0.9	1025	97.9	97
25	1.2127	0.3740	1354	43	6	0.6	1031	98.5	98
26	1.3532	0.3758	1370	43	6	0.6	1037	99.0	99
27	1.4951	0.3778	1386	44	2	0.2	1039	99.2	99
28	1.6389	0.3808	1403	44	2	0.2	1041	99.4	99
29	1.7857	0.3860	1420	45	1	0.1	1042	99.5	99
30	1.9378	0.3949	1438	46	3	0.3	1045	99.8	99
31	2.0993	0.4097	1456	47	0	0.0	1045	99.8	99
32	2.2761	0.4331	1477	50	0	0.0	1045	99.8	99
33	2.4787	0.4695	1500	54	1	0.1	1046	99.9	99
34	2.7250	0.5268	1529	61	1	0.1	1047	100.0	99
35	3.0500	0.6202	1566	72	0	0.0	1047	100.0	100
36	3.5344	0.7855	1622	91	0	0.0	1047	100.0	100
37	4.4122	1.1261	1724	130	0	0.0	1047	100.0	100
38	5.8476	1.9280	1890	223	0	0.0	1047	100.0	100