

The Connecticut Mastery Test: Technical Report

**Prepared by
Irene Hendrawan & Arianto Wibowo**

November 2009



Table of Contents

Table of Contents	2
List of Charts	3
List of Figures	4
List of Tables	5
Part 1: Introduction	6
1.1. General Description of CMT.....	6
1.2. 2009 CMT Test Design.....	7
1.3. 2009 CMT Test Forms.....	9
Part 2: Test Development	10
2.1. Content Standards.....	10
2.2. Item Development.....	10
2.3. Forms Construction.....	11
Part 3: Validity	12
3.1. Content Validity Survey.....	12
3.2. Scoring Quality Assurance Procedures Undertaken during Development.....	13
3.3. Item Quality Analysis Undertaken during Development.....	13
Part 4: CMT4 Achievement Standards	15
4.1. Standards for CMT3.....	15
4.2. Establishment of Standards for CMT4.....	15
4.3. Establishment of Science Standards for CMT4.....	16
4.4. Levels of Achievement and Cut Scores.....	18
Part 5: Scaling and Equating	20
5.1. Calibration Process.....	20
Part 6: Item and Test Statistics	23
6.1. Reliability.....	23
6.2. Classification Consistency and Accuracy.....	24
Part 7: Vertical Scale Score Development for CMT4	26
7.1. Overview.....	26
7.2. Data Collection and Design.....	27
7.3. Methodology.....	28
7.4. Results.....	29
References	36
Appendix A: Rasch Values for Editing and Revising Form S	37
Appendix B: Item Analysis	39
Appendix C: Raw Score, Theta, and Scale Score	51
Appendix D: 2007 Vertical Scaling Design	57
Appendix E: 2007 Vertical Scaling Item Parameters	60

List of Charts

Chart 1: Calibration Design for 2009 CMT Operational Writing	20
---	----

List of Figures

Figure 1: Theta Distributions for Mathematics across Grades.....	31
Figure 2: Mean of Theta Values for Mathematics across Grades.....	32
Figure 3: Theta Distributions for Reading across Grades.....	33
Figure 4: Mean of Theta Values for Reading across Grades.....	33
Figure 5: Relationship of Mathematics Cut Scores for Each Proficiency Level across Grades.....	35
Figure 6: Relationship of Reading Cut Scores for Each Proficiency Level across Grades.....	35

List of Tables

Table 1: 2009 CMT Operational Test Design	8
Table 2: Summary of Round 1	17
Table 3: Summary of Round 2	17
Table 4: Summary of Round 3	18
Table 5: Percentages of Students at Each Level, Using Round 3 Cut Scores	18
Table 6: CMT4 Achievement Levels and Scale Score Ranges	18
Table 7: 2008 CMT Equating Constants	20
Table 8: Summary of Weighting for Reading and Writing	21
Table 9: Scaling Coefficients	21
Table 10: Summary of Item Analysis Writing Form S'	23
Table 11: 2009 CMT Writing Cronbach's Alpha	24
Table 12: 2009 CMT Scale Score Summary Statistics	24
Table 13: Classification Consistency	24
Table 14: Classification Accuracy	25
Table 15: False Negative Classification	25
Table 16: False Positive Classification	25
Table 17: An Example of Scale Values, Cut Scores, and Performance Levels	26
Table 18: An Example of Scale Score Growth Expectations at Proficient	26
Table 19: Number and Item Types for Mathematics across Grades	27
Table 20: Number and Item Types for DRP across Grades	27
Table 21: Number and Item Types for RC across Grades	27
Table 22: Common Item and Student Design	28
Table 23: Number of Items Removed based on P-value Reversal	30
Table 24: Vertical Scale Cut Scores in Mathematics and Reading at Each Proficiency Level for Grades 3-8 for CMT4	34

Part 1: Introduction

1.1. General Description of CMT

Connecticut General Statutes Section 10-14n mandates a statewide mastery test to be administered annually in March to all public school students enrolled in grades 3-8. In accordance with that mandate, the Connecticut Mastery Test (CMT) was designed to measure student performance in the areas of Mathematics, Reading, Writing, and Science. The assessment focuses on content that students at each grade level can reasonably be expected to have mastered. Although the legislation specifically prohibits the use of test results as the sole criterion for promotion or graduation, the CMT provides information about achievement that is used for many purposes. Some purposes of the CMT are to:

- set high expectations and standards for student achievement;
- test a comprehensive range of academic skills;
- disseminate useful test achievement information about students, schools, and districts;
- identify students in need of intervention;
- assess equitable educational opportunities; and
- continually monitor student progress in grades 3-8 over time.

The CMT has measured growth in achievement for Connecticut students since 1985, when it was first administered. A second generation of the CMT was introduced in 1993 and a third generation in fall 2000. A fourth generation, which is the version currently in use statewide, was introduced in March 2006. New generations of the test offer an opportunity to adjust content, re-establish standards, and reflect changes in philosophy and technology that have occurred since the previous generation was developed.

The CMT is one important measure for determining student achievement in all of Connecticut's public elementary and middle schools. In 1994, the Connecticut Academic Performance Test (CAPT) was instituted for all Connecticut 10th-grade students as the logical extension of the CMT in the high school. Together, the CMT and CAPT provide a comprehensive system of monitoring and reporting on the academic progress of Connecticut students.

All Connecticut public school students are required to participate in the CMT except for a small number of students with very limited English proficiency that may be exempted from the test. The CMT results provide the opportunity to publicly account for statewide student achievement on the skills and knowledge that Connecticut considers to be important.

The content of the CMT was selected to represent the most important Mathematics, Reading, Writing, and Science skills for students at each of the grades tested. The test content reflects the standards of Connecticut's Curriculum Frameworks. This document, combined with the CMT, aids educators throughout Connecticut in designing instructional programs across all grades to bring about continued improvement in student achievement.

The interpretation of CMT results does not depend on comparing students against one another in terms of performance. Instead, the best way to understand CMT scores is to compare student performance against the established achievement standards. While scores are reported for each of the five tests on the CMT, achievement standards have been established in the four broad areas of Mathematics, Reading, Writing, and Science. In 2000, three achievement standards were established by the Connecticut Board of Education (CSBE), creating four levels of achievement. In 2002, a fourth standard was added to the previous three by CSBE, creating five levels of performance: Advanced, Goal, Proficient, Basic, and Below Basic. The top two levels (Advanced and Goal) define the Goal Range, which is the same as what has historically been referred to as "at or above goal."

There are other ways in which student results are presented. The sections of the test differ in breadth and complexity for each grade and content area. For this reason, student performance is reported in various ways for

each section, most frequently in relation to content strand mastery standards. This information will be explained in more detail in later sections.

The CMT requires more from students than most traditional tests in the areas of mathematics, reading, writing, and science. While traditional assessments typically measure what students know, the CMT also employs performance tasks to measure what students can do with what they know. For example, instead of just doing mathematical calculations, students are asked to apply calculation skills to solve everyday problems. In writing, students are asked to demonstrate their communication skills by producing an essay on a grade-appropriate topic.

The CMT is intended to support high-quality classroom instruction by providing useful feedback to teachers. By administering the CMT in grades 3-8, school districts can gain a comprehensive picture of student achievement. This information can be used for such purposes as individual student diagnosis and placement, curriculum alignment, instructional programs, and communication with parents about student progress.

The CMT plays an important role in education at the school and district levels. CMT results are reported for each school, each school district, and the state as a whole. They are available to the press and to the public on the website www.cmtreports.com.

An aligned assessment program reinforces educational priorities established by Connecticut educators. The CMT provides important feedback to schools and school districts as they work to improve the effectiveness of their educational programs. Many initiatives are in place to support the use of CMT results and to guide instruction toward greater effectiveness.

1.2. 2009 CMT Test Design

The content of the 2009 CMT was selected to represent the most important Mathematics, Reading, Writing, and Science skills for students at each of the grades tested. The test content reflects the standards of Connecticut's Curriculum Frameworks. From Connecticut's Curriculum Frameworks, assessment standards were developed for the CMT.

The spring 2009 administration was the fourth operational (OP) administration of CMT4. Each administration comprises the following content areas:

1. Mathematics
Mathematics (MA) consists of a single test administered in two sessions for grades 3 and 4 and three sessions for grades 5 through 8. The tests contain dichotomously scored multiple-choice (MC) items, grid-in (GR) response items, and open-ended (OE) items scored on a 0-1, 0-2, or 0-3 scale.
2. Reading
Reading (RD) consists of two subtests:
 - 2.1. Degrees of Reading Power
Degrees of Reading Power (DRP) has a single session of MC items.
 - 2.2. Reading Comprehension
Reading Comprehension (RC) consists of MC items and OE items scored on a 0-2 scale. RC has two sessions.
3. Writing
Writing (WR) consists of two subtests:
 - 3.1. Editing & Revising
Editing & Revising (ER) has only MC items and one session.
 - 3.2. Direct Assessment of Writing
Direct Assessment of Writing (DAW) has a single prompt test scored on a 2-12 scale.

4. Science

Science (SC), which is administered in grades 5 and 8 only, consists of MC items and OE items scored on 0-2 scale.

The 2009 CMT Operational forms for Mathematics, Reading, and Science are the same forms used in the 2008 CMT Operational (Forms S and Q'). For Writing, form S' consists of Editing & Revising (ER) items from 2008 Operational form S and a new DAW prompt (EX) scored on a 2-12 scale. Form Q' from 2008 Operational will again be used in 2009 Operational test.

Table 1: 2009 CMT Operational Test Design

Content	Subject	Grade	Number of Items				Total Items	Score Points	
			MC	GR	OE	ER			
Mathematics	Mathematics	3	76		18		94	0 – 106	
		4	80		16		96	0 – 110	
		5	80	13	20		113	0 – 132	
		6	71	18	27		116	0 – 140	
		7	70	19	31		120	0 – 146	
		8	61	20	36		117	0 – 146	
Reading	Degree of Reading Power	3	42				42	0 – 42	
		4	42				42	0 – 42	
		5	49				49	0 – 49	
		6	49				49	0 – 49	
		7	49				49	0 – 49	
		8	49				49	0 – 49	
	Reading Comprehension	3	22		9		31	0 – 40	
		4	24		8		32	0 – 40	
		5	22		9		31	0 – 40	
		6	22		9		31	0 – 40	
		7	20		10		30	0 – 40	
		8	20		10		30	0 – 40	
	Writing	Editing & Revising	3	32				32	0 – 32
			4	32				32	0 – 32
5			36				36	0 – 36	
6			36				36	0 – 36	
7			40				40	0 – 40	
8			40				40	0 – 40	
Direct Assessment of Writing		3				1	1	2 - 12	
		4				1	1	2 - 12	
		5				1	1	2 - 12	
		6				1	1	2 - 12	
		7				1	1	2 - 12	
		8				1	1	2 - 12	
Science		Science	5	36		3		39	0 - 42
			8	45		3		48	0 - 51

1.3. 2009 CMT Test Forms

The design of CMT forms reflects two critical goals of: 1) maintaining the horizontal link (year to year) from 2008 to the 2009 tests and 2) piloting new items for future years of CMT testing. Both of these goals are accomplished while maintaining the same high standards of CMT testing from previous years. The 2009 Operational forms have links to the 2008 Operational forms. Form Q' is used as the breach form in 2009 CMT.

Part 2: Test Development

The process by which each form of the CMT is developed is extensive, spanning a two- or three-year period and going through many stages. The development process is led and overseen by staff members in the Bureau of Student Assessment at the Connecticut State Department of Education (CSDE), but it also involves many other people who represent a wide variety of perspectives and areas of expertise. CSDE curriculum specialists and content experts play a critical role and work closely with the assessment staff throughout the process. In addition, a major testing company and other organizations and individuals with experience in educational assessment are involved at appropriate points in the development process.

Advisory committees of Connecticut educators are particularly important throughout the development of the CMT. Advisory committees are composed of Connecticut educators with respected knowledge in particular content areas. A separate advisory committee is established for each part of the CMT: Mathematics, Reading, Writing and Science. Additionally, a Fairness Committee screens all test material to ensure that all groups of examinees are validly assessed. Educators are carefully selected for the advisory committees to be representative of school districts throughout Connecticut.

2.1. Content Standards

The first and most critical stage of test development is the basic conceptual design of the test, determining what the most important content to assess is and how that content can best be assessed given the present resources and constraints. These decisions have important implications for the direction of education in Connecticut and for the manner in which the progress of students, schools, and school districts will be measured for several years. These basic decisions are based on the collective expertise of both assessment specialists and curriculum specialists at CSDE, along with input from the CMT advisory committees. Current educational research in the content areas, current assessment research, and current policies and priorities for education in Connecticut form the bases for these decisions. For example, the content tested on the CMT is directly aligned with the content outlined in *The Connecticut Framework: K-12 Curricular Goals and Standards*.

Once content is determined, other issues must be decided. Test formats (i.e., the types of questions used) must be selected. Also, the methods of scoring the questions and performance tasks must be established. These factors are directly related to the skills and knowledge being assessed. There is, therefore, great variation between and within CMT tests, each uniquely designed to assess specific abilities.

When decisions have been made about test content and test format, they are referred to as “test specifications.” Test specifications serve as the rules for developing the actual test questions. Clear test specifications ensure that test material is not only consistent with the priorities of Connecticut educators, but also that test forms are comparable from year to year. Hundreds of Connecticut citizens and educators responded to surveys that identified the content intended to be included on each test form, validating the appropriateness of the material for students at each grade.

2.2 Item Development

Test items for the CMT4 were carefully developed in accordance with the established test specifications and test blueprint for each grade to reflect content standards in the Connecticut Curriculum Frameworks for mathematics, reading/language arts, and science. After test items were developed according to the test specifications, they underwent extensive review by the testing company, CMT content advisory committees, and the fairness committee before being piloted with Connecticut students in grades 3 through 8. The content advisory committees included content experts, regular and special education teachers, Connecticut State Department of Education curriculum and assessment content specialists, who are knowledgeable about grade appropriate educational content and processes. For the CMT4, the fairness committee was responsible for determining whether items were appropriate and fair to all examinees. Items that did not pass the scrutiny of the either committee were eliminated from the pool of pilot items.

After committee reviews, field test forms were created and piloted on a representative sample, stratified by scale score distribution, of approximately 2000 students per form. During pilot testing, representative samples of students in grades 3 through 8 try out new test questions for the purpose of identifying potential problems with the questions. Questions that are being piloted do not count toward a student's score. The utility of the potential test questions is evaluated based on the results of the pilot testing. Estimated pilot statistics such as the mean, point biserial, and Rasch difficulty, misinterpretation or confusion on the part of the test takers, and performance of various demographic groups are reviewed by CSDE assessment content staff and psychometricians. A judgment is made as to whether each test question enabled students to demonstrate the required skills and knowledge. In addition, for constructed response items that require hand-scoring, the contractor provides qualitative summaries about whether students appeared to have sufficient contextual knowledge to be able to fully respond to the item. Based on these pilot results, flawed items were removed from the item pool, including those showing test item bias or inappropriate levels of difficulty, some were revised for re-piloting, and some became candidates for inclusion on a future form of the CMT.

2.3 Forms Construction

With test specifications as a guide, test forms are carefully constructed, taking into consideration the difficulty of the items and the balance of content. Because a new form of the CMT is developed and administered every few years, it is critical that the forms are "parallel," that is, as similar as possible in terms of both content coverage and test difficulty. This parallelism allows meaningful comparisons to be made from one test form to another. Any slight differences in difficulty among test forms that remain are accounted for through the equating process.

In Connecticut, we think in terms of "generations" of our testing program to allow predictable points where the testing process can be reevaluated and revised as necessary. A "generation" of a Connecticut test spans about five to seven years. During those years, every effort is made to create test forms, score student work, and interpret results in the same way from year to year. The first generation of CMT began in 1985, the second generation began in 1993, and the third generation began in fall 2000. The current, fourth generation CMT began in March 2006. Each new generation of the CMT involves a process similar to the one described above.

Based on the CMT4 blueprints, all test forms of equivalent difficulty per grade were then simultaneously constructed from the grade level pool of items that met all the review criteria, using eMetric's proprietary software, TestBuilder. Every effort was made to ensure that strand level difficulties were comparable and that the items reflected the range of content within the strands across the generation.

Part 3: Validity

According to the 1999 AERA, APA, NCME *Standards*, “It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) the assembly and evaluation of the test for operational use.

In the development and maintenance of CMT each of these phases is carefully planned and implemented. The following section details the critical psychometric procedures undertaken to ensure a strong validity argument for the use and interpretation of CMT (Kane, 2006; Messick, 1989).

3.1. Content Validity Survey

To examine the validity of the CMT for its intended applications, a number of studies have been conducted. The first focused on establishing content validity of each part of the CMT. In October 1984 (the year before the first administration of the grade 4 CMT), a survey of the objectives proposed for the grade 4 CMT was sent to more than 3,000 Connecticut educators. The purpose of the survey was to determine (1) the importance of the proposed mathematics and reading/writing objectives and (2) whether the objectives were taught prior to the fall administration of grade 4. Similar surveys of objectives proposed for grades 6 and 8 were sent to more than 8,000 Connecticut educators in October 1985.

For the third generation, another survey was developed and distributed in January 2000 for the same purpose. The respondents characterized the objectives as important educational outcomes to which students would be instructed prior to being tested. In addition to the test objective validation process, a two-step validation process was carried out. First, content experts reviewed all objectives and test items, examining the relationship between each item and its associated objective. Second, content experts judged how well each item and objective measured the purported content domain.

With the development of CMT4, CSDE commissioned Assessment and Evaluation Concepts, Inc. (AEC) to undertake a comprehensive survey of the Language Arts and Mathematics items to determine the match between item content and respective content strands, as well as the categorical concurrence between the test items and the broader content standards. In their summary report, AEC concluded that CSDE “has done a solid, quality job in matching the test items included on the CMT4 with the relevant content strands and standards of the Language Arts and Mathematics Curriculum Framework.” Such evidence, provided by an external reviewer, enhances the validity argument that the CMT4 content is relevant and representative of the constructs being measured.

When establishing validity for a newly developed test, it is common to correlate the examinee scores of the new test with the scores of other tests intended to measure similar content. The two tests need not be parallel or interchangeable, nor do they need to be used for the same purpose. Accordingly, the seventh edition of the Metropolitan Achievement Test (MAT7) was correlated with the CMT in 1993. In 2000, the Metropolitan Achievement Test, eighth edition (MAT8) was used during the first administration of the third generation CMT. Data from each of the four sections of the MAT (Total Language, Reading Comprehension, Math Concepts and Math Procedures) were used to compute the correlations among CMT tests and MAT sections. These correlations provided additional evidence to establish concurrent validity of the CMT.

The Direct Assessment of Writing portion of the CMT was additionally analyzed in another way. This was done because the Direct Assessment of Writing is a single, extended-response measure and, therefore, considerably different from the rest of the CMT tests. Validity concerns in this measure include the relation of the writing sample with the other language arts scores. Correlations between the Direct Assessment of Writing test and the other Language Arts tests (i.e., Degrees of Reading Power, Reading Comprehension, and Editing & Revising) were calculated to establish evidence of construct and concurrent validity.

3.2. Scoring Quality Assurance Procedures Undertaken during Development

Much of the following discussion applies to procedures undertaken during field testing and test construction phases of development work. Of course quality control is applied during the operational administration, but not with the aim of selecting or removing items.

In order to ensure the validity of inferences made from the CMT tests is to make certain there are quality control procedures in place for the scoring of the test. One such quality assurance component is to check the MC answer keys for MC items several times prior to test administration and one final time during the first run of live results. Items yielding low point-biserial correlations are checked a final time for miskeying.

For constructed-response (CR) items, CMT staff and contractor staff work with Connecticut educators to establish score boundaries in a process known as “range finding”. The score point examples and training sets so established are carried forward into operational scoring and elaborated with new samples of student responses. Reader training lasts up to several days, and readers must qualify by matching scores to several sets of prescored student responses. Once scoring begins, validity packets are used to maintain reader accuracy. These are packets of student responses with scores pre-assigned by CMT staff and Connecticut educators. Readers periodically receive these packets, and their responses are compared to the pre-assigned scores. If a reader assigns too many discrepant scores, that reader is retrained or removed from the project. Other QA procedures include a 100% second read for the writing prompts (DAW). There is a 20% second read for short answer and extended response items in mathematics and reading comprehension.

3.3. Item Quality Analysis Undertaken During Development

Another part of assessing the quality and validity of inferences made from an instrument is to assess the quality of the items on the test. This quality is typically assessed by examining the classical item statistics as well as the potential for item bias. Item bias could lead to less valid inferences made for certain subgroups.

Item specifications. CMT employs *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of the tests. The introduction to the 1999 *Standards* best describes how those *Standards* are and will be used in the development and evaluation of CMT tests:

Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4)

Thus, the terms ‘target’ and ‘goal’ are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to scrap a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

Item statistics. Because the CMT tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Target reliability coefficients of .90 (or higher) are therefore set for the important cut points of each test.

Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below:

For Multiple-Choice (MC) Items

Percent correct: greater than or equal to .25
Point biserial correlation with total score: greater than or equal to .20
Mantel-Haenszel: No Category C items (see below)

For Constructed-Response (CR) Items

Difficulty: any level as long as all score points are well represented

Correlation with total score: greater than or equal to .20

Generalized Mantel-Haenszel: No chi-square significant at .05 level of alpha

It should be pointed out that the point biserial correlations for MC items and the correlations for CR items refer to total scores of the field test form with the influence of the item in question removed.

Differential item functioning. The Mantel_Haenszel statistic computes an odds ratio for each item that compares item performance for a reference group and a focal group (for whom bias may be an issue). Specifically, the M-H statistic is a ratio of the probability of success on an item for the reference group to the probability of success on the same item for the focal group. When the ratio is greater than one, the probability of success on the item favors the reference group over the focus group. Note that M-H and other methods for identifying statistical bias are flagging mechanisms that do not necessarily mean that the performance difference is due to unfairness in the item. Instead, the standard procedure is for the bias committee review the items to make a final judgmental determination as to whether or not the item is actually biased.

Since its introduction in the field of epidemiology in 1959, Mantel-Haenszel statistics have been employed by many test developers, and several refinements have been added. Educational Testing Service (ETS) uses the Mantel-Haenszel statistic and calculates a D statistic which permits grouping of test items into three categories (Zieky, 1993). The D statistic is a function of the case-control odds estimator of risk generated by SAS's PROC FREQ. The D statistic is calculated as follows:

1. α = case-control estimate of risk (odds ratio)
2. β = natural log of α
3. $D = -2.35 * \beta$

Camilli and Shepard (1994, p. 121) describe three categories of items with respect to D:

- A D does not significantly differ from zero using Mantel-Haenszel chi-square, or D's absolute value is less than 1
- B D significantly differs from 0 and D has either (a) an absolute value less than 1.5 or (b) an absolute value not significantly different from 1
- C D's absolute value is significantly greater than or equal to 1.5

Camilli and Shepard note that Category B items are typically investigated for potential bias, while Category C items are typically removed. Others treat Category C items only as candidates for elimination, pending a reprieve from the committee. In other words, Category C items are considered unusable unless specifically declared usable by the committee. It should be noted that an item that allowed a target group to break out of a pattern of trailing behind the reference group on all other items would tend to fall into Category C. The committee would likely want to keep such an item, in spite of its Mantel-Haenszel status.

DIF occurs when an item shows different results by group (e.g., by race or sex) that cannot be explained by known differences in the overall achievement levels of the two groups. Overall achievement level is typically taken as score on an operational test, assuming that the operational test is itself free of bias. While committee members are free to examine all field-tested items, they must review all items with a Category C rating. Unless the committee specifically calls for the inclusion of any such item, that item is removed from the pool.

Part 4: CMT4 Achievement Standards

To continue to comply with the *No Child Left Behind* (NCLB) accountability requirements, the Connecticut State Department of Education (CSDE) carried over from the third generation Connecticut Mastery Test (CMT3) to the fourth generation (CMT4) the previously adopted achievement standards: Below Basic, Basic, Proficient, Goal and Advanced. The CMT3 was last administered in fall 2004 to students in Grades 4, 6 and 8 in mathematics, reading and writing. The CMT4 was first administered in Grades 3 through 8 in spring 2006 in the same three content areas.

The purpose of this section is to summarize the procedures used to accomplish the task of carrying over CMT3 standards to CMT4 and to recommend for approval the CMT4 achievement standards for each grade and content area. The recommendations take into consideration the results from a statistical intergenerational equating study, historical results from past CMT3 administrations, and input from our CMT Standards Review Panel composed of a diverse group of Connecticut educators. All procedures were discussed with and approved by our Technical Advisory Committee (TAC) prior to implementation. The TAC is composed of nationally recognized experts in the measurement field.

4.1. Standards for CMT3

In June 2002, the State Board of Education approved revisions to the standards for the CMT3 in Grades 4, 6 and 8. Standards were established based on scale scores (100-400) in three areas: mathematics, reading and writing. In all content areas, the standards define the different academic performance levels, denoted as Below Basic, Basic, Proficient, Goal and Advanced. The state goal has been an important benchmark for judging the quality of education in Connecticut for more than a decade. The proficient standard is used for accountability purposes as required by NCLB to make determinations about Adequate Yearly Progress (AYP) and schools in need of improvement.

4.2. Establishment of Standards for CMT4

When standards were being established for first and second generation CMT, a judgmental standard setting process called Modified Angoff, was employed. Through that process, groups of educators who were familiar with the performance of students at a particular grade level in a particular content area were asked to predict how students who just meet a particular standard (e.g., remedial standard) would perform on many different CMT items. Using the judgment of these groups of educators in consideration with other validity checks, appropriate state goal and remedial standards were recommended by the Department and adopted by the State Board of Education.

The third generation standards were developed through department staff working with a CMT3 Standards Advisory Panel composed of technical experts, district content experts and district research and testing specialists. The CMT3 standards were set to be as rigorous as the CMT2 standards and to be equivalent across grade levels and across content areas as much as possible.

The process of carrying over CMT3 standards to the CMT4 was based on an intergeneration linking study, consideration of historical results from the CMT3, and judgmental input from the CMT Standards Review Panel. The purpose of the linking study was to equate standards from Grades 3, 5 and 7 of CMT4 with Grades 4, 6 and 8 of CMT3 in order to maintain the same performance standards for NCLB purposes. The equating not only adjusted for differences in difficulty between CMT3 and CMT4, but also for differences due to the change in the testing window. The CMT4 standards for Grades 4, 6 and 8 were then derived through interpolation and extrapolation procedures by examining the previously established trends in standards across Grades 3, 5 and 7.

The Standards Review Panel assisted in the identification of acceptable and valid test standards for each content area of CMT4. Committee membership was broadly constituted to be representative of the state and to include a variety of stakeholders. The CMT Standards Review Panel was given an overview of the CMT3 including the content covered, score weighting, and reporting conventions. Differences between CMT3 and CMT4 were also discussed. Copies of the complete CMT4 were available for reference. In addition, the procedures for carrying

CMT3 standards over to CMT4 were presented in detail so that committee members would better understand their role in the process. They reviewed data from several related analyses and discussed implications from both an educational perspective and a technical perspective. They were asked particularly to provide input in the following three areas:

- Review results from the intergenerational linking procedure to ensure that standards are reasonable and appropriate across grades and content areas,
- Provide subjective input about the effect of changing testing from fall to spring and losing instructional time in March through June for CMT4 examinees, so that the CMT3 standards are maintained across the two generations of testing, and
- Provide subjective input about the reasonableness and consistency of the standards for all grades and content areas.

The full standard-setting report contains the projected percentages of students who will score at or above the CMT4 standards along with the comparative data from the 2004 CMT3 administration. Based on the best data that were available at the time the standards panel was convened, we were able to estimate the scale score cut points that correspond to the projected percentages. The scale score cut scores for 2008 CMT along with those used with CMT3 are presented in Tables 3 (below). The same scale score cut points established for the 2006 administration will be applied in all future CMT4 administrations.

4.3. Establishments of Science Standards for CMT4

On May 20-22, 2008, the Connecticut State Department of Education (CSDE) conducted standard setting for the Science Test component of the Connecticut Mastery Test (CMT). CSDE staff had invited 25 science educators from around the state to participate in this activity and to recommend cut scores for the tests for grades 5 and 8. Measurement Incorporated (MI), the contractor for CMT, served as facilitator for the session employing the bookmark procedure (Cizek & Bunch, 2007).

With the Bookmark procedure, panelists examine test items in an ordered-item booklet and determine whether or not a minimally Basic, Proficient, Goal, or minimally Advanced student would have a 2/3 chance of answering the item correctly (for MC items) or obtain the given score point (for CR items). The ordered-item booklet consists of the items from the actual test but arranged in order of difficulty, with the easiest item on the first page and the most difficult item on the last page. MC and SCR items appear only once in the booklet, but ECR items and writing prompts appear once for each score point. An item worth three points would appear three times, the first time with a sample response representing one point, later with a sample response representing two points, and so on.

Each page contains essential information about the item, including its position in the ordered booklet, its position in the original booklet, and the achievement level (theta) required for a student to have a 2/3 chance of answering correctly or obtaining that point. These theta values are derived from analysis of the student responses to the items through the use of item response theory (IRT) procedures. Specifically, for CMT Science, MI uses the Rasch model for item calibration and test construction. This model allows for the calibration of all items and students on a common scale. This common calibration allows for the calculation of a probability of a correct response to a given item by a given student from information about the student's achievement level (theta) and the items difficulty level (delta).

Panelists enter four bookmarks on a special form, one each for the last page they believe a threshold student would have a 2/3 chance of answering correctly. The page number is associated with a theta required for a 2/3 chance of answering correctly. These theta values are averaged across all panelists. The mean theta is then translated into a score via a table from the Rasch analysis of the live test results. The tabled raw score closest to this value becomes the cut score

After panelists entered their bookmarks (page numbers) in Round 1, MI research assistants (one for each test) entered them into spreadsheets set up to convert the page numbers to theta values for the operational test and

calculate mean theta for each cut, as well as standard deviation and range. The facilitators shared these results with the panelists as described above at the beginning of Round 2. Table 2 summarizes the results of Round 1.

Table 2: Summary of Round 1

	Grade 5	Grade 8
Range of bookmarks for Basic	3-5	4-6
Range of bookmarks for Proficient	7-10	10-21
Range of bookmarks for Goal	17-28	23-39
Range of bookmarks for Advanced	34-37	42-48
Mean theta cut for Basic (S.D.)	-0.549 (0.151)	-0.294 (0.075)
Mean theta cut for Proficient (S.D.)	0.046 (0.068)	0.143 (0.109)
Mean theta cut for Goal (S.D.)	0.923 (0.104)	0.737 (0.265)
Mean theta cut for Advanced (S.D.)	1.801 (0.131)	1.945 (0.181)
Mean cut score for Basic	15	22
Mean cut score for Proficient	20	27
Mean cut score for Goal	28	33
Mean cut score for Advanced	34	43

Panelists received the feedback from Round 1 and discussed the range of bookmarks for each proficiency level, as well as the range of cut scores. Panelists then evaluated the items in the ordered item booklets. On this occasion, they skipped over the items at the beginning of the booklet that had not been delineated by Basic bookmarks by anyone in Round 1 and stopped on the last page delineated by an advanced bookmark in Round 1. Otherwise, procedures in Round 2 were exactly as those in Round 1. Table 3 shows the results of Round 2.

Table 3: Summary of Round 2

	Grade 5	Grade 8
Range of bookmarks for Basic	3-4	4-5
Range of bookmarks for Proficient	7-10	11-17
Range of bookmarks for Goal	20-28	24-32
Range of bookmarks for Advanced	35-38	42-48
Mean theta cut for Basic (S.D.)	-0.786 (0.209)	-0.349 (0.082)
Mean theta cut for Proficient (S.D.)	-0.006 (0.048)	0.117 (0.031)
Mean theta cut for Goal (S.D.)	0.922 (0.064)	0.667 (0.137)
Mean theta cut for Advanced (S.D.)	1.855 (0.185)	2.043 (0.256)
Mean cut score for Basic	13	21
Mean cut score for Proficient	20	26
Mean cut score for Goal	28	32
Mean cut score for Advanced	34	43

Panelists returned for Round 3 and received the information in Table 3 as well as a graphic showing the percentages of students classified at each level, based on their Round 2 cut scores. Drs. Bunch and Deville led discussions of the range of cut scores as well as the impact data and answered panelists' questions. After this discussion, panelists completed their Readiness Forms, indicating that they were ready to go on to Round 3.

In Round 3, panelists entered not only the page numbers where they would place their bookmarks, but their cut scores and percentages of students scoring at or above that cut. Results are summarized in Table 4. Impact data are summarized in Table 5.

Table 4: Summary of Round 3

	Grade 5	Grade 8
Range of bookmarks for Basic	4-5	4-5
Range of bookmarks for Proficient	9-21	10-19
Range of bookmarks for Goal	26-28	24-34
Range of bookmarks for Advanced	37-39	41-50
Mean theta cut for Basic (S.D.)	-0.505 (0.058)	-0.360 (0.079)
Mean theta cut for Proficient (S.D.)	0.188 (0.275)	0.106 (0.075)
Mean theta cut for Goal (S.D.)	0.977 (0.030)	0.667 (0.142)
Mean theta cut for Advanced (S.D.)	2.137 (0.111)	2.058 (0.331)
Mean cut score for Basic	16	21
Mean cut score for Proficient	21	26
Mean cut score for Goal	28	32
Mean cut score for Advanced	36	43

Table 5: Percentages of Students at Each Level, Using Round 3 Cut Scores

	Grade 5*	Grade 8
Below Basic	7.4	14.6
Basic	11.0	10.0
Proficient	26.4	16.6
Goal	40.6	43.8
Advanced	14.7	14.9

After Round 3, MI staff shared results with panelists who then evaluated the process and outcomes. Before leaving the workshop, MI staff delivered the final cut scores and a summary of results to CSDE for final review and approval.

4.4. Levels of Achievement and Cut Scores

Table 6 shows the range of scale scores in each performance category.

Table 6: CMT4 Achievement Levels and Scale Score Ranges

Content Area	Grade	Below Basic	Basic	Proficient	Goal	Advanced
Mathematics	3	100 - 186	187 - 209	210 - 241	242 - 287	288 - 400
	4	100 - 193	194 - 214	215 - 244	245 - 289	290 - 400
	5	100 - 190	191 - 214	215 - 244	245 - 292	293 - 400
	6	100 - 189	190 - 213	214 - 243	244 - 284	285 - 400
	7	100 - 190	191 - 215	216 - 245	246 - 289	290 - 400
	8	100 - 190	191 - 213	214 - 244	245 - 286	287 - 400
Reading	3	100 - 201	202 - 216	217 - 234	235 - 278	279 - 400
	4	100 - 212	213 - 226	227 - 243	244 - 294	295 - 400
	5	100 - 202	203 - 214	215 - 229	230 - 278	279 - 400
	6	100 - 206	207 - 219	220 - 235	236 - 288	289 - 400
	7	100 - 193	194 - 207	208 - 221	222 - 272	273 - 400
	8	100 - 205	206 - 218	219 - 231	232 - 281	282 - 400

Content Area	Grade	Below Basic	Basic	Proficient	Goal	Advanced
Writing	3	100 - 187	188 - 211	212 - 239	240 - 286	287 - 400
	4	100 - 184	185 - 208	209 - 236	237 - 280	281 - 400
	5	100 - 185	186 - 208	209 - 237	238 - 283	284 - 400
	6	100 - 184	185 - 210	211 - 236	237 - 283	284 - 400
	7	100 - 191	192 - 212	213 - 235	236 - 269	270 - 400
	8	100 - 188	189 - 211	212 - 235	236 - 282	283 - 400
Science	5	100 - 187	188 - 212	213 - 247	248 - 299	300 - 400
	8	100 - 201	202 - 220	221 - 243	244 - 298	299 - 400

Part 5: Scaling and Equating

5.1 Calibration Process

The 2009 CMT test forms were scaled and equated using the Rasch model. The WINSTEPS software, written by Linacre (Mesa Press, 2005) was used to estimate the latent trait difficulty of each item on the test. WINSTEPS is a WINDOWS-based program that is widely used for similar high stakes tests. WINSTEPS (the Rasch model), allows for the estimation of item difficulty for multiple-choice, open-ended, and extended response items on a single scale. Using these item difficulties, the model is able to estimate the ability (theta) of each student corresponding to each student's raw score.

All scaling and equating analyses were undertaken by three independent groups: Measurement Incorporated (MI), the contractor, the Connecticut State Department of Education (CSDE), and H. Jane Rogers and H. Swaminathan from the University of Connecticut (UCONN). Results were compared and cross-checked to the fourth decimal point to ensure accuracy.

The 2009 CMT Operational forms for Mathematics, Reading, and Science are the same forms used in the 2008 CMT Operational (Forms S and Q'). CSDE has decided to use 2008 score tables (raw-to-scale score and raw-to-vertical scale score) for these subject areas. Please refer to the 2008 Technical Report.

For Writing, form S' consists of Editing & Revising (ER) items from 2008 Operational form S and a new DAW prompt (EX) scored on a 2-12 scale. New raw score to scale score tables for Writing were constructed.

The Writing equating was accomplished using a common item equating design. The purpose of the equating was to place the difficulty estimates of the Form S' items on the same scale as Form S (CMT 2008 Live). The Writing equating was accomplished in the following steps:

1. For Writing, calibrate the 2009 OP with Form S (see Chart 1 for sample calibration data matrix). This step is a free run calibration. For DAW two points are subtracted from each score so that scores are on a scale from 0 to 10.

Chart 1: Calibration Design for 2009 CMT Operational Writing

Grade 3 - 8 Form S'

S_ER	2009_DAW
------	----------

Note:

S_ER = Form S Editing & Revising

2009_DAW = new DAW prompt

2. Select common items between forms S and S'. Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and Form S' values (see Appendix A for the Rasch values of common items). This is an iterative process in which each item, starting with the one with the greatest absolute value difference, is removed until all items fulfill the criterion for use. Using the remaining items the difference between the scale means from Form S and Step 1 yields the equating constant. Table 7 shows the equating constants.

Table 7: 2008 CMT Equating Constants

Grade	Writing
3	-0.1687
4	-0.0394
5	-0.1006
6	-0.0323
7	-0.3397
8	-0.1403

3. Using the item output files from Step 1 and anchoring these b-values, perform another run for each combination of forms, i.e., employ only those items from a given form in order to obtain theta values for each group of students administered a particular form. For Writing, the appropriate weights were included (see Table 8).

Table 8: Summary of Weighting for Reading and Writing

Content/Subject	Grade	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
Editing & Revising	3	0 – 32	40%	1.00		0 – 32
	4	0 – 32	40%	1.00		0 – 32
	5	0 – 36	40%	1.00		0 – 36
	6	0 – 36	40%	1.00		0 – 36
	7	0 – 40	40%	1.00		0 – 40
	8	0 – 40	40%	1.00		0 – 40
Direct Assessment of Writing	3	2 – 12	60%	4.80	(DAW-2)*4.80	0 – 48
	4	2 – 12	60%	4.80	(DAW-2)*4.80	0 – 48
	5	2 – 12	60%	5.40	(DAW-2)*5.40	0 – 54
	6	2 – 12	60%	5.40	(DAW-2)*5.40	0 – 54
	7	2 – 12	60%	6.00	(DAW-2)*6.00	0 – 60
	8	2 – 12	60%	6.00	(DAW-2)*6.00	0 – 60
Total Writing	3	2 – 44				0 – 80
	4	2 – 44				0 – 80
	5	2 – 48				0 – 90
	6	2 – 48				0 – 90
	7	2 – 52				0 – 100
	8	2 – 52				0 – 100

4. Compute scale score (SS) and scale score standard error (SSE) for each forms

$$SS = \left(\frac{T + EQ - T_{mean}}{T_{SD}} \right) * 45 + 250 \text{ and } SSE = \frac{T_{err}}{T_{SD}} * 45$$

where

T and T_{err} are the ability score and the standard error of the ability from the score file in Step 3.

EQ is the difference between the mean of difficulties estimates of the linking items on Form S and mean of difficulties estimates of the common items on Form S', called the equating constant. This value was obtained in Step 2.

T_{mean} and T_{SD} are the scaling coefficients from CMT3 and 2006 CMT (see Table 9).

Table 9: Scaling Coefficients

Content	Grade	T_mean	T_SD
Mathematics	3	2.11972	1.06174
	4	1.715231	1.127543
	5	1.56796	1.08322
	6	1.206464	1.184245
	7	0.79254	1.18006
	8	0.742283	1.223549

Content	Grade	T_mean	T_SD
Reading	3	0.99235	1.24795
	4	1.149643	1.173126
	5	1.19747	1.20351
	6	1.107734	1.241252
	7	1.52062	1.19801
	8	1.379708	1.271917
Writing	3	0.97123	1.24615
	4	1.405899	1.303604
	5	1.06359	1.23642
	6	1.200022	1.203568
	7	1.21748	1.36516
	8	1.123911	1.2611

Note: Scaling coefficients for grade 3, 5, and 7 came from CMT3 and for grade 4, 6, and 8 came from 2006 CMT.

The minimum SS will be 100 and the maximum SS will be 400. SS less than 100 will be reported as 100 and SS greater than 400 will be reported as 400.

Appendix C contains the results of raw scores, theta, and scale score for Form S' Writing. Please contact CSDE for other subjects, forms and combinations.

Part 6: Item and Test Statistics

Table 10 and Appendix B present a summary and detailed of item analysis (item quality) data for grades 3-8 Writing Form S, respectively. The following information is presented in each item analysis:

Classical and IRT difficulties: Item difficulty is fundamentally a ratio of the proportion of examinees who answered the item correctly. Thus, an easy item has a high p-value and a difficult item has a low p-value. If an item has a very high p-value it may be so easy that it does not provide much information about what most examinees know or can do, while an item with a very low p-value may be so difficult that it is beyond the range of what most students know or can do. Therefore, items with very high or very low p-values may be rejected, unless content relevance overrides that concern.

Item Discriminations: The point biserial correlation or item-total correlations measure the strength of the relationship between the particular item score and the total score. Thus, item discrimination reflects how well a particular item differentiates between high and low total test performers. When the correlation is high, examinees that do well on the item also tend to do well on the entire test and correspondingly, examinees that do not do well on the item also tend not to do well on the total test.

Distractor Frequencies: The proportion of students who answered each option (A-E, 0-3, and 2-12) are presented for the multiple-choice items, open-ended and extended response, respectively. The percent of students at each score point is presented for extended response (2-12).

Table 10: Summary of Item Analysis Writing Form S'

Subject	Grade	Rasch		P-value		Point biserial	
		Mean	Std	Mean	Std	Mean	Std
Editing & Revising	3	-0.0231	0.8996	0.71	0.14	0.40	0.08
	4	-0.0193	0.7526	0.75	0.11	0.39	0.09
	5	-0.0178	0.7106	0.73	0.11	0.38	0.09
	6	-0.0232	0.7328	0.73	0.11	0.38	0.10
	7	-0.0180	0.8525	0.74	0.13	0.36	0.09
	8	-0.0118	0.9895	0.72	0.15	0.39	0.09
Direct Assessment of Writing	3	0.7400		8.02		0.53	
	4	0.6177		8.44		0.56	
	5	0.6421		8.10		0.50	
	6	0.8366		7.94		0.58	
	7	0.7200		8.18		0.60	
	8	0.4711		8.54		0.60	

6.1. Reliability

Reliability is a statistical index of the consistency of test performance over repeated trials. The simplest model for conveying the concept of reliability is to describe the test re-test method. If a test is administered to a group of examinees and then re-administered to the same examinees a short time later, the correlation of the scores across both test administrations estimates the reliability of the test. To measure reliability using a single administration, the test items are split using various techniques into half-length tests and those scores are then correlated. Cronbach's alpha estimates the lower-bound estimate of an infinite combination of split-halves and therefore is regarded as a very conservative method for assessing test reliability.

Table 11 summarizes reliability estimates for 2009 CMT Writing. The reliability coefficients are based on Cronbach’s alpha measure of internal consistency. When evaluating these results it is important to remember that reliability is partially a function of test length and thus reliability is likely to be greater for clusters that have more items. Table 12 presents the mean and standard deviation of students’ scale scores.

Table 11: 2009 CMT Writing Cronbach’s Alpha

Grade	Writing
3	0.887
4	0.881
5	0.884
6	0.881
7	0.884
8	0.895

Table 12: 2009 CMT Scale Score Summary Statistics

Grade	MA		Reading		Writing		Science	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
3	256.45	51.84	239.19	40.80	254.24	45.57		
4	262.46	49.09	254.78	43.31	253.26	46.06		
5	268.61	50.41	243.61	41.06	255.13	43.89	254.27	46.11
6	264.20	44.76	258.42	45.80	252.87	46.67		
7	264.77	45.79	250.57	45.18	245.25	38.97		
8	260.03	43.76	251.26	40.23	254.12	45.59	252.58	45.88

6.2. Classification Consistency and Accuracy

Classification Consistency and Accuracy studies were measured using the IRT-Class program (see Lee, Hanson, and Brennan, 2002), developed by [CASMA](#) (Center for Advanced Studies in Measurement and Assessment) at the University of Iowa. The classification consistency and accuracy can be assessed based on the given ability distribution and the difficulty of the items (IRT parameters). Tables 13-16 contain the results of these analyses.

The results of decision consistency and accuracy computations show that for the most part, decisions are highly consistent (see Table 13). The consistency ratings at each cut score are generally in the upper 90s. This tends to tail off at the highest cut score (i.e., the upper end of the distributions). The cumulative effect of applying all cut scores simultaneously yields an average consistency of around mid 90s. The classification accuracy estimates show (see Table 14), similarly, that the accuracy ratings at each cut score are generally in the upper 90s.

The program also computes the false negative rates for the test, which in effect are an estimate of those students that may have been misclassified to a performance category lower than their true performance category. The results of the false negatives, found in Table 15, indicate that a very small number of students may have been negatively misclassified in this way. Table 16 shows the false positive classification.

Table 13: Classification Consistency

Content Area	Grade	Overall Classification Consistency	Cut 1	Cut 2	Cut 3	Cut 4
Writing	3	0.96537	0.96608	0.96610	0.96610	0.96611

Content Area	Grade	Overall Classification Consistency	Cut 1	Cut 2	Cut 3	Cut 4
	4	0.97128	0.97133	0.97133	0.97133	0.97137
	5	0.96801	0.96810	0.96807	0.96807	0.96807
	6	0.95932	0.96199	0.96315	0.96315	0.96315
	7	0.97538	0.97539	0.97539	0.97539	0.97539
	8	0.97330	0.97341	0.97341	0.97341	0.97345

Table 14: Classification Accuracy

Content Area	Grade	Overall Classification Accuracy	Cut 1	Cut 2	Cut 3	Cut 4
Writing	3	0.94510	0.96185	0.97377	0.97588	0.97481
	4	0.95098	0.96786	0.97939	0.97950	0.97715
	5	0.95011	0.96790	0.97478	0.97724	0.97438
	6	0.94073	0.95622	0.96970	0.97306	0.97432
	7	0.96675	0.97845	0.98315	0.98315	0.98106
	8	0.94817	0.97396	0.98100	0.98100	0.97042

Table 15: False Negative Classification

Content	Grade	Overall False Negative	Cut 1	Cut 2	Cut 3	Cut 4
Writing	3	0.03676	0.03629	0.02070	0.01364	0.00718
	4	0.03068	0.03067	0.01491	0.01435	0.00456
	5	0.02983	0.02977	0.02037	0.01514	0.00557
	6	0.04335	0.04144	0.02536	0.01918	0.01055
	7	0.01900	0.01899	0.01015	0.01015	0.00469
	8	0.02418	0.02416	0.01365	0.01365	0.00200

Table 16: False Positive Classification

Content	Grade	Overall False Positive	Cut 1	Cut 2	Cut 3	Cut 4
Writing	3	0.01814	0.00185	0.00553	0.01048	0.01801
	4	0.01833	0.00147	0.00570	0.00615	0.01829
	5	0.02006	0.00232	0.00485	0.00762	0.02005
	6	0.01593	0.00234	0.00494	0.00776	0.01513
	7	0.01425	0.00255	0.00670	0.00670	0.01425
	8	0.02765	0.00187	0.00536	0.00536	0.02758

Part 7: Vertical Scale Score Development for CMT4

7.1. Overview

Vertical scaling is used to place test scores from assessments that vary in difficulty, but measure similar constructs, on the same scale. For example, students in grades 3-8 who take their state’s reading achievement assessments, whereby each grade level has its own test can be provided vertically scaled scores so that a given student’s achievement can be compared to students’ scores from the same grade as well as across the grades. In addition, a vertical scale allows one to track a student’s growth, e.g., in reading from year to year. Vertically scaled scores can also be aggregated, so that one could also track scores at the grade, school, or district level.

This type of scale can also be used to track student growth, relate the content and skills in items across grades, and examine the relationship of performance standards from grade to grade (see hypothetical values in Tables 17 and 18). Such a scale might also afford the state of Connecticut an additional method for reporting student achievement for purposes of No Child Left Behind, or simply as another approach to investigating and interpreting test scores for purposes of tracking growth and development.

The hypothetical numbers in Table 17 illustrate growth in two directions. First within a grade, e.g., grade 3, the raw and scale scores needed to attain Basic, Proficient, and Advanced proceed from 48 to 65 to 80 (raw) and 330 to 500 to 654 (scale). Looking across grades within a level, e.g., at the Proficient level, a grade 3 student must obtain a scale score of 500, while a grade 4 students needs a score of 559, etc., up to grade 8 where a student must score 700. (Raw scores are not relevant when examining growth across grades within a proficiency level.)

Table 18 illustrates, again using hypothetical numbers, the level of growth or the amount of score change needed when moving from grade to grade. As just described, at the Proficiency level, a score change of 59 points would be required. Likewise, a 45-point score change between grades 4 and 5 is needed to maintain a performance level of Proficient.

In summary, a vertical scale can be a useful tool to examine the growth of individual students or aggregates of students (e.g., schools). The scale can provide information regarding students’ progress across grades as well as within a grade across proficiency levels.

Table 17: An Example of Scale Values, Cut Scores, and Performance Levels

Grade	Basic		Proficient		Advanced	
	Raw	Scale	Raw	Scale	Raw	Scale
3	48	330	65	500	80	654
4	42	354	64	559	80	748
5	39	382	62	604	81	799
6	44	417	69	641	83	823
7	43	426	65	673	80	867
8	47	507	64	700	81	914

Table 18: An Example of Scale Score Growth Expectations at Proficient

Grade Progression	Gain
3 to 4	59 points
4 to 5	45 points
5 to 6	37 points
6 to 7	32 points
7 to 8	27 points
3 to 8	200 points

In Spring 2007, the Connecticut State Department of Education (CSDE) decided to investigate the possibility of using vertical scales in its statewide testing program. This part provides information with respect to the vertical scaling analyses undertaken by the state’s contractor, Measurement Incorporated (MI).

7.2. Data Collection and Design

Data were collected as part of the regular testing administration in Spring 2007. Test scores from the regular, operational administration (Form P’) were used, as well as scores from shorter, supplemental exams. Items from the operational tests were used to construct all supplemental exams. Tables 19-21 provide the numbers and types of items from the Form P’ operational tests across grades 3-8. The Math tests were comprised of multiple-choice (MC), grid-in (GR), and open-ended (OE) questions. The Reading test is a combination of two separate parts, the Degree of Reading Power (DRP) and the Reading Comprehension (RC) test.

Table 19: Number and Item Types for Mathematics across Grades

Grade	Number of Items			Total Items
	MC	GR	OE	
3	76		18	94
4	80		16	96
5	80	13	20	113
6	71	18	27	116
7	70	19	31	120
8	61	20	36	117

Table 20: Number and Item Types for DRP across Grades

Grade	Number of MC Items	Total Items
3	42	42 of 73
4	42	42 of 74
5	49	49 of 80
6	49	49 of 80
7	49	49 of 79
8	49	49 of 79

Table 21: Number and Item Types for RC across Grades

Grade	Number of Items		Total Items
	MC	OE	
3	22	9	31 of 73
4	24	8	32 of 74
5	22	9	31 of 80
6	22	9	31 of 80
7	20	10	30 of 79
8	20	10	30 of 79

During the 2007 CMT administration, students in grades 3-8 were given a supplemental exam in addition to the regular, operational assessments. The supplemental exams were constructed so that the students could be tested ‘off grade’, meaning that, for example, grade 5 students were administered a supplemental test that contained either grade 4 or grade 6 operational items. The supplemental tests were shorter than the operational exams (students took only one section within the supplemental content area), but enough supplemental forms were created and administered to include all operational items. So for a given grade-level operational test, all items

were also administered to students in the adjacent grades via the supplemental exams. The design called for the administration of each grade-level item to approximately 1,500 students from each adjacent grade. This common item and student design permits vertical linking of performance across grades (see Table 22). The diagonal (boldface) fields represent the on-level items at a given grade level, while the off-diagonal fields represent the off-grade administration of the operational items to adjacent grades (the upper diagonal are the supplemental exams administered to adjacent lower grades, while the lower are the tests given to the adjacent higher grades).

Table 22: Common Item and Student Design

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3	OP33	SU34				
	Grade 4	SU43	OP44	SU45			
	Grade 5		SU54	OP55	SU56		
	Grade 6			SU65	OP66	SU67	
	Grade 7				SU76	OP77	SU78
	Grade 8					SU87	OP88
	Grade 8						

Notation: OP=Operational test; SU=Supplemental test; Numerals=grade-level students and test level taken, e.g., SU56 refers to fifth-grade students who took the supplemental exam containing grade-six operational items.

7.3. Methodology

Only students who were administered Form P' (main form) during the 2007 Spring administration were included in the analyses. Equating analyses for the 2007 operational forms for the six grades and three content areas were performed and cross-validated (see 2007 CMT Technical Report).

Before beginning with the linking, we first examined the classical difficulties (p-values) from the on-grade data and values for the same items from the next higher grade. If an item had a p-value for the on-grade students that was 5%*maximum score or greater than that obtained from students in the next higher grade, we removed it from subsequent analyses. (Experience has shown that p-values from on-grade students are almost always higher than those obtained from students in the next lower grade, e.g., grade 4 students administered grade 4 test materials will, in general, perform better on every item than 3rd-grades taking those same 4th-grade items.) Because students in higher grades may have forgotten material learned the previous year, some items are likely to exhibit this 'reverse' pattern of difficulty. The item can work well when measuring on-grade performance, but may not be suitable for modeling a vertically linked continuum of learning. For this reason, we removed such items.

The linking plan follows the scheme represented in Appendix D. As explained below, there are two linking paths to follow, meaning we obtained two sets of item parameters for each grade level. The strength of this design is that we were able to determine how well the two links result in convergent values for the rescaled parameters.

Having obtained the two sets of parameters from following the upper and lower linking paths we then examined the item parameters to determine how similar they were by obtaining a correlation coefficient between the sets of parameters and used Fisher's z-test to determine if the differences were significant. An item that exhibited

very different parameter values was to be removed from further analysis. We then used the mean of the two parameter values for each item to proceed. The following steps detail the analyses.

Step 1 (see Appendix D). Based on advice from the Technical Advisory Committee, we set a middle grade as the base scale, namely grade 5. We first did a free run on the OP55 items and obtained item and person parameters.

There are common items linking OP55 and SU45. By anchoring the grade 5 items we obtained ‘new’ (i.e., different from what these grade 4 students would have from having taken OP44) thetas for the SU45 test takers. We then fixed these new theta values. Using the PAFILE command in WINSTEPS, i.e., anchoring students’ thetas, we then obtained item parameter estimates by linking to OP44.

Again starting from OP55, we had theta values for the grade-5 students from their on-grade testing and the initial free run. Linking to SU54 via the common students, we anchor their grade-5 theta values and obtained parameter estimates for the grade-4 items.

At this point we have two sets of parameters for the grade-4 items, each set linked to the grade-5 scale. We compared the two sets using a Pearson correlation and the Fisher Z-test. We expected $r > 0.90$. Fisher’s Z is calculated by:

$Z = .5 \log \left(\frac{1+(r-0.9)}{1-(r-0.9)} \right)$, where

$Z \sim N \left(0, 1/\sqrt{n-3} \right)$, and where n is the number of observations.

Our plan was to remove ‘outliers’ until $Z < 1.96 / \sqrt{n-3}$, then calculate the average of the remaining item parameters. These estimates were then be used to obtain thetas for the OP44 students.

Step 2. At this point we had the OP44 item parameters and thetas, linked to the grade-5 scale, and proceeded as in the first step. There are common OP44 items linked to SU34. By fixing those item values, we obtained thetas for the test takers in SU34. We then anchored their theta values and linked to OP33, obtaining a set of item parameters for the grade-3 items.

Similar to the grades 5 and 4 connection, there are common students between OP44 and SU43. We anchored the theta values in OP44 and linked to SU43, giving us a second set of item parameter estimates for grade 3.

We then went through the same procedure described above to determine if the two linking paths and procedures gave us similar results. Finally, we used the average item parameter estimates to obtain thetas for the grade 3 students.

Step 3. The same procedures were used to link the higher grades. Again, we started with the free run of OP55, using those item parameter and theta estimates as the starting point. Common items link OP55 to SU65. Fixing the grade 5 item parameters, we obtained theta estimates for SU65. By fixing these theta values, we linked to OP66 to obtain grade-6 item parameters.

The link from OP55 to SU56 is the common students. We fixed the students’ theta values from their on-grade testing, i.e., OP55, and obtained item parameter estimates for the grade-6 items.

The items were examined to identify problematic ones, which were to be discarded. For the remaining items we calculated the mean of the two parameters and used that to get thetas for the grade-6 students.

Steps 4 and 5. The same procedures were used as just delineated for grades 7 and 8. When finished we had items and students on the same Rasch scale using grade 5 as the base.

Using the final item parameter and theta estimates a vertical, developmental scale was created to demonstrate what growth would look like across the grades in Math and Reading. It is emphasized here that the choice of a scale was somewhat arbitrary and was undertaken without consultation with CSDE or the TAC. The scale is for illustrative purpose only.

7.4. Results

Table 23 presents the number of items that were removed because of item p-value reversals, i.e., where the p-value for the item taken by the on-grade students was 5%*maximum score or higher than the p-value for the students at the higher adjacent grade. Noteworthy is that few items were removed, especially at the lower grades 3 and 4. More Math items were removed than Reading items. With respect to Reading, no items were

removed until grade 6, where 10 of 80 had to be discarded for further analysis. Grade 6 also saw the most Math items removed. The TAC and CSDE discussed why so many items from this particular grade level showed reversals and whether the content of the items might play a role. In addition, having removed a larger number of items, especially in Reading, likely affected the subsequent vertical scaling, although to what extent would be very difficult to determine. While removing these ‘misfitting’ items likely results in better vertical scales, further analysis, interpretation, and justification is needed to improve our understanding of how this procedure affects vertical scaling.

Table 23: Number of Items Removed based on P-value Reversal

Grade	# Items Removed	
	Mathematics	Reading
3		
4	2 / 96	
5	4 / 113	
6	8 / 116	10 / 80
7	4 / 120	3 / 79

The WINSTEPS runs were performed in the manner described above in Section 7.2.2. The two linking paths were followed linking grade to grade. The resulting two sets of item parameters were compared using Pearson’s correlation and the Fisher Z-test. No items were removed based on these analyses. The TAC suggested that the method of comparing the parameters may not have been stringent enough. An investigation into what other procedures might be more appropriate would be a worthwhile research project.

The final Rasch item parameters, using grade 5 as the base scale, can be found in Appendix E. Figure 1 is output from SAS that shows the distributions of thetas across grades in Mathematics based on the vertical scaling using the obtained Rasch values. The mean thetas increase across grades, from a mean of 0.3021 for grade 3 to a high of 2.9339 for grade 8. The variability in the distributions is quite similar, with standard deviations between 1.2 and 1.3. The range of the thetas across the six grades is approximately 10 logits, from -3.5 to 6.0.

Figure 1: Theta Distributions for Mathematics across Grades

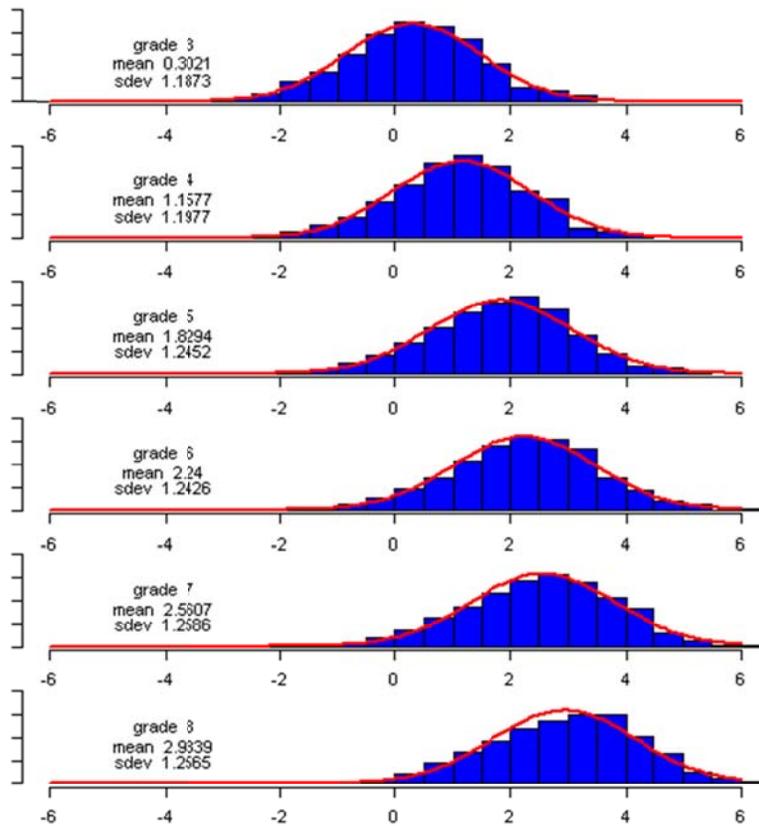
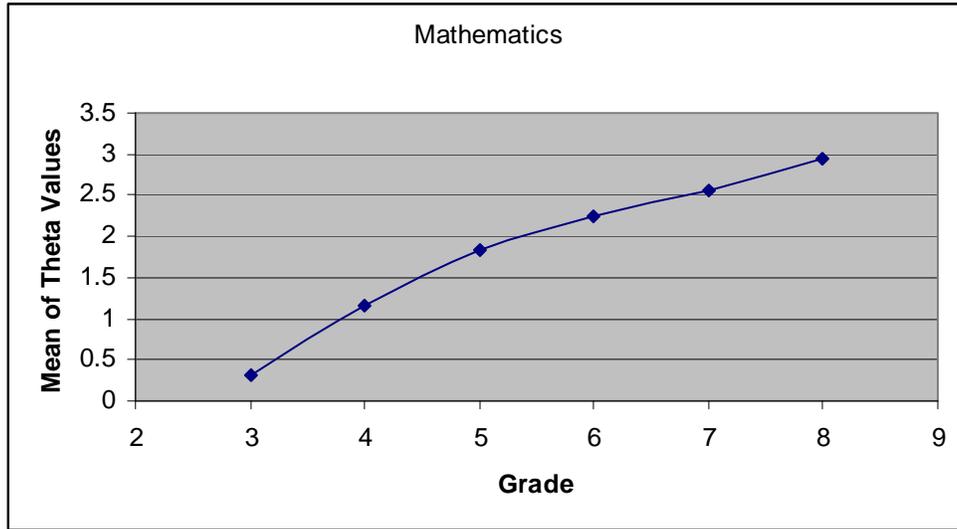


Figure 2 provides a graphic depiction of the increasing mean theta values in Mathematics across grades 3-8. Growth, as depicted here, appears to be steeper at the lower grades and becomes somewhat flatter in the upper grades. In other words, growth appears to slow as the students get older.

Figure 2: Mean of Theta Values for Mathematics across Grades



The pattern of thetas for Reading was similar to the pattern for the Mathematics (Figures 3-4). Again, the mean thetas increase across grades, from a mean of -0.0991 for grade 3 to a high of 2.0393 for grade 8 in Reading. The standard deviations were between 1.1 and 1.3 in Reading. Similar to Mathematics, the range of the thetas for Reading across the six grades is approximately 10 logits, from -3.5 to 6.0. Figure 4 show the increasing mean theta values in Reading across grades 3-8 respectively. Once again, growth appears to be steeper at the lower grades and becomes somewhat flatter in the upper grades, i.e., growth looks to slow as the students get older.

For illustrative purposes we constructed a vertical scale score in order to demonstrate what growth would look like across such a scale, and just as importantly, what the relationship would be across the grades when examining the performance levels. MI did not consult with CSDE or the TAC to generate this scale, although the results appear to be very promising.

At the outset of constructing the scale, we discovered that we could not use the score files given by WINSTEPS because some items were not included (a number of items had been removed due to p-value reversals). So we anchored the thetas and recalibrated all items. Having done that, we then recalibrated the thetas using all items with parameters obtained above.

The scale range chosen was 100-800. These somewhat arbitrary values come from simply doubling the present score scale used for all CMT tests (i.e., 100-400). At this point we have a theta for each student in grades 3-8. The student's vertical scale score (VS) is equal to:

$$VS = 100 + 700 * ((\text{theta} - \min(\text{theta})) / (\max(\text{theta}) - \min(\text{theta})))$$

Figure 3: Theta Distributions for Reading across Grades

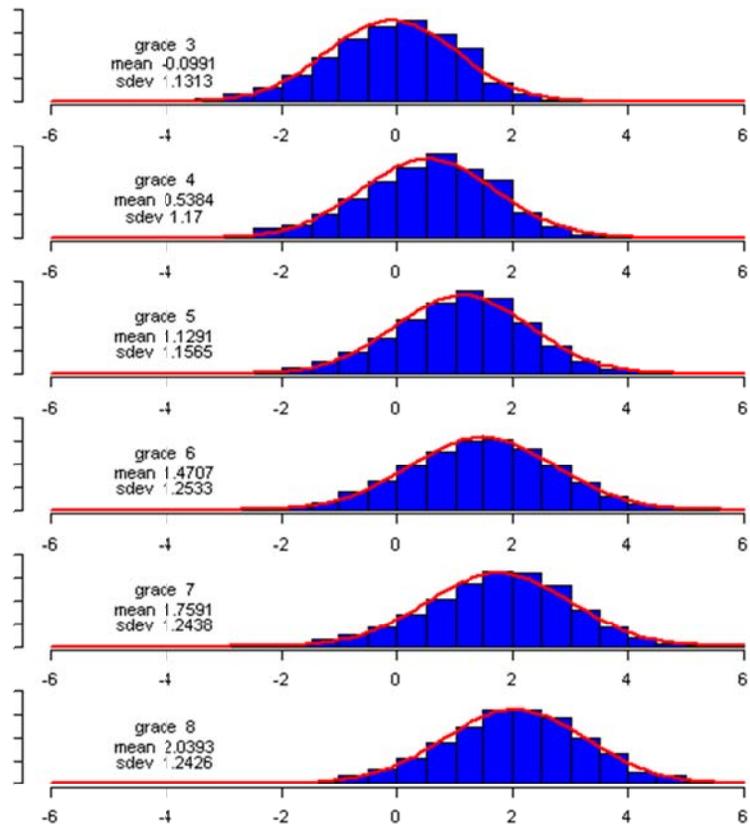
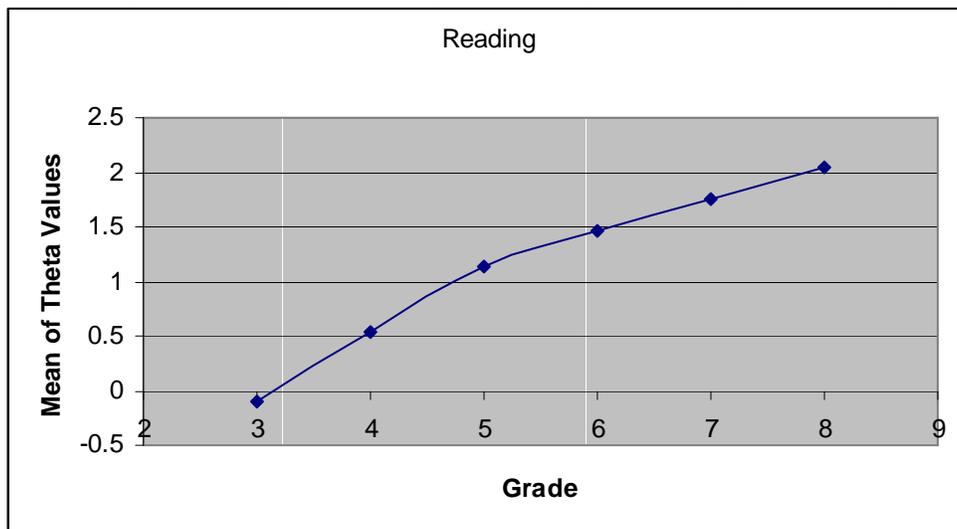


Figure 4: Mean of Theta Values for Reading across Grades



The vertical scores of the above formula have resulted in inconsistent vertical scale scores changes at the lower and upper end of the scale, which may lead to misinterpretation. University of Connecticut (H. Swaminathan and H. Jane Rogers) suggested and have implemented the necessary adjustments.

To obtain the vertically scaled cut scores, we started with the Mathematics and Reading scale score cut points for the different performance levels used operationally in CMT4 (where the scale is 100-400), and then found the corresponding raw cut scores. We then looked at the score file out of the final vertical scaling run in WINSTEPS to obtain the theta value that corresponds to the raw score. Finally, the theta values were inserted into the scale score formula above to obtain a student's VS on the vertical scale of 100-800 (Table 24).

Figures 5-6 depict the relationship between the vertically scaled cut scores across the proficiency levels for Mathematics and Reading respectively. Growth increases across the grades as do the cut scores. From the graphs it is clear that the cut scores for Advanced, especially in Reading, set this group well apart from the others.

Some degree of caution is advisable when interpreting the extent of growth, the speed of growth, and the extent of differences across grades. A vertical scale is most helpful when looking at such information across years and not simply for a single year, as presented in this report. That said, it appears these initial results indicate that a vertical scale may add another, and important, dimension for Connecticut's educators to interpret test scores.

Based on vertical scaling in CMT 2007, CSDE has decided to use the available results to generate the conversion tables for the whole generation of CMT4. In order to generate conversion tables in subsequent years, conversion tables mapping the conventional scale score to the vertical scale score will be used as lookup tables to determine the appropriate vertical scale score for a given conventional scale score.

Table 24: Vertical Scale Cut Scores in Mathematics and Reading at Each Proficiency Level for Grades 3-8 for CMT4

Content Area	Grade	Basic	Proficient	Goal	Advanced
Mathematics	3	396	418	452	499
	4	429	453	483	531
	5	450	476	506	558
	6	466	492	526	572
	7	483	509	543	593
	8	496	523	559	608
Reading	3	382	400	425	481
	4	410	427	447	507
	5	436	449	467	525
	6	439	455	475	545
	7	453	472	489	550
	8	466	483	500	564

Figure 5: Relationship of Mathematics Cut Scores for Each Proficiency Level across Grades

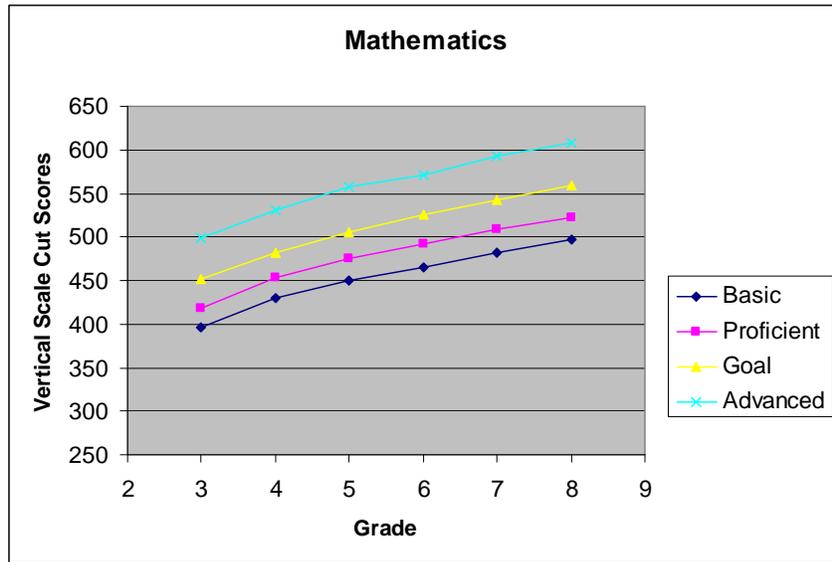
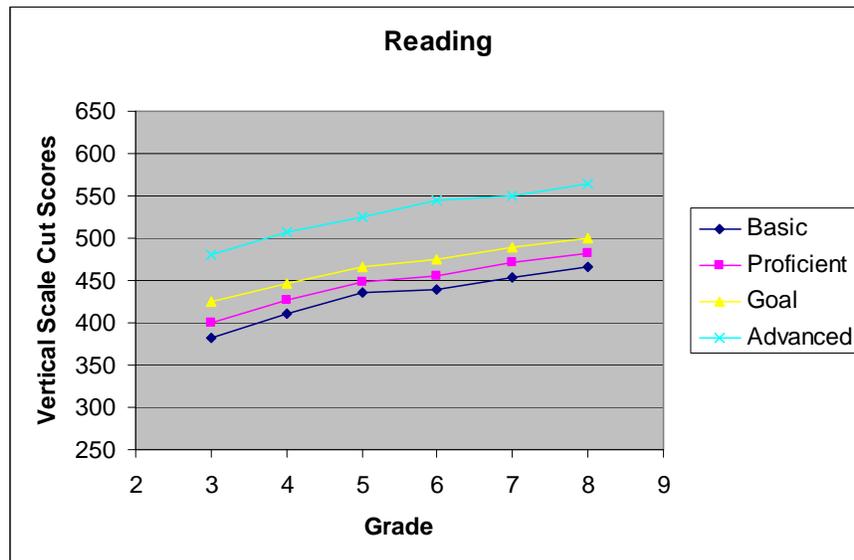


Figure 6: Relationship of Reading Cut Scores for Each Proficiency Level across Grades



REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600), Washington, DC: American Council on Research.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.
- Lee, W-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412-432.
- 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.
- The Connecticut Framework: K-12 Curricular Goals and Standards. (1998). Connecticut State Department of Education.
- Winsteps. (1991-2006©). Linacre, John M.
- Zikey, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix A: Rasch Values for Editing and Revising Form S

Item	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
1	-0.5483	0.5554	1.3955	-0.8389	-0.9427	-0.6274
2	1.3217	-0.7236	0.3708	-0.2709	-0.7008	-1.7331
3	1.0441	0.7501	-0.5437	0.8413	-1.3928	1.0168
4	-0.3751	0.3203	-0.6583	0.2304	-0.1316	-1.2188
5	-1.3891	-1.0708	-0.3718	-1.2587	-1.0609	-1.9518
6	-1.2191	0.2449	0.3089	-0.4959	-2.4253	0.5916
7	-1.3894	0.3563	0.3151	-0.7942	-1.0403	-1.0068
8	0.5923	-0.2287	-0.8446	-0.4377	-0.4851	0.6943
9	-1.9338	-0.4346	-0.3881	1.4383	-0.7500	-1.7976
10	1.2905	0.5741	-0.0021	-1.6949	0.6468	-1.2091
11	-0.0211	0.1737	-1.0424	0.6294	0.3614	1.4723
12	1.0377	-0.1568	0.0877	-0.2775	0.0260	-1.4563
13	-0.1903	1.0924	0.7279	0.5179	-0.4298	-0.0844
14	-0.6519	-0.6267	-0.4289	0.8928	-0.2998	-0.6961
15	-0.3919	-1.1124	0.5811	-0.0861	0.2042	-0.0895
16	-0.1168	-0.6959	-0.9024	-0.9866	-0.4000	-1.1860
17	-0.6789	-1.1990	-0.9589	-0.3262	-1.3713	0.1002
18	-1.0954	-0.5509	0.6409	0.9202	-0.9273	0.8733
19	0.8770	0.7982	-0.5251	0.3532	-0.8605	-0.6875
20	-1.5242	-1.2098	0.8358	0.2468	-0.3080	0.9212
21	-0.7092	-0.6089	-0.5370	-0.4156	1.3791	-1.1932
22	-0.7911	-0.2981	0.7290	-1.0665	0.4220	-0.6445
23	0.4841	-0.9456	0.1007	-0.0176	-0.6279	-0.7570
24	-0.7743	0.8974	-0.1705	0.1868	-0.3537	-0.3713
25	-0.6738	-0.3566	-0.3633	0.3739	0.4200	-0.5708
26	-0.0191	0.3493	-0.3471	0.0922	-0.7859	-0.2981
27	0.0875	-0.1733	-1.0004	0.1440	0.5235	-0.9355
28	1.4458	0.3727	-0.5325	0.2902	0.2117	-0.1732
29	-0.5478	0.2889	-0.5323	0.5880	0.6285	1.5619
30	0.6301	-0.5230	-0.3112	-0.5766	-0.0788	1.4368
31	-0.3282	0.0867	1.6791	-0.6262	0.9179	-0.1492
32	0.4206	2.1772	0.8417	-0.4137	-1.2151	-0.3999
33			-0.3448	-0.2453	2.1390	0.4334
34			-1.4400	-0.3415	-0.6726	1.0166
35			-0.6175	-0.3031	-0.7412	-0.1115

Item	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
36			-0.0140	1.7273	-0.2841	1.1042
37					-1.1573	0.2609
38					-0.4184	-0.0191
39					-1.3585	0.1852
40					-0.9696	1.6180

Appendix B: Item Analysis

Editing and Revising Form S Grade 3 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	-0.4579	0.79	0.32
2	MC	1.5003	0.45	0.35
3	MC	1.1093	0.53	0.26
4	MC	-0.2107	0.76	0.39
5	MC	-1.2203	0.88	0.33
6	MC	-1.0377	0.86	0.54
7	MC	-1.2414	0.88	0.40
8	MC	0.7708	0.59	0.31
9	MC	-1.7856	0.92	0.39
10	MC	1.4902	0.45	0.22
11	MC	0.0961	0.71	0.51
12	MC	1.2288	0.50	0.37
13	MC	-0.0172	0.73	0.51
14	MC	-0.4583	0.79	0.46
15	MC	-0.2372	0.76	0.38
16	MC	0.1772	0.70	0.43
17	MC	-0.5179	0.80	0.44
18	MC	-0.9427	0.85	0.40
19	MC	1.0754	0.53	0.29
20	MC	-1.3346	0.89	0.41
21	MC	-0.5558	0.81	0.43
22	MC	-0.6034	0.81	0.58
23	MC	0.6186	0.62	0.38
24	MC	-0.5922	0.81	0.48
25	MC	-0.4908	0.80	0.47
26	MC	0.1915	0.70	0.37
27	MC	0.2696	0.68	0.52
28	MC	1.5470	0.44	0.34
29	MC	-0.3833	0.78	0.40
30	MC	0.8386	0.58	0.50

Item	Type	Rasch	PC	RPB
31	MC	-0.1408	0.75	0.37
32	MC	0.5745	0.63	0.34

Direct Assessment of Writing Form S Grade 3 Item Analysis

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.7400	8.02	0.53	0.01	0.00	0.03	0.03	0.12	0.11	0.32	0.16	0.15	0.04	0.02

Editing and Revising Form S Grade 4 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	0.5727	0.67	0.21
2	MC	-0.7015	0.85	0.46
3	MC	0.8304	0.63	0.22
4	MC	0.3659	0.71	0.28
5	MC	-1.0566	0.89	0.41
6	MC	0.3404	0.71	0.42
7	MC	0.4111	0.70	0.41
8	MC	-0.2264	0.80	0.38
9	MC	-0.3560	0.81	0.37
10	MC	0.6130	0.67	0.48
11	MC	0.1683	0.74	0.40
12	MC	-0.1363	0.78	0.45
13	MC	1.1423	0.57	0.35
14	MC	-0.5367	0.83	0.40
15	MC	-1.0355	0.88	0.42
16	MC	-0.5535	0.84	0.37
17	MC	-1.1758	0.90	0.43
18	MC	-0.6157	0.84	0.49
19	MC	0.7700	0.64	0.30
20	MC	-1.1220	0.89	0.45
21	MC	-0.5778	0.84	0.18
22	MC	-0.2569	0.80	0.48
23	MC	-0.9014	0.87	0.39
24	MC	0.9720	0.60	0.30
25	MC	-0.3723	0.81	0.54
26	MC	0.3154	0.72	0.45
27	MC	-0.0370	0.77	0.40
28	MC	0.4527	0.69	0.49
29	MC	0.3443	0.71	0.45
30	MC	-0.4645	0.82	0.45
31	MC	0.1141	0.75	0.47
32	MC	2.0954	0.39	0.33

Direct Assessment of Writing Form S Grade 4 Item Analysis

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.6177	8.44	0.56	0.00	0.00	0.01	0.01	0.06	0.10	0.37	0.20	0.15	0.06	0.03

Editing and Revising Form S Grade 5 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	1.4859	0.47	0.22
2	MC	0.4541	0.66	0.45
3	MC	-0.4275	0.80	0.44
4	MC	-0.5648	0.82	0.40
5	MC	-0.2697	0.78	0.40
6	MC	0.4235	0.67	0.28
7	MC	0.4185	0.67	0.26
8	MC	-0.7286	0.84	0.44
9	MC	-0.3779	0.80	0.24
10	MC	0.0349	0.73	0.39
11	MC	-0.9211	0.86	0.35
12	MC	0.2567	0.70	0.41
13	MC	0.8210	0.60	0.31
14	MC	-0.3656	0.79	0.32
15	MC	0.6182	0.63	0.38
16	MC	-0.7329	0.84	0.43
17	MC	-0.8590	0.85	0.35
18	MC	0.7781	0.60	0.39
19	MC	-0.4559	0.81	0.38
20	MC	0.9258	0.58	0.38
21	MC	-0.4369	0.80	0.50
22	MC	0.8289	0.59	0.18
23	MC	0.2102	0.71	0.27
24	MC	-0.0008	0.74	0.44
25	MC	-0.2637	0.78	0.40
26	MC	-0.3022	0.79	0.47
27	MC	-0.9003	0.86	0.51
28	MC	-0.4193	0.80	0.56
29	MC	-0.4813	0.81	0.32
30	MC	-0.1640	0.77	0.50
31	MC	1.7811	0.41	0.32
32	MC	0.9580	0.57	0.31

Item	Type	Rasch	PC	RPB
33	MC	-0.2228	0.77	0.47
34	MC	-1.3715	0.90	0.38
35	MC	-0.5010	0.81	0.48
36	MC	0.1296	0.72	0.49

Direct Assessment of Writing Form S Grade 5 Item Analysis

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.6421	8.10	0.50	0.00	0.00	0.02	0.02	0.10	0.12	0.43	0.16	0.11	0.04	0.02

Editing and Revising Form S Grade 6 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	-0.7981	0.84	0.31
2	MC	-0.2477	0.77	0.41
3	MC	0.9246	0.57	0.31
4	MC	0.2274	0.70	0.22
5	MC	-1.2641	0.89	0.21
6	MC	-0.4039	0.80	0.37
7	MC	-0.7567	0.84	0.43
8	MC	-0.4467	0.80	0.41
9	MC	1.5336	0.45	0.33
10	MC	-1.6646	0.92	0.30
11	MC	0.6212	0.63	0.50
12	MC	-0.2275	0.77	0.29
13	MC	0.6275	0.63	0.31
14	MC	0.9216	0.57	0.40
15	MC	-0.0920	0.75	0.42
16	MC	-0.9262	0.86	0.46
17	MC	-0.3259	0.79	0.41
18	MC	0.9937	0.56	0.13
19	MC	0.3972	0.67	0.41
20	MC	0.3015	0.69	0.47
21	MC	-0.3802	0.79	0.42
22	MC	-0.9843	0.86	0.44
23	MC	0.0745	0.72	0.31
24	MC	0.1444	0.71	0.35
25	MC	0.4297	0.66	0.43
26	MC	0.0814	0.72	0.51
27	MC	0.0478	0.73	0.16
28	MC	0.3477	0.68	0.44
29	MC	0.6466	0.62	0.35
30	MC	-0.5234	0.81	0.54
31	MC	-0.6011	0.82	0.40
32	MC	-0.4070	0.80	0.45

Item	Type	Rasch	PC	RPB
33	MC	-0.2073	0.77	0.47
34	MC	-0.3237	0.79	0.35
35	MC	-0.2455	0.77	0.41
36	MC	1.6689	0.43	0.46

Direct Assessment of Writing Form S Grade 6 Item Analysis

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.8366	7.94	0.58	0.01	0.00	0.03	0.03	0.12	0.14	0.36	0.15	0.11	0.04	0.02

Editing and Revising Form S Grade 7 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	-0.6633	0.85	0.35
2	MC	-0.2886	0.80	0.39
3	MC	-1.0266	0.88	0.43
4	MC	0.2919	0.71	0.36
5	MC	-0.7907	0.86	0.38
6	MC	-2.0817	0.95	0.33
7	MC	-0.7120	0.85	0.45
8	MC	-0.0594	0.77	0.42
9	MC	-0.4573	0.82	0.33
10	MC	1.0043	0.58	0.28
11	MC	0.6714	0.64	0.27
12	MC	0.3487	0.70	0.30
13	MC	-0.0580	0.77	0.42
14	MC	0.1025	0.74	0.44
15	MC	0.5915	0.66	0.41
16	MC	-0.0036	0.76	0.40
17	MC	-1.0900	0.89	0.46
18	MC	-0.5333	0.83	0.37
19	MC	-0.5165	0.83	0.42
20	MC	0.0499	0.75	0.25
21	MC	1.6285	0.45	0.20
22	MC	0.6794	0.64	0.33
23	MC	-0.4035	0.82	0.35
24	MC	-0.0468	0.77	0.26
25	MC	0.7254	0.63	0.38
26	MC	-0.3487	0.81	0.33
27	MC	0.8811	0.60	0.34
28	MC	0.5366	0.67	0.20
29	MC	0.9418	0.59	0.23
30	MC	0.3518	0.70	0.25
31	MC	1.2867	0.52	0.23
32	MC	-0.9116	0.87	0.48

Item	Type	Rasch	PC	RPB
33	MC	2.5222	0.28	0.22
34	MC	-0.3873	0.81	0.48
35	MC	-0.3398	0.81	0.50
36	MC	-0.0252	0.76	0.41
37	MC	-0.8581	0.87	0.46
38	MC	-0.0444	0.77	0.50
39	MC	-1.0035	0.88	0.43
40	MC	-0.6840	0.85	0.54

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Direct Assessment of Writing Form S Grade 7 Item Analysis

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.7200	8.18	0.60	0.01	0.00	0.02	0.02	0.10	0.11	0.36	0.18	0.14	0.05	0.02

Editing and Revising Form S Grade 8 Item Analysis

Multiple-choice Items

PC = Proportion Correct

RPB = Point-biserial correlation for keyed answer

A – D = Proportion answering each distractor; answer key is shaded

Item	Type	Rasch	PC	RPB
1	MC	-0.4970	0.81	0.40
2	MC	-1.5922	0.92	0.41
3	MC	1.1907	0.53	0.27
4	MC	-1.0830	0.88	0.37
5	MC	-1.8210	0.93	0.37
6	MC	0.7254	0.62	0.33
7	MC	-0.8870	0.86	0.45
8	MC	0.8567	0.60	0.39
9	MC	-1.6210	0.92	0.43
10	MC	-1.1565	0.88	0.41
11	MC	1.7476	0.42	0.17
12	MC	-1.2498	0.89	0.39
13	MC	-0.0068	0.75	0.42
14	MC	-0.5675	0.82	0.37
15	MC	0.0515	0.74	0.47
16	MC	-1.0695	0.88	0.39
17	MC	0.2410	0.71	0.35
18	MC	0.9906	0.57	0.46
19	MC	-0.6214	0.83	0.48
20	MC	1.1069	0.55	0.10
21	MC	-1.0673	0.88	0.32
22	MC	-0.5400	0.82	0.47
23	MC	-0.6480	0.83	0.50
24	MC	-0.2650	0.78	0.39
25	MC	-0.4360	0.81	0.37
26	MC	-0.1277	0.76	0.47
27	MC	-0.8108	0.85	0.43
28	MC	-0.0565	0.75	0.40
29	MC	1.6720	0.44	0.49
30	MC	1.5825	0.46	0.39
31	MC	-0.0502	0.75	0.39
32	MC	-0.2031	0.78	0.29

Item	Type	Rasch	PC	RPB
33	MC	0.5374	0.66	0.53
34	MC	1.1834	0.53	0.43
35	MC	0.0794	0.73	0.33
36	MC	1.2340	0.52	0.42
37	MC	0.4835	0.67	0.23
38	MC	0.1272	0.73	0.48
39	MC	0.3559	0.69	0.25
40	MC	1.7406	0.43	0.39

Direct Assessment of Writing Form S Grade 8 Item Analysis

Extended Response

Mean = Mean EX score

Corr = Item-total correlation

2 – 12 = Percent of students at each score point

Item	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
1	EX	0.4711	8.54	0.60	0.01	0.00	0.01	0.01	0.07	0.08	0.34	0.20	0.18	0.07	0.03

Appendix C: Raw Score, Theta, and Scale Score

Writing Grade 3 Form S'

Raw Score	Theta	Scale Score
0	-5.3723	100
1	-4.2146	100
2	-3.5782	100
3	-3.2197	100
4	-2.9707	108
5	-2.7792	115
6	-2.6226	120
7	-2.4889	125
8	-2.3715	129
9	-2.2658	133
10	-2.1689	137
11	-2.0789	140
12	-1.9941	143
13	-1.9135	146
14	-1.8362	149
15	-1.7615	151
16	-1.6890	154
17	-1.6181	156
18	-1.5485	159
19	-1.4801	161
20	-1.4124	164
21	-1.3453	166
22	-1.2787	169
23	-1.2124	171
24	-1.1463	174
25	-1.0802	176
26	-1.0142	178
27	-0.9482	181

Raw Score	Theta	Scale Score
28	-0.8819	183
29	-0.8155	185
30	-0.7488	188
31	-0.6818	190
32	-0.6145	193
33	-0.5467	195
34	-0.4785	198
35	-0.4097	200
36	-0.3404	203
37	-0.2703	205
38	-0.1995	208
39	-0.1278	210
40	-0.0552	213
41	0.0186	216
42	0.0935	218
43	0.1698	221
44	0.2477	224
45	0.3272	227
46	0.4084	230
47	0.4917	233
48	0.5770	236
49	0.6645	239
50	0.7543	242
51	0.8465	245
52	0.9410	249
53	1.0377	252
54	1.1367	256
55	1.2379	260

Raw Score	Theta	Scale Score
56	1.3410	263
57	1.4462	267
58	1.5533	271
59	1.6624	275
60	1.7737	279
61	1.8875	283
62	2.0043	287
63	2.1246	292
64	2.2492	296
65	2.3791	301
66	2.5151	306
67	2.6584	311
68	2.8101	316
69	2.9711	322
70	3.1422	328
71	3.3241	335
72	3.5175	342
73	3.7236	349
74	3.9454	357
75	4.1887	366
76	4.4644	376
77	4.7950	388
78	5.2317	400
79	5.9415	400
80	7.1456	400

Writing Grade 4 Form S'

Raw Score	Theta	Scale Score
0	-5.1698	100
1	-4.0091	100
2	-3.3709	100
3	-3.0134	100
4	-2.7680	106
5	-2.5821	112
6	-2.4326	117
7	-2.3072	122
8	-2.1990	126
9	-2.1033	129
10	-2.0169	132
11	-1.9379	135
12	-1.8645	137
13	-1.7955	139
14	-1.7302	142
15	-1.6676	144
16	-1.6073	146
17	-1.5486	148
18	-1.4913	150
19	-1.4349	152
20	-1.3791	154
21	-1.3237	156
22	-1.2683	158
23	-1.2128	160
24	-1.1570	162
25	-1.1006	163
26	-1.0435	165
27	-0.9856	167

Raw Score	Theta	Scale Score
28	-0.9267	169
29	-0.8667	172
30	-0.8055	174
31	-0.7430	176
32	-0.6792	178
33	-0.6139	180
34	-0.5472	183
35	-0.4791	185
36	-0.4094	187
37	-0.3381	190
38	-0.2652	192
39	-0.1907	195
40	-0.1144	198
41	-0.0363	200
42	0.0437	203
43	0.1258	206
44	0.2100	209
45	0.2964	212
46	0.3854	215
47	0.4769	218
48	0.5710	221
49	0.6680	225
50	0.7679	228
51	0.8708	232
52	0.9766	235
53	1.0853	239
54	1.1968	243
55	1.3109	247

Raw Score	Theta	Scale Score
56	1.4273	251
57	1.5457	255
58	1.6657	259
59	1.7871	263
60	1.9096	267
61	2.0333	272
62	2.1580	276
63	2.2841	280
64	2.4120	285
65	2.5420	289
66	2.6750	294
67	2.8118	299
68	2.9535	303
69	3.1011	309
70	3.2561	314
71	3.4204	320
72	3.5960	326
73	3.7860	332
74	3.9949	339
75	4.2296	347
76	4.5022	357
77	4.8356	368
78	5.2823	384
79	6.0113	400
80	7.2365	400

Writing Grade 5 Form S'

Raw Score	Theta	Scale Score
0	-5.3699	100
1	-4.2174	100
2	-3.5863	100
3	-3.2329	100
4	-2.9895	102
5	-2.8043	109
6	-2.6548	115
7	-2.5291	119
8	-2.4202	123
9	-2.3237	127
10	-2.2365	130
11	-2.1566	133
12	-2.0825	135
13	-2.0129	138
14	-1.9470	140
15	-1.8840	143
16	-1.8235	145
17	-1.7649	147
18	-1.7077	149
19	-1.6517	151
20	-1.5967	153
21	-1.5422	155
22	-1.4881	157
23	-1.4343	159
24	-1.3804	161
25	-1.3265	163
26	-1.2722	165
27	-1.2176	167
28	-1.1625	169
29	-1.1068	171
30	-1.0505	173

Raw Score	Theta	Scale Score
31	-0.9934	175
32	-0.9356	177
33	-0.8771	179
34	-0.8177	182
35	-0.7576	184
36	-0.6966	186
37	-0.6348	188
38	-0.5723	190
39	-0.5089	193
40	-0.4448	195
41	-0.3798	197
42	-0.3140	200
43	-0.2472	202
44	-0.1795	205
45	-0.1106	207
46	-0.0407	210
47	0.0305	212
48	0.1031	215
49	0.1772	218
50	0.2530	220
51	0.3306	223
52	0.4102	226
53	0.4921	229
54	0.5763	232
55	0.6629	235
56	0.7522	239
57	0.8444	242
58	0.9394	245
59	1.0373	249
60	1.1380	253
61	1.2414	256

Raw Score	Theta	Scale Score
62	1.3473	260
63	1.4553	264
64	1.5651	268
65	1.6760	272
66	1.7878	276
67	1.8998	280
68	2.0118	285
69	2.1236	289
70	2.2350	293
71	2.3462	297
72	2.4572	301
73	2.5684	305
74	2.6804	309
75	2.7936	313
76	2.9085	317
77	3.0261	321
78	3.1470	326
79	3.2725	330
80	3.4036	335
81	3.5420	340
82	3.6899	346
83	3.8502	351
84	4.0274	358
85	4.2283	365
86	4.4647	374
87	4.7591	384
88	5.1627	399
89	5.8421	400
90	7.0262	400

Writing Grade 6 Form S'

Raw Score	Theta	Scale Score
0	-5.2165	100
1	-4.0474	100
2	-3.4005	100
3	-3.0365	100
4	-2.7853	101
5	-2.5941	108
6	-2.4396	114
7	-2.3095	119
8	-2.1967	123
9	-2.0965	127
10	-2.0059	130
11	-1.9228	133
12	-1.8456	136
13	-1.7731	139
14	-1.7044	141
15	-1.6388	144
16	-1.5757	146
17	-1.5146	149
18	-1.4552	151
19	-1.3972	153
20	-1.3402	155
21	-1.2840	157
22	-1.2285	159
23	-1.1734	161
24	-1.1187	163
25	-1.0641	165
26	-1.0095	167
27	-0.9550	169
28	-0.9003	171
29	-0.8454	174
30	-0.7902	176

Raw Score	Theta	Scale Score
31	-0.7347	178
32	-0.6788	180
33	-0.6225	182
34	-0.5658	184
35	-0.5086	186
36	-0.4508	188
37	-0.3925	190
38	-0.3335	193
39	-0.2740	195
40	-0.2138	197
41	-0.1528	199
42	-0.0911	202
43	-0.0286	204
44	0.0348	206
45	0.0991	209
46	0.1646	211
47	0.2311	214
48	0.2989	216
49	0.3681	219
50	0.4388	222
51	0.5111	224
52	0.5851	227
53	0.6611	230
54	0.7392	233
55	0.8195	236
56	0.9021	239
57	0.9872	242
58	1.0749	245
59	1.1652	249
60	1.2581	252
61	1.3535	256

Raw Score	Theta	Scale Score
62	1.4514	259
63	1.5515	263
64	1.6535	267
65	1.7572	271
66	1.8623	275
67	1.9684	279
68	2.0754	283
69	2.1831	287
70	2.2914	291
71	2.4005	295
72	2.5106	299
73	2.6220	303
74	2.7352	307
75	2.8508	312
76	2.9694	316
77	3.0918	321
78	3.2190	325
79	3.3522	330
80	3.4927	336
81	3.6422	341
82	3.8030	347
83	3.9781	354
84	4.1722	361
85	4.3924	369
86	4.6508	379
87	4.9705	391
88	5.4040	400
89	6.1202	400
90	7.3365	400

Writing Grade 7 Form S'

Raw Score	Theta	Scale Score
0	-5.7523	100
1	-4.5601	100
2	-3.8860	100
3	-3.5009	100
4	-3.2340	103
5	-3.0313	110
6	-2.8687	115
7	-2.7334	120
8	-2.6174	124
9	-2.5159	127
10	-2.4255	130
11	-2.3436	133
12	-2.2687	135
13	-2.1993	137
14	-2.1344	140
15	-2.0732	142
16	-2.0150	143
17	-1.9594	145
18	-1.9058	147
19	-1.8540	149
20	-1.8035	150
21	-1.7542	152
22	-1.7057	154
23	-1.6579	155
24	-1.6106	157
25	-1.5636	158
26	-1.5168	160
27	-1.4699	161
28	-1.4230	163
29	-1.3758	165
30	-1.3283	166
31	-1.2803	168
32	-1.2319	169
33	-1.1829	171

Raw Score	Theta	Scale Score
34	-1.1332	173
35	-1.0829	174
36	-1.0319	176
37	-0.9802	178
38	-0.9277	179
39	-0.8745	181
40	-0.8206	183
41	-0.7660	185
42	-0.7108	186
43	-0.6548	188
44	-0.5983	190
45	-0.5412	192
46	-0.4835	194
47	-0.4252	196
48	-0.3663	198
49	-0.3067	200
50	-0.2466	202
51	-0.1857	204
52	-0.1241	206
53	-0.0617	208
54	0.0016	210
55	0.0658	212
56	0.1311	214
57	0.1975	216
58	0.2652	219
59	0.3342	221
60	0.4046	223
61	0.4766	226
62	0.5501	228
63	0.6254	230
64	0.7023	233
65	0.7809	236
66	0.8613	238
67	0.9432	241

Raw Score	Theta	Scale Score
68	1.0269	244
69	1.1119	247
70	1.1983	249
71	1.2858	252
72	1.3744	255
73	1.4638	258
74	1.5540	261
75	1.6448	264
76	1.7363	267
77	1.8285	270
78	1.9215	273
79	2.0153	276
80	2.1102	279
81	2.2065	283
82	2.3045	286
83	2.4045	289
84	2.5069	293
85	2.6122	296
86	2.7210	300
87	2.8337	303
88	2.9513	307
89	3.0744	311
90	3.2043	315
91	3.3425	320
92	3.4909	325
93	3.6526	330
94	3.8319	336
95	4.0359	343
96	4.2764	351
97	4.5763	361
98	4.9875	374
99	5.6779	397
100	6.8731	400

Writing Grade 8 Form S'

Raw Score	Theta	Scale Score
0	-5.6565	100
1	-4.4587	100
2	-3.7770	100
3	-3.3855	100
4	-3.1136	100
5	-2.9076	106
6	-2.7431	112
7	-2.6070	117
8	-2.4913	121
9	-2.3908	125
10	-2.3020	128
11	-2.2224	131
12	-2.1500	133
13	-2.0835	136
14	-2.0218	138
15	-1.9641	140
16	-1.9096	142
17	-1.8578	144
18	-1.8083	145
19	-1.7606	147
20	-1.7144	149
21	-1.6695	150
22	-1.6256	152
23	-1.5825	153
24	-1.5399	155
25	-1.4978	156
26	-1.4560	158
27	-1.4143	159
28	-1.3725	161
29	-1.3306	162
30	-1.2885	164
31	-1.2460	165
32	-1.2030	167
33	-1.1594	169

Raw Score	Theta	Scale Score
34	-1.1151	170
35	-1.0701	172
36	-1.0242	173
37	-0.9774	175
38	-0.9296	177
39	-0.8808	178
40	-0.8309	180
41	-0.7800	182
42	-0.7278	184
43	-0.6745	186
44	-0.6200	188
45	-0.5643	190
46	-0.5074	192
47	-0.4494	194
48	-0.3900	196
49	-0.3294	198
50	-0.2675	200
51	-0.2043	203
52	-0.1398	205
53	-0.0739	207
54	-0.0067	210
55	0.0620	212
56	0.1321	215
57	0.2037	217
58	0.2767	220
59	0.3511	222
60	0.4270	225
61	0.5042	228
62	0.5828	231
63	0.6626	234
64	0.7436	236
65	0.8257	239
66	0.9087	242
67	0.9925	245

Raw Score	Theta	Scale Score
68	1.0771	248
69	1.1623	251
70	1.2482	254
71	1.3345	258
72	1.4213	261
73	1.5087	264
74	1.5967	267
75	1.6853	270
76	1.7747	273
77	1.8651	276
78	1.9568	280
79	2.0498	283
80	2.1445	286
81	2.2414	290
82	2.3405	293
83	2.4425	297
84	2.5477	301
85	2.6564	305
86	2.7693	309
87	2.8869	313
88	3.0098	317
89	3.1390	322
90	3.2754	327
91	3.4204	332
92	3.5762	338
93	3.7456	344
94	3.9330	350
95	4.1455	358
96	4.3953	367
97	4.7055	378
98	5.1284	393
99	5.8329	400
100	7.0403	400

Appendix D: 2007 Vertical Scaling Design

Step 1: Grades 5 and 4

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3						
	Grade 4		OP44	SU45			
	Grade 5		SU54	OP55			
	Grade 6						
	Grade 7						
	Grade 8						

Step 2: Grades 4 and 3

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3	OP33	SU34				
	Grade 4	SU43	OP44				
	Grade 5						
	Grade 6						
	Grade 7						
	Grade 8						

Step 3: Grades 5 and 6

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3						
	Grade 4						
	Grade 5			OP55	SU56		
	Grade 6			SU65	OP66		
	Grade 7						
	Grade 8						

Step 4: Grades 6 and 7

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3						
	Grade 4						
	Grade 5						
	Grade 6				OP66	SU67	
	Grade 7				SU76	OP77	
	Grade 8						

Step 5: Grades 7 and 8

		Items					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Students	Grade 3						
	Grade 4						
	Grade 5						
	Grade 6						
	Grade 7					OP77	SU78
	Grade 8					SU87	OP88

Appendix E: 2007 Vertical Scaling Item Parameters

Mathematics Grade 3

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1	-2.8289	31	-1.0542	61	-1.5802	91	-1.6397
2	-2.8289	32	-1.9134	62	-1.9039	92	-1.3445
3	-2.3674	33	-2.3248	63	-3.2659	93	0.5030
4	-1.9349	34	-0.5533	64	-3.1784	94	0.0523
5	-3.1565	35	-0.5234	65	-3.6727		
6	-1.9165	36	-1.6831	66	-2.1237		
7	-2.9605	37	-2.2273	67	-2.4766		
8	-3.5134	38	-0.7116	68	-2.8326		
9	-2.7590	39	-0.0255	69	-2.1604		
10	-3.5899	40	-0.3118	70	-2.0956		
11	-2.9749	41	-0.6549	71	-1.5840		
12	-2.4750	42	-0.5423	72	-2.2827		
13	-2.2828	43	-0.3999	73	-1.9825		
14	-1.4639	44	-2.3204	74	-2.7443		
15	-2.6452	45	-0.5826	75	-0.2788		
16	-2.6303	46	-3.7232	76	-1.2143		
17	-4.1474	47	-3.6099	77	-1.8231		
18	-2.2995	48	-3.1113	78	-2.1420		
19	-3.1516	49	-2.7093	79	-2.8222		
20	-1.3555	50	-0.5573	80	-1.5169		
21	-1.6911	51	-1.8210	81	-1.6199		
22	-2.1563	52	-1.7251	82	-1.5752		
23	-3.5267	53	-1.3208	83	-1.7558		
24	-1.4084	54	-1.1930	84	-0.7891		
25	-2.6770	55	-1.6570	85	-0.3757		
26	-1.9881	56	-0.9115	86	-1.2356		
27	-1.7598	57	-2.1981	87	0.1570		
28	-0.9878	58	-2.1344	88	-2.2464		
29	-1.4148	59	-1.4836	89	-1.6672		
30	-1.6346	60	-2.1705	90	-1.4303		

Mathematics Grade 4

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1	-1.3568	31	-1.5592	61	-1.8457	91	-0.0013
2	-1.2408	32	-1.8041	62	-1.5735	92	-0.7767
3	-0.8814	33	-0.0663	63	-2.7019	93	0.6334
4	-0.2297	34	-1.1577	64	-1.9994	94	-0.1058
5	-0.4677	35	-0.1147	65	-2.6518	95	0.0392
6	-0.6352	36	-1.1128	66	-1.5477	96	1.4420
7	-2.8457	37	-1.4850	67	-2.1216		
8	-1.0952	38	-0.7946	68	-1.5919		
9	-1.9060	39	-1.0541	69	-1.2198		
10	-0.9105	40	-1.3542	70	0.2857		
11	0.4287	41		71	-1.1513		
12	0.6741	42	-0.0857	72	-0.7915		
13	0.5269	43	-1.6381	73	-0.6878		
14	-0.4459	44	-0.6923	74	-0.6521		
15	-0.5082	45	-1.5974	75	-0.4266		
16	0.1620	46	-0.2013	76	-0.7894		
17	-0.1681	47	-2.9179	77	-0.9344		
18	0.5636	48	-1.7287	78	-0.5967		
19	-1.3851	49	-1.2675	79	-1.9560		
20	-0.3557	50	-1.1861	80	0.4345		
21	-1.1024	51	-0.4039	81	-0.9651		
22	-0.6729	52	-0.1272	82	-0.0384		
23	-1.3184	53	-1.2915	83	-1.5419		
24	-2.1129	54	-0.7645	84	-0.9469		
25	0.6601	55	-0.2368	85	0.4911		
26		56	0.0142	86	0.4323		
27	0.1883	57	-2.5592	87	0.5196		
28	-0.0949	58	0.6236	88	1.0539		
29	0.9666	59	-2.4591	89	0.6946		
30	0.1688	60	1.3347	90	-2.1113		

Mathematics Grade 5

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1	0.2430	31	0.2563	61	-1.2513	91	-2.5352
2	-1.1374	32	0.2721	62	0.1707	92	0.9364
3	-1.4107	33	1.0004	63	0.5742	93	-0.4448
4	-0.6478	34	1.1034	64	0.7699	94	-1.6185
5	-1.4418	35	1.5466	65	0.3421	95	1.6568
6	-0.7330	36	0.8918	66	1.0528	96	
7	0.0795	37	-0.6235	67	-0.1787	97	-2.3061
8	-1.3261	38	-0.5573	68	0.6137	98	-1.7419
9	-1.8281	39	-1.0064	69	0.9084	99	0.0980
10	-1.7528	40	-0.5668	70	0.4264	100	0.7395
11	0.4860	41	-0.5533	71	1.4481	101	0.1134
12	-0.2347	42	-0.7844	72	0.7345	102	-0.0673
13	-0.7828	43	1.0470	73	2.1093	103	0.3062
14	-0.9799	44	1.7387	74	-0.1804	104	0.6890
15	-0.6139	45	0.4207	75	0.3146	105	
16	0.4722	46	0.9318	76	-1.0892	106	-0.7996
17	0.8383	47	1.1792	77	-1.9681	107	-0.3627
18	0.0323	48	1.2777	78	0.0812	108	0.1124
19	0.7017	49	0.6913	79	-2.1912	109	
20	0.5022	50		80	0.0181	110	1.1478
21	-0.2339	51	-0.7644	81	0.7413	111	0.2395
22	0.2839	52	-0.4867	82	1.7422	112	0.7315
23	-1.6341	53	-1.0874	83	0.8614	113	0.7131
24	-1.1040	54	-0.1301	84	0.1272		
25	0.5426	55	-1.8012	85	1.4567		
26	-1.5710	56	-1.5072	86	0.1992		
27	0.5407	57	-1.1452	87	1.1201		
28	0.5169	58	1.9280	88	0.6921		
29	-0.3732	59	-0.6823	89	2.0091		
30	-0.4858	60	0.1865	90	1.0146		

Mathematics Grade 6

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1		31	-0.7939	61	0.6704	91	-0.5719
2		32	0.0366	62	1.1852	92	
3		33	-0.5048	63	1.5758	93	-0.6010
4	0.4495	34	-0.1615	64	1.6527	94	1.2601
5		35	1.3895	65	-0.2652	95	0.4007
6	0.0215	36	1.0923	66	1.6452	96	0.8343
7	0.3735	37	1.0030	67	1.2201	97	-0.3165
8	0.3375	38	1.3515	68	-0.6254	98	-0.3921
9	-0.5731	39	0.8677	69	-0.3384	99	1.6105
10	-0.5632	40	0.5838	70	1.8522	100	0.8216
11	1.9470	41	0.2909	71	0.9939	101	1.4601
12		42	1.0855	72	0.0884	102	2.3511
13	2.3273	43	2.1570	73	1.3038	103	1.2280
14	1.9761	44	1.2492	74	0.5231	104	0.1563
15	2.1892	45	1.4157	75	1.5465	105	2.9840
16	4.0231	46	1.9082	76	1.0031	106	0.7693
17	1.3887	47	1.3593	77	2.7053	107	-0.1900
18	0.8666	48	1.0128	78	2.5778	108	-0.4902
19	0.2592	49	-0.9193	79	1.1900	109	1.1349
20	0.7919	50	1.4976	80	1.1546	110	0.7909
21	0.9325	51	1.6723	81	1.2215	111	0.1987
22	1.3947	52	-0.0015	82	0.8889	112	
23	-0.8950	53	0.7809	83	0.8497	113	0.8111
24	0.4462	54	-0.3470	84	1.3505	114	2.3412
25	0.4503	55	0.1295	85	0.9314	115	1.7668
26		56	-1.1697	86	2.3232	116	2.3128
27	0.9129	57	-1.7431	87	0.9812		
28	2.4046	58	0.3350	88	1.6998		
29	0.1812	59	2.3287	89	2.3203		
30	0.0879	60	1.6443	90	1.8628		

Mathematics Grade 7

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1	1.6452	31	2.7910	61	0.6478	91	1.6412
2	0.9978	32	0.6919	62	2.1939	92	0.5607
3	1.3189	33	2.0991	63	1.0881	93	1.5911
4	-0.0182	34	0.4326	64	1.3910	94	1.2450
5	0.3476	35	0.9154	65	1.6935	95	-0.1923
6	1.3623	36	2.6043	66	1.1128	96	0.4864
7	1.4797	37	2.8881	67	1.7952	97	1.1370
8	1.3931	38	1.4167	68	2.2867	98	3.2034
9	1.0928	39	1.3385	69	0.4971	99	0.3455
10	2.0582	40	1.2454	70	1.0727	100	-0.3284
11		41	0.8032	71	1.1418	101	1.2226
12	-0.4667	42	2.2834	72	1.5024	102	0.7798
13		43	1.0035	73	2.1359	103	0.8478
14	0.8611	44	1.2344	74	1.9791	104	1.8159
15	1.2069	45	2.3294	75	0.6566	105	2.1487
16	1.1550	46		76	1.6999	106	1.9329
17	1.7152	47	2.2300	77	3.1412	107	2.8423
18	2.1062	48	2.1005	78	3.6413	108	1.6330
19		49	1.1848	79	2.6361	109	1.9439
20	0.7116	50	1.8001	80	2.7781	110	1.5873
21	1.7317	51	-0.5794	81	4.1535	111	1.4191
22	2.2566	52	2.1742	82	3.3380	112	1.5634
23	1.7549	53	2.0938	83	1.3214	113	2.2074
24	1.4677	54	1.7031	84	2.2025	114	0.9626
25	1.9449	55	1.5473	85	3.2240	115	1.9166
26	1.3225	56	2.0468	86	1.7455	116	2.8778
27	2.9138	57	1.4236	87	1.2290	117	1.0076
28	1.2778	58	0.4976	88	2.5369	118	1.6561
29	0.6929	59	1.5459	89	2.7677	119	2.3484
30	1.9322	60	1.4187	90	2.2121	120	1.4833

Mathematics Grade 8

Item	Rasch	Item	Rasch	Item	Rasch	Item	Rasch
1	0.7142	31	2.4179	61	2.0525	91	1.3803
2	-0.1553	32	3.7860	62	2.1408	92	2.4175
3	1.5577	33	2.6466	63	1.5228	93	1.1683
4	1.2378	34	1.8214	64	2.8161	94	2.3084
5	2.4511	35	1.4226	65	1.5893	95	1.9837
6	1.4852	36	1.4770	66	3.0423	96	2.2688
7	2.2539	37	2.6153	67	2.8613	97	0.2606
8	2.1423	38	2.3687	68	2.7289	98	1.0831
9	1.2587	39	3.6861	69	3.1240	99	1.5549
10	2.1683	40	2.2154	70	3.5272	100	2.3493
11	2.4212	41	2.0238	71	1.6793	101	2.4743
12	2.1638	42	1.7555	72	2.0206	102	3.4871
13	1.2769	43	1.2808	73	1.6825	103	2.8794
14	3.3054	44	1.8446	74	2.8544	104	1.5141
15	1.0512	45	1.8863	75	3.3514	105	1.5069
16	3.2360	46	1.8680	76	2.2902	106	3.3533
17	4.0854	47	1.1331	77	3.0849	107	0.9235
18	2.0304	48	1.5075	78	2.1684	108	2.6986
19	2.2099	49	1.6237	79	2.8776	109	1.3295
20	1.7169	50	2.2365	80	1.6935	110	1.5370
21	2.0504	51	1.6284	81	2.6452	111	2.6014
22	1.9332	52	1.3946	82	-0.3455	112	3.0736
23	1.8264	53	1.6339	83	2.4389	113	2.5804
24	2.2983	54	2.1826	84	0.6018	114	0.9604
25	1.1153	55	1.3707	85	0.8505	115	2.3058
26	1.0233	56	3.1479	86	0.2512	116	3.2384
27	1.7201	57	1.4934	87	1.3091	117	3.8168
28	1.9424	58	1.6362	88	2.7917		
29	3.3517	59	1.8087	89	1.1807		
30	3.1742	60	2.0301	90	0.9951		

Reading Grade 3

Item	Rasch	Item	Rasch	Item	Rasch
1	-2.2752	31	-1.2373	61	-0.6064
2	-1.6975	32	-2.2198	62	0.0588
3	-1.8020	33	-3.4067	63	0.2450
4	-1.7297	34	0.1471	64	2.0209
5	-2.9475	35	-3.0766	65	-0.1800
6	-0.9245	36	-2.2975	66	-0.4630
7	-0.2724	37	-2.2905	67	-0.0562
8	-2.6627	38	-2.0748	68	0.3684
9	-0.7832	39	-0.4722	69	0.6888
10	-1.7056	40	-2.1046	70	-0.2298
11	0.0274	41	-0.3336	71	0.1948
12	-1.1025	42	-0.2262	72	2.9546
13	-0.7582	43	-1.3892	73	0.0359
14	-0.8707	44	-2.0555		
15	-2.7154	45	-2.4293		
16	-2.2477	46	0.1517		
17	-1.3746	47	-1.1577		
18	-1.2063	48	-1.2690		
19	0.3762	49	-2.5622		
20	-1.6417	50	-1.0729		
21	-0.2227	51	-0.4442		
22	-0.9508	52	1.8735		
23	-1.8797	53	-0.5480		
24	-2.2929	54	0.1520		
25	-1.1315	55	-0.9777		
26	-2.9637	56	-0.5158		
27	-0.9948	57	-0.9274		
28	-0.0592	58	0.5213		
29	0.0879	59	0.6859		
30	0.6637	60	-0.1596		

Reading Grade 4

Item	Rasch	Item	Rasch	Item	Rasch
1	-0.7089	31	0.1402	61	-0.6392
2	-0.1836	32	0.3316	62	-1.2143
3	-1.6700	33	-2.7555	63	0.0138
4	-0.4203	34	-0.4246	64	0.6999
5	-0.3009	35	-2.5469	65	0.9780
6	-0.9574	36	-2.4156	66	0.6070
7	-2.5057	37	-2.9201	67	-0.3407
8	0.4718	38	0.7037	68	1.3876
9	0.0449	39	-1.3012	69	1.0430
10	0.6372	40	-0.8776	70	1.2334
11	-0.2606	41	-0.3881	71	0.1436
12	-1.6258	42	-0.6515	72	0.2120
13	-0.9046	43	-1.9602	73	0.6090
14	-1.7263	44	-1.2707	74	0.7324
15	0.7509	45	-2.0750		
16	-0.0586	46	-1.3029		
17	-2.1638	47	-0.4253		
18	0.6454	48	-1.9542		
19	-0.3429	49	-0.4621		
20	-1.0316	50	-0.9550		
21	-1.6007	51	0.1880		
22	-0.8506	52	-1.3278		
23	-0.6032	53	-1.1946		
24	0.2684	54	-0.2376		
25	-1.0976	55	-0.9079		
26	-0.9530	56	0.1344		
27	-1.3296	57	-0.3709		
28	-2.0063	58	-0.8250		
29	-0.3332	59	0.1093		
30	0.2858	60	0.2162		

Reading Grade 5

Item	Rasch	Item	Rasch	Item	Rasch
1	0.4761	31	0.0701	61	-0.4597
2	-1.2916	32	-0.9537	62	0.6785
3	0.0686	33	-0.0553	63	1.3191
4	-0.6705	34	-0.4766	64	1.1737
5	-0.4715	35	-1.9012	65	0.0610
6	-1.0044	36	-0.6014	66	-0.3632
7	-1.0657	37	-2.1300	67	0.7273
8	-1.0273	38	0.0412	68	2.7605
9	-0.0134	39	-0.1381	69	1.0828
10	0.5252	40	-0.8039	70	2.1633
11	-1.0322	41	-0.4789	71	0.3668
12	-0.7421	42	0.4830	72	0.1463
13	0.2260	43	-0.6770	73	1.7036
14	-0.0905	44	-1.1085	74	2.7479
15	-1.1737	45	-0.0930	75	0.0783
16	0.4392	46	-0.2227	76	0.9983
17	0.3024	47	-0.5283	77	0.6077
18	-0.8568	48	-1.2816	78	1.5340
19	-0.2929	49	0.5868	79	2.5326
20	-0.5229	50	0.3920	80	1.0631
21	-1.2249	51	0.1618		
22	-0.6656	52	-0.5918		
23	1.3514	53	-0.1984		
24	0.3625	54	0.0811		
25	-0.4868	55	-0.3852		
26	-0.3425	56	0.7193		
27	-0.6531	57	0.6612		
28	0.1220	58	-0.5023		
29	-0.1804	59	0.3689		
30	0.8700	60	-0.2702		

Reading Grade 6

Item	Rasch	Item	Rasch	Item	Rasch
1	0.3440	31		60	-0.0156
2	0.2273	32	-0.8286	61	-0.1455
3	-0.2147	33	-0.9278	62	-0.3798
4	-0.1598	34	-1.5433	63	1.9635
5	-1.3337	35	-2.1579	64	-0.1416
6	-0.7368	36	0.8403	65	0.7694
7		37	0.2109	66	1.1086
8		38	-0.0994	67	2.4647
9		39	-0.2018	68	0.1839
10		40	0.2185	69	1.5365
11	-0.3305	41	-1.6415	70	1.9504
12	0.6884	42	-0.5216	71	0.7398
13	1.4165	43	1.2414	72	0.7333
14	0.4371	44	-0.8555	73	1.9378
15	0.9828	45	0.1595	74	1.7457
16	-0.1848	46	0.0015	75	2.5750
17		47	0.9341	76	1.7614
18		48	-0.0080	77	1.1785
19	-0.6227	49	0.4917	78	1.6963
20		50	-0.1116	79	1.3511
21	-0.3355	51	-0.6936	80	0.7651
22	-0.9129	52	-0.2615		
23		53	-0.9350		
24	0.3786	54	-0.4850		
25	-0.1675	55	-0.0487		
26	1.5991	56	0.4576		
27	0.6001	57	0.4803		
28	0.3259	58	0.9775		
29	0.6243	59	2.5622		
30					

Reading Grade 7

Item	Rasch	Item	Rasch	Item	Rasch
1	0.8380	31	-0.8242	61	0.7137
2	-0.1030	32	2.0739	62	0.7800
3	0.2115	33	-0.7778	63	0.5800
4	-0.5060	34	0.0685	64	0.8840
5	0.9224	35	-1.0503	65	3.8243
6	-0.2342	36	-1.2692	66	0.9976
7	1.6791	37	-0.6115	67	3.6574
8	0.9395	38	0.4427	68	2.8367
9	1.2416	39	-0.5769	69	1.1184
10	1.2094	40	-0.8495	70	0.3200
11	0.6480	41	1.1862	71	-0.0672
12	1.6882	42	-1.1745	72	2.0750
13	0.7562	43	-0.9524	73	2.6772
14	0.0093	44	2.1948	74	1.2851
15	1.1575	45	-0.6199	75	0.7597
16		46	2.5497	76	0.9321
17	0.8410	47	-0.0239	77	2.2661
18	0.4500	48	0.2340	78	1.4464
19	0.5534	49	0.2219	79	1.3375
20	-0.2206	50	0.2804		
21	1.4109	51	0.3534		
22		52	0.0063		
23	1.2756	53	0.2784		
24	-1.0703	54	0.9002		
25	0.3006	55	1.0462		
26	0.7926	56	0.4724		
27	-0.0052	57	0.0568		
28	0.3838	58	0.3424		
29	0.8190	59	0.1761		
30		60	0.2078		

Reading Grade 8

Item	Rasch	Item	Rasch	Item	Rasch
1	1.0162	31	0.4008	61	2.1807
2	-0.1109	32	-0.8264	62	0.5464
3	-0.1922	33	-0.4446	63	0.9193
4	0.9395	34	-1.1448	64	1.2376
5	-0.5964	35	-1.0438	65	0.3397
6	0.9717	36	0.2674	66	2.1689
7	0.6273	37	1.6623	67	0.4247
8	0.6985	38	-1.2522	68	1.4031
9	-0.0418	39	0.5921	69	-0.2368
10	0.6438	40	-1.2833	70	1.3233
11	0.9471	41	-0.5995	71	1.3018
12	2.6963	42	0.7230	72	2.9400
13	2.1340	43	0.0812	73	1.9515
14	1.5278	44	0.5894	74	1.2048
15	1.1405	45	0.4144	75	1.7735
16	1.6237	46	-0.2184	76	1.1724
17	0.8520	47	-0.5136	77	1.4446
18	1.5780	48	-1.0386	78	1.8979
19	0.2812	49	-0.0661	79	2.7678
20	1.7631	50	0.9103		
21	2.0243	51	-0.1334		
22	0.9542	52	1.7172		
23	0.5519	53	0.0844		
24	1.1824	54	1.1377		
25	-0.2073	55	-0.2513		
26	0.3962	56	-0.5704		
27	0.1540	57	1.1669		
28	0.1187	58	1.0414		
29	0.9496	59	1.1940		
30	1.2782	60	0.5009		