Wisconsin Knowledge and Concepts Examinations-CRT

December 2004
Field Test/Standardization
Technical Report

CTB McGraw-Hill

June 30, 2005

CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, California 93940

Foreword

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

Table of Contents

Foreword 1
Part 1: Overview 5
Part 2: Content Rationale 6
General Test Specifications 6
Content Rationale 6
Development Procedures 7
Content Areas 9
Part 3: Scale Development 14
Description of Vertical and Horizontal Linkages 14
Description of the Sample of Students Recruited for the Study 17
General Description of IRT Methods Used to Calibrate Items 22
Item Response Theory Models 22
Advantages of the 3PL and 2PPC Models 23
Parameter Estimation 23
Fit Rating 25
Description of the Calibration/Scaling Methods Used 26
Vertical Scaling Method 1 26
Vertical Scaling Method 2 27
Vertical Scaling Method 3 27
Vertical Scaling Method 4 27
Horizontal Scaling/Form Equating 27
Description of Vertical Linking Set Performance 29
Summary of Results of Alternative Vertical Scaling Options 39
Final Scale Selection 42
Part 4: Description of the First Operational Form 59
Classical Form Summary 59
IRT Form Summary 61
DIF Summary 68

Tables

Table 1 Reporting Categories for WKCE-CRT Mathematics 11
Table 2 Reporting Categories for WKCE-CRT Reading 13
Table 3 Total Group and Subgroup Frequencies by Grade and Form For Mathematics..... 18
Table 4 Total Group and Subgroup Proportions by Grade and Form For Reading..... 19
Table 5 Summary Statistics For Common Items by Grade and Form for Mathematics..... 20
Table 6 Summary Statistics For Common Items by Grade and Form for Reading..... 21
Table 7 Summary of Vertical Linking Set Performance – Mathematics..... 31
Table 8 Summary of Vertical Linking Set Performance – Reading 32
Table 9 Raw Score Descriptive Statistics for Mathematics Vertical Scaling Forms..... 39
Table 10 Raw Score Descriptive Statistics for Reading Vertical Scaling Forms 39
Table 11 Summary of Vertical Scaling Outcomes for Mathematics 40
Table 12 Summary of Vertical Scaling Outcomes for Reading 41
Table 13 Classical Statistics for Mathematics Fall 2005 Operational Forms Grade 3–Grade 8/Grade 10..... 59
Table 14 Classical Statistics for Reading Fall 2005 Operational Forms Grade 3–Grade 8/Grade 10..... 60
Table 16 Mathematics Items Flagged for DIF by Focal Group..... 70
Table 17 Reading Items Flagged for DIF by Focal Group 71

Figures

Figure 1 Description of Vertical Linkages Between Grades 15
Figure 2 Example of Horizontal Linkages within a Grade 16
Figure 3 Math Grade 03/04 Anchors 33
Figure 4Math Grade 04/05 Anchors 33
Figure 5 Math Grade 05/06 Anchors 34
Figure 6 Math Grade 06/07 Anchors 34
Figure 7 Math Grade 07/08 Anchors 35
Figure 8 Math Grade 08/10 Anchors 35
Figure 9 Reading Grade 03/04 Anchors 36
Figure 10 Reading Grade 04/05 Anchors 36
Figure 11 Reading Grade 05/06 Anchors 37
Figure 12 Reading Grade 06/07 Anchors 37
Figure 13 Reading Grade 07/08 Anchors 38
Figure 14 Reading Grade 08/10 Anchors 38
Figure 15 Test Characteristic Curves for Method 1 Mathematics 43
Figure 16 Test Characteristic Curves for Method 2 Mathematics 44
Figure 17 Test Characteristic Curves for Method 3 Mathematics 45
Figure 18 Test Characteristic Curves for Method 1 Mathematics 46
Figure 19 Cumulative Distribution Functions for Method 1 Mathematics 47
Figure 20 Cumulative Distribution Functions for Method 2 Mathematics 48
Figure 21 Cumulative Distribution Functions for Method 3 Mathematics 49
Figure 22 Cumulative Distribution Functions for Method 4 Mathematics 50
Figure 23 Test Characteristic Curves for Method 1 Reading 51
Figure 24 Test Characteristic Curves for Method 2 Reading 52
Figure 25 Test Characteristic Curves for Method 3 Reading 53
Figure 26 Test Characteristic Curves for Method 4 Reading 54
Figure 27 Cumulative Distribution Functions for Method 1 Reading 55
Figure 28 Cumulative Distribution Functions for Method 2 Reading 56
Figure 29 Cumulative Distribution Functions for Method 3 Reading 57
Figure 30 Cumulative Distribution Functions for Method 4 Reading 58
Figure 31 Test Characteristics Curves for WKCE-CRT Math 2005–06 Operational Form 62
Figure 32 Standard Error Curves for WKCE-CRT Math 2005–06 Operational Form 63
Figure 33 Information Curves for WKCE-CRT Math 2005–06 Operational Form 64
Figure 34 Test Characteristic Curves for WKCE-CRT Reading 2005–06 Operational Form 65
Figure 35 Standard Error Curves for WKCE-CRT Reading 2005–06 Operational Form 66
Figure 36 Information Curves for WKCE-CRT Reading 2005–06 Operational Form 67

Part 1: Overview

Beginning in the 2005–2006 school year, the federal No Child Left Behind Act requires all states to test all students in reading and mathematics in Grades 3 through 8 and once in high school (Grade 10 under Wisconsin law s. 118.30). These tests developed in response to this legislation are referred to as the Wisconsin Knowledge and Concepts Examination – Criterion-Referenced Tests (WKCE-CRT) and will replace WKCE reading and mathematics tests beginning in fall 2005. Student performance on these tests is reported in proficiency categories and used to determine the adequate yearly progress of students at the school, district and state levels.

As part of the development of these tests, CTB/McGraw-Hill conducted item pilot testing in May 2004 and forms calibration in December 2004 based on a stratified sampling design drawing from all public schools in the state. The Wisconsin Department of Public Instruction also contracted with three national experts to evaluate the work of CTB/McGraw-Hill as well as to advise the department on issues of validity and reliability of the new WKCE-CRT test design for reading and mathematics.

This report summarizes the information about the Calibration/Scaling Field Test that was conducted as part of the development of the Wisconsin Knowledge and Concepts Examinations-Criterion Reference Test (CRT) in December 2004. This Field Test focused exclusively on two content areas: mathematics and reading. The field test was conducted for two primary purposes: (a) develop new vertical reporting scales for both mathematics and reading that span Grades 3 though 8 and Grade 10, and (b) calibrate and scale a pool of items for inclusion in future forms. To that end, this report consists of the following 4 sections.

1. Overview;
2. Content Rationale;
3. Scale Development;
4. Description of the First Operational Form.

Part 2: Content Rationale

General Test Specifications

All test items developed for the new WKCE-CRT tests in reading and mathematics are of one of two formats: selected-response/multiple-choice (MC) or constructed-response (CR). The test reporting categories and items assigned to measure each reporting category are aligned to the Wisconsin Model Academic Standards in reading and mathematics with grade-level appropriate descriptors supporting learning expectations for tests administered in the fall semester. The test design draws approximately 80% of the total score points from selected-response items and 20% of the score points from student-generated constructed-response items.

All students in Grades 3 through 8 and 10 will be tested in reading and mathematics using these new customized WKCE-CRT tests beginning in fall 2005. Students with disabilities will be allowed accommodations during these tests unless an alternate assessment is required based on an IEP process. Students whose English language proficiency as tested on state-approved language proficiency examinations is level three or higher will take the WKCE-CRT tests with allowable accommodations. English language learners with language proficiency scores less than three will take an alternate assessment. All alternate assessments are aligned to state standards.

Content Rationale

Establishing the Content Rationale

The following principles guided the test development process to establish the content for WKCE-CRT tests:

- provide valid, equitable measurement of achievement in the areas of reading and mathematics;
- offer multiple ways of measuring student progress;
- give information useful for improving student's understanding of key concepts;
- engage and motivating students so they will perform their best work; and
- reflect current curricula and state standards.

Establishing the content framework and eligible test content began in August 2003 with a workshop with Wisconsin educators. At this workshop facilitated by CTB and DPI staff, committees of educators carefully considered what knowledge and skills students should have by the fall of each school year. Committees then defined the eligible test content and framework documents, ensuring that the test framework they designed incorporated the content and performance standards enumerated in the Wisconsin Model Academic Standards.

During the August workshop, the Wisconsin educators reviewed the framework documents to determine which could be efficiently and effectively measured using MC items and which were best measured using CR items and made recommendations regarding how much emphasis should be given to each content standard on any given test form. The WKCE-CRT tests sample commonly taught processes, skills, and knowledge. They do not measure all of the skills that make up an educational domain. The outcomes of the workshops were the test framework for each grade and content area. Following the workshop, the DPI conducted follow-up meetings with educators to refine and articulate the content and sub-skills in the test framework across grade levels. CTB researchers provided guidelines regarding the number of items needed to achieve reliable test scores. The result was a draft test blueprint that specified the amount of testing time required for each content area, the total number of score points for each test, the total number of score points for each content standard, and the numbers of MC and CR test items that would comprise each form. CTB and DPI then reviewed the draft blueprints to ensure that the tests would provide a balanced measure of the eligible performance standards and yield highly reliable and valid scores; modifications were made as necessary to achieve appropriate content coverage and balance. Together, the Wisconsin educators, DPI, and CTB reviewed a variety of sample test items and discussed the characteristics of the types of items that would be best suited for inclusion on the WKCE-CRT field test.

Development Procedures

Designing Assessment Specifications

Using the principles outlined in the content rationale, WKCE-CRT test developers wrote detailed reading passage and stimulus specifications and test item specifications. These specifications ensured that stimulus materials and items met the content criteria established for the tests and that they were well constructed and written in language appropriate for the various testing levels. The specifications were applied to all aspects of development, including the creation of tryout materials, analysis of tryout data, and selection of materials for the final tests.

Writing and Developing Assessment Materials

A staff of professional item writers—many of them experienced teachers—researched, collected, and wrote the tryout material. All assessment materials were carefully reviewed for content and editorial accuracy by test development specialists and the content specialists at the Wisconsin Department of Public Instruction and Wisconsin classroom teachers. Artists and designers worked with the writers during development to ensure graphic and textual consistency.

Item development for the WKCE-CRT operational test forms began with selecting a variety of literary, informational, and everyday text reading passages. The emphasis was on selecting reading passages that are engaging to students and contain appropriate subject matter, but are not familiar to the students (which would create a potential source of bias). Materials were reviewed and approved by committees of Wisconsin educators.

The test developers wrote approximately 2,400 items for each academic area. Items were reviewed and edited by content specialists and also reviewed for editorial style. An important element in creating the test and writing the test items was to use a learning and thinking skills framework. The skills, sub-skills, and framework documents of the reading content framework incorporate thinking skills. The mathematics items were written to address one of four levels of cognitive demand:

Level 1: Recognizing and Recalling

Level 1 tasks require the student to recognize and recall basic facts, terms, concepts, and definitions of the content and processes of mathematics. Level 1 tasks also include computation with whole numbers, fractions, decimals, and integers.

Level 2: Using Fundamental Concepts and Procedures

Level 2 tasks require the student to describe or apply basic facts, terms, rules, concepts and definitions of the content and processes of mathematics.

Level 3: Concluding and Explaining

Level 3 tasks require the student to demonstrate an understanding of complex ideas, to draw conclusions based on this understanding, and to communicate ideas and conclusions effectively.

Level 4: Evaluating, Extending, and Making Connections

Level 4 tasks require the student to synthesize skills and techniques from various concepts of mathematics to solve multifaceted problems, and to justify conclusions using mathematical definitions, properties, and principles. Level 4 tasks also include supporting mathematical arguments with definitions, properties, and principles.

Matching mathematics items to a thinking skill level helped ensure that test items required a range of cognitive skills, and not simply recall.

Documenting Content

An integral part of the development process was documentation of content using Wisconsin's curriculum framework. This procedure ensured that items would be sound in content and format, and targeted appropriately to the grade level in which the associated concepts are typically taught. Items were written to measure the eligible academic standards and also the more specific framework documents, as defined in the item specifications.

All items were reviewed by committees of Wisconsin educators at meetings held in Wisconsin. During the item content review meetings, committee members considered whether items measured a content framework documents and that items were at an appropriate level of difficulty and cognitive demand and represented the breadth and depth of content taught in the academic areas. Careful attention was given to ensure that the test items measured only the content and skills that every student in Wisconsin has an opportunity to learn.

Minimizing Bias

Four procedures are used to reduce bias in WKCE-CRT items. The first is based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias can occur only if the test is measuring different things for different groups. If the test includes irrelevant skills or knowledge (however common), the possibility of bias is increased. Thus, careful attention is paid to content validity.

The second step is to follow the McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers are directed to two McGraw-Hill publications: Guidelines for Bias-Free Publishing (McGraw-Hill, 1983) and Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals (Macmillan/McGraw-Hill, 1993). Whenever CTB staff review items, these guidelines are kept in mind. Such internal editorial reviews are completed by at least four separate people: the content editor, the style editor, the development supervisor, and the proofreader.

In the third procedure, Wisconsin educators trained in equity issues reviewed the items from the perspective of various ethnic groups and groups of students. These educators identified assessment materials that might reflect possible bias in language, subject matter, or representation of people. Their comments and suggestions were considered carefully during the revision and selection of reading passages, stimulus materials, and items for the final tests.

It is believed that these three procedures improve the quality of the items and reduce bias. However, current evidence suggests that expertise in this area is no substitute for data; reviewers are sometimes wrong about which items work to the disadvantage of a group, apparently because some of their ideas about how students will react to items may be faulty (Sandoval & Mille, 1979; Jensen, 1980; Scheuneman, 1984). Thus, differential item functioning statistics were calculated for every item based on field test data. CTB reviewed item data to determine whether each item became part of the pool of items eligible for use on live test forms.

Content Areas

Mathematics

Development of the mathematics content framework was aligned with the National Council of Teachers of Mathematics (NCTM) Standards, just as the Wisconsin Model Academic Standards for mathematics were aligned with NCTM standards. More emphasis was placed on a balance of skills, concepts, knowledge, and problem solving than on procedural and computational processes. This emphasis produced assessment materials rich and varied in context and relevant to students' lives.

Mathematics is presented as a useful tool for making sense of the world. Items reinforce the connections between mathematical principles and their applications in everyday life. A broadening and deepening of higher-order thinking skills accompany adherence to NCTM Standards, as does an emphasis on the contextual integration of mathematics with other content areas such as social studies and science. The reporting categories for mathematics can be found in Table 1.

Reading

The development of the reading test is based on the position that reading is a complex, interactive process that is a primary means of acquiring and using information. Because reading is fundamental to the mastery of other school subjects, students at all grade levels must learn to understand what they read. They must know and use various strategies-ways

of unlocking the meaning of words and larger blocks of text-to become successful readers. The reading test requires that students construct meaning from a single text or across two texts, including understanding word meaning and stated information, making inferences, and evaluating the author's craft and ideas in the text.

Among the reading materials are excerpts from traditional and contemporary literature, informational selections from current publications, and real-life documents and graphics. In recognition of the varied background of American students, reading selections represent a range of authors, perspectives, and cultural experiences. Consistent with classroom practice, comprehension items focus on the central meaning of a passage rather than just on surface details. Item progression reflects the reading process by moving from initial understanding through development of interpretation and on to evaluation and extension of concepts.

Reading items should focus on the Wisconsin WKCE-CRT reporting categories, sub-skills, and framework documents that were created to be consistent with the Wisconsin Model Academic Standards. Items should emphasize both academic and real-world situations and not be strictly theoretical for the sake of being academic. For example, items should not test recall of terminology or discrete facts in isolation. Rather, as much as possible, realistic and practical contexts should be used for assessing the eligible content. On the other hand, items should not be at a minimal, basic level that might be characteristic of a functional level or minimum competency test. Items should address a range of difficulty and cognitive demand so that distinctions can be made among students at the basic, proficient, or advanced level. The reporting categories for reading can be found in Table 2.

Table 1
Reporting Categories for WKCE-CRT Mathematics

Sub-skills	Objective	Score Points						
		G03	G04	G05	G06	G07	G08	G10
	A. Mathematical Processes							
		10	12	14	14	14	16	12
Aa	Reasoning							
Ab	Communication							
Ac	Connections							
Ad	Representation							
Ae	Problem Solving							
	B. Number Operations and Relationships							
		14	13	15	14	15	10	7
Ba	Number Concepts							
Bb	Number Computation							
	C. Geometry							
		6	6	7	7	7	3	2
Ca	Describing Figures							
Cb	Special Relationships and Transformations							
Cc	Coordinate System							
	D. Measurement							
		10	10	12	12	11	13	12
Da	Measurable Attributes							
Db	Direct Measurement							
Dc	Indirect Measurement							

Table 1 (Continued)
Reporting Categories for WKCE-CRT Mathematics

Sub-skills	Objective	Score Points G03	G04	G05	G06	G07	G08	G10
	E. Statistics and Probability							
		10	10	12	12	11	10	15
Statistics and Probability								
Ea	Data Analysis and Statistics							
Eb	Probability							
	F. Algebraic Relationships							
		10	12	12	12	12	15	15
Algebraic Relationships								
Fa	Patterns, Relations, and Functions							
Fb	Expressions, Equations, and Inequalities							
Fc	Properties							
	Total Score Points of Mathematics All Items	***65***	***68***	***76***	***76***	***76***	***74***	***69***
	Total Number of Selected Response Items	***50***	***50***	***55***	***55***	***55***	***50***	***55***
	Total Number of Constructed Response Items	***5***	***6***	***7***	***7***	***7***	***8***	***6***

Table 2
Reporting Categories for WKCE-CRT Reading

Sub-skills	Objective	Score Points						
		G03	G04	G05	G06	G07	G08	G10
	Aa. Determines meaning of words or phrases in context	16	13	13	12	12	12	15
Aa.1	Uses context clues to determine meaning of words or phrases							
Aa.2	Uses knowledge of word structure to determine meaning of words							
Aa.3	Uses word reference materials to determine meaning of words and phrases							
	Ab. Understands Text	19	18	17	16	16	16	15
Ab.1	Demonstrates understanding of literal meaning by identifying stated information in literary text							
Ab.2	Demonstrates understanding of literal meaning by identifying stated information in informational text							
Ab.3	Demonstrates understanding of explicitly stated sequence of events in literary and informational text							
	Ac. Analyzes Text	23	26	26	25	25	25	30
Ac.1	Analyzes literary text							
Ac.2	Analyzes informational text							
Ac.3	Analyzes author's use of language in literary and informational text							
	Ad. Evaluates and Extends Text	8	9	13	16	16	16	24
Ad.1	Evaluates and extends literary text							
Ad.2	Evaluates and extends informational text							
Ad.3	Evaluates and extends author's use of language in literary and informational text							
	Total Score Points	***66***	***66***	***69***	***69***	***69***	***69***	***84***
	Total Number of Selected Response Items	***60***	***60***	***60***	***60***	***60***	***60***	***60***
	Total Number of Constructed Response Items	***2***	***2***	***3***	***3***	***3***	***3***	***8***

Part 3: Scale Development

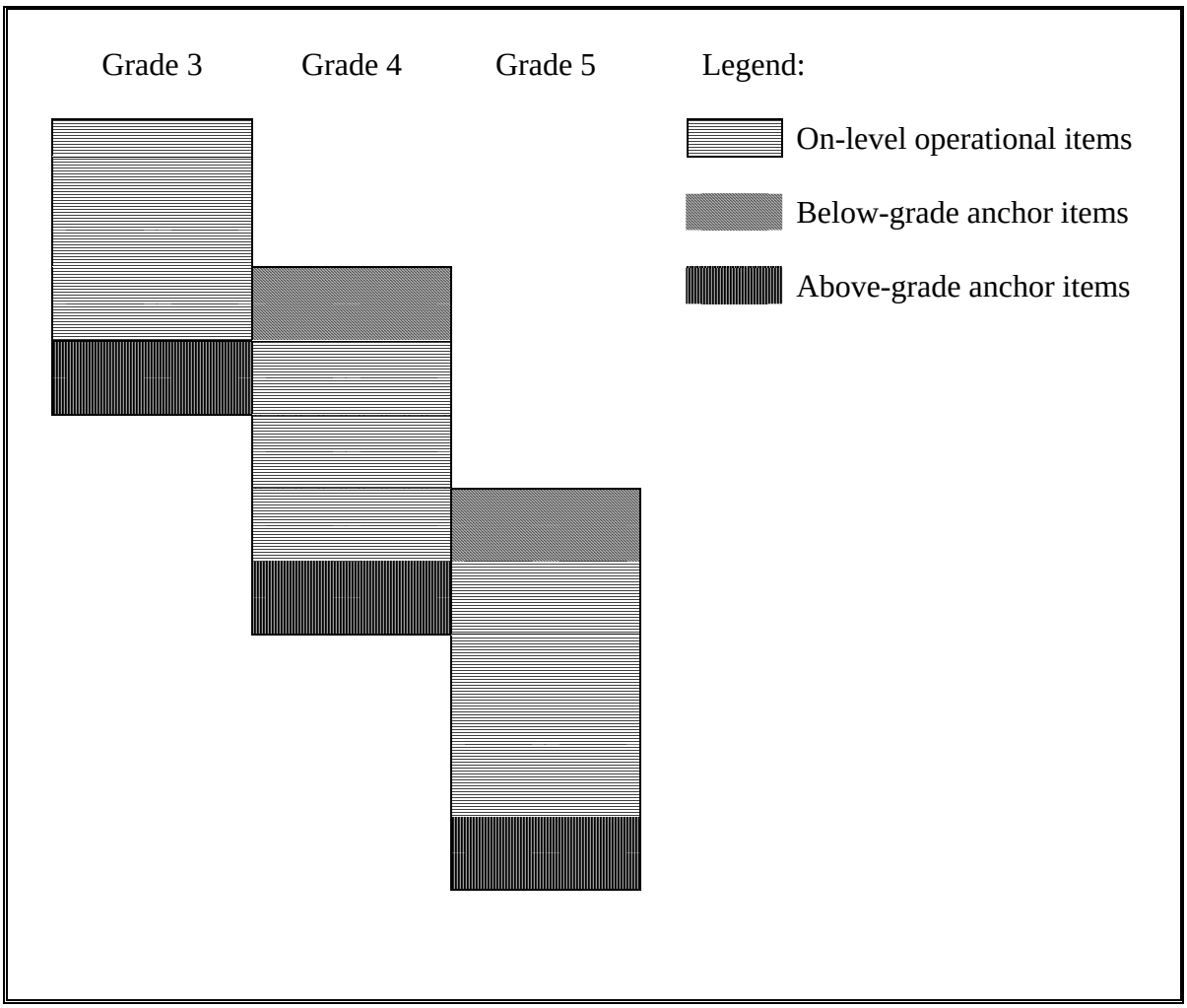
Description of Vertical and Horizontal Linkages

The basic design called for multiple forms to be delivered at each grade level. In general, each form consisted of three components. The bulk of each form was unique and was comprised of items that were designated as operational. These collections of items were assembled to conform to a specific content blueprint, and many, but not all, of these items had been field tested at a previous administration. Each delivered form was then finalized by augmenting the operational core with two additional sets of items. First, within a grade level and content area, the same set of common items was added. This set of items was included to provide a horizontal link of all of the items for each grade level to the vertical scale. Second, an additional set of items was added that varied in composition across the forms. These items were either new items being field tested or operational items from an adjacent grade that were being delivered to assist in the development of a vertical scale.

A single figure to depict all of the linking relationships would be quite complex. Thus, separate figures are used to provide examples of the horizontal and vertical linkages. Figure 1 presents an excerpt of the complete vertical linking pattern. This figure shows the linkages for Grades 3, 4 and 5. The same linking design was used for both mathematics and reading. The only distinction between content areas and grades is the specific number of items used to form each link. By virtue of being at the ends of the vertical scale, forms for Grades 3 and 10 were slightly different. Forms for these two grades needed only be linked to the vertical scale at one end.

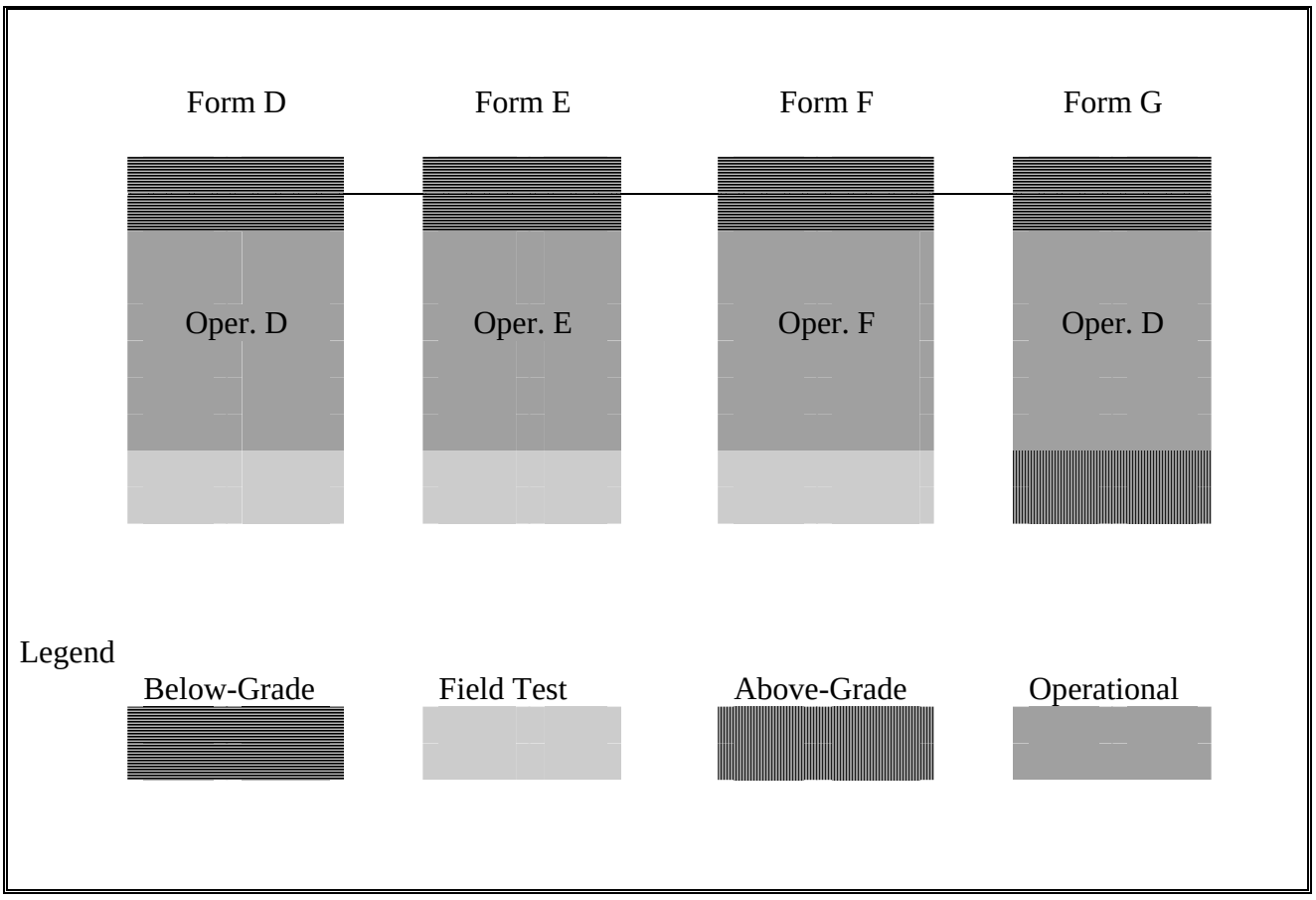
The forms to support development of the vertical scales were designed such that the linkage between any pair of grades did not rely solely on one grade being tested on out-of-level items. Thus, for each grade (Grade 3 and Grade 10 being exceptions), the base operational form was augmented with two collections of items. One collection was a group of operational items from the grade below and was designated as the below-grade anchor. The second collection was a group of operational items from the grade above and was designated as the above-grade anchor. Note that for any pair of grades, the total linking set was defined as the union of the set of items designated as the above-grade anchors in the lower grade and the set of items designated as the below-grade anchor in the higher grade.

Figure 1
Description of Vertical Linkages Between Grades



For Grades 4-8, Figure 2 depicts the collection of four forms that were used. Form G was designated as the vertical linking form, containing both below-grade and above-grade anchors. Note that the operational core items from Form D are repeated in Form G. Thus Forms D and G differ only in that Form D contains new items being field-tested and Form G instead has above-grade linking items. Also note that the below-grade anchor items define the horizontal link. For Grade 10, the vertical linking form is Form D. There was no need to make a special form because Grade 10 needed only to link to Grade 8. The situation for Grade 3 was similar to that for Grade 10. Form D was again designated as the vertical linking form; however, by virtue of being linked at the opposite end of the scale, the above-grade linking items were used to form the horizontal link.

Figure 2
Example of Horizontal Linkages within a Grade



Description of the Sample of Students Recruited for the Study

Because the new scales include grades that were not included in the Fall 2004 operational testing, schools needed to be specially recruited. The sampling frame for this study was developed using the 2002 Fall General Research Files (GRFs) for the WKCE Grades 4, 8, and 10 programs. Given the close temporal proximity of May 2004 and December 2004 data collections, the sample for the May 2004 Field Test and the sample for the December 2004 Forms Calibration / Vertical Scaling Studies were chosen simultaneously. Only public schools were considered for selection.

2002-03 Fall WKCE performance data at Grades 4, 8, and 10 (the only grades with such data available) were used to obtain a representative sample in terms of scale-score means and standard deviations. Initially, schools with fewer than 20 students were ineligible for the study, due to the instability of their mean scale scores associated with such small sample sizes. However, at the DPI's request, these small schools became eligible midway through the study, helping to ensure a more diverse and representative sample. School size was used as a stratifying variable, with three categories (small, medium, and large) defined so as to minimize standard deviations of scale scores within cells. Schools were randomly selected without replacement from each cell, proportional to the population proportion in that cell. Replacement schools were also chosen to account for the schools declining to participate. Throughout the process, three constraints were strictly enforced: 1) any given student could only take one form within each study (May or December); 2) a school could not be selected for more than four grades for the same study; and 3) no school of the same content/grade level could be drawn for both May and December. Before actual obtainment of the sample, CTB verified that the schools chosen were reasonably comparable to the population both in terms of achievement (Fall 2003 scale-score means and standard deviations) and ethnic breakdowns (proportions of Asian-American, African-American, Hispanic, American Indian/Alaskan Native, and White students).

The December study involved administering four forms in Grades 4-8 and three forms in Grade 3 and Grade 10. Forms were spiraled within classrooms to achieve maximally uniform samples among different forms. The target sample size was 2,300 per form for the December study. Table 1 and Table 2 display the total sample sizes and subgroup compositions for mathematics and reading, respectively. The values under each subgroup heading represent the proportion of the total group accounted for by the subgroup. Non-responses are not included in the table.

Obtained sample sizes were smaller than the stated target of 2300. However, all sample sizes were easily adequate for providing stable item parameter estimates. Additionally, the spiraling of forms within classrooms produced reasonably equivalent samples within each content/grade level combination. This is evidenced by the similarity of the subgroup composition of students delivered each form (see Table 3 and Table 4), and the similarity in student performance on the common set of out-of level items appended to each of the operational forms within a content/grade level (see Table 5 and Table 6).

Table 3
Total Group and Subgroup Frequencies by Grade and Form For Mathematics

Grade	Form	Total	Gender		Ethnicity				
			Female	Male	As-Am	Af-Am	Hispanic	AI-AN	White
03	D	1942	0.47	0.52	0.04	0.12	0.06	0.02	0.75
	E	1898	0.49	0.50	0.04	0.12	0.07	0.02	0.74
	F	1912	0.46	0.53	0.03	0.11	0.07	0.02	0.76
04	D	1636	0.47	0.53	0.03	0.13	0.08	0.01	0.75
	E	1646	0.50	0.50	0.04	0.14	0.06	0.01	0.74
	F	1816	0.50	0.50	0.04	0.11	0.07	0.01	0.76
	G	1836	0.49	0.51	0.04	0.11	0.06	0.01	0.76
05	D	2118	0.48	0.51	0.03	0.16	0.06	0.02	0.72
	E	2085	0.49	0.51	0.03	0.16	0.06	0.01	0.72
	F	1888	0.50	0.49	0.03	0.14	0.06	0.01	0.74
	G	1900	0.48	0.51	0.04	0.14	0.05	0.01	0.74
06	D	1803	0.50	0.50	0.05	0.11	0.07	0.02	0.74
	E	1794	0.47	0.52	0.05	0.11	0.08	0.03	0.73
	F	2091	0.49	0.50	0.05	0.13	0.07	0.03	0.72
	G	2099	0.48	0.51	0.05	0.11	0.08	0.03	0.73
07	D	1873	0.47	0.52	0.05	0.09	0.06	0.02	0.76
	E	1853	0.49	0.50	0.06	0.08	0.05	0.02	0.77
	F	2062	0.50	0.49	0.06	0.10	0.07	0.02	0.74
	G	2066	0.49	0.51	0.06	0.10	0.05	0.02	0.75
08	D	1629	0.48	0.51	0.05	0.09	0.06	0.01	0.76
	E	1609	0.52	0.48	0.06	0.10	0.07	0.01	0.75
	F	1452	0.46	0.53	0.05	0.09	0.06	0.01	0.77
	G	1437	0.47	0.52	0.07	0.07	0.06	0.01	0.76
10	D	1730	0.47	0.52	0.04	0.06	0.04	0.03	0.82
	E	1737	0.49	0.50	0.04	0.06	0.03	0.02	0.84
	F	1752	0.50	0.49	0.03	0.06	0.03	0.02	0.84

Table 4
Total Group and Subgroup Proportions by Grade and Form For Reading

Grade	Form	Total	Gender		Ethnicity				
			Female	Male	As-Am	Af-Am	Hispanic	AI-AN	White
03	D	1660	0.50	0.50	0.04	0.10	0.06	0.01	0.77
	E	1670	0.50	0.50	0.04	0.09	0.05	0.01	0.79
	F	1678	0.46	0.53	0.04	0.10	0.06	0.01	0.77
04	D	1811	0.47	0.53	0.04	0.14	0.08	0.01	0.72
	E	1794	0.50	0.50	0.04	0.15	0.08	0.01	0.71
	F	1556	0.52	0.48	0.03	0.13	0.08	0.01	0.74
	G	1546	0.47	0.53	0.05	0.12	0.07	0.01	0.75
05	D	1640	0.49	0.51	0.03	0.08	0.05	0.02	0.80
	E	1651	0.49	0.51	0.04	0.08	0.04	0.02	0.80
	F	1868	0.49	0.50	0.04	0.08	0.04	0.03	0.79
	G	1881	0.47	0.52	0.04	0.08	0.04	0.03	0.79
06	D	1954	0.49	0.51	0.03	0.07	0.05	0.01	0.82
	E	1927	0.48	0.51	0.02	0.08	0.05	0.01	0.81
	F	1651	0.48	0.51	0.02	0.08	0.05	0.01	0.82
	G	1654	0.49	0.51	0.02	0.08	0.06	0.01	0.82
07	D	1835	0.48	0.51	0.02	0.08	0.06	0.02	0.81
	E	1827	0.49	0.51	0.03	0.09	0.05	0.02	0.80
	F	1611	0.48	0.52	0.02	0.08	0.06	0.02	0.83
	G	1611	0.52	0.48	0.02	0.07	0.06	0.02	0.82
08	D	1827	0.48	0.52	0.03	0.12	0.06	0.01	0.78
	E	1828	0.48	0.52	0.03	0.11	0.07	0.01	0.78
	F	1499	0.49	0.51	0.03	0.11	0.04	0.02	0.80
	G	1492	0.50	0.50	0.03	0.11	0.06	0.02	0.79
10	D	1792	0.49	0.50	0.05	0.05	0.01	0.00	0.87
	E	1790	0.48	0.51	0.04	0.06	0.02	0.01	0.86
	F	1804	0.49	0.51	0.04	0.07	0.01	0.01	0.87

Table 5
Summary Statistics For Common Items
by Grade and Form for Mathematics

Grade	Form	N	Mean	SD
03 Anchor=15	D	1942	9.32	2.89
	E	1898	9.15	2.88
	F	1912	9.27	2.94
04 Anchor=15	D	1636	10.83	2.60
	E	1646	10.91	2.56
	F	1816	10.93	2.65
	G	1836	10.91	2.62
05 Anchor=19	D	2118	14.70	3.92
	E	2085	14.64	3.84
	F	1888	14.82	3.98
	G	1900	14.93	3.90
06 Anchor=15	D	1803	11.51	3.45
	E	1794	11.57	3.43
	F	2091	11.87	3.36
	G	2099	11.64	3.37
07 Anchor=19	D	1873	13.52	4.33
	E	1853	13.38	4.50
	F	2062	13.18	4.48
	G	2066	13.21	4.54
08 Anchor=18	D	1629	13.62	4.50
	E	1609	13.61	4.48
	F	1452	13.41	4.43
	G	1437	13.70	4.38
10 Anchor=19	D	1730	11.78	4.61
	E	1737	11.60	4.70
	F	1752	11.65	4.65

Table 6
Summary Statistics For Common Items
by Grade and Form for Reading

Grade	Form	N	Mean	SD
03 Anchor=16	D	1660	12.16	3.65
	E	1670	12.10	3.69
	F	1678	12.16	3.61
04 Anchor=16	D	1811	11.76	4.54
	E	1794	11.89	4.44
	F	1556	11.80	4.60
	G	1546	11.88	4.51
05 Anchor=20	D	1640	13.96	4.34
	E	1651	13.80	4.29
	F	1868	13.51	4.67
	G	1881	13.85	4.52
06 Anchor=20	D	1954	13.83	4.39
	E	1927	13.90	4.42
	F	1651	13.88	4.41
	G	1654	13.95	4.23
07 Anchor=20	D	1835	13.69	4.55
	E	1827	13.80	4.52
	F	1611	14.23	4.38
	G	1611	14.06	4.27
08 Anchor=20	D	1827	12.51	4.87
	E	1828	12.22	4.92
	F	1499	12.57	4.86
	G	1492	12.66	4.82
10 Anchor=21	D	1792	13.35	6.45
	E	1790	13.62	6.48
	F	1804	13.76	6.50

General Description of IRT Methods Used to Calibrate Items

Items were scaled and calibrated using item response theory (IRT) procedures similar to those followed in the development of the *TerraNova* (CTB/McGraw-Hill, 1997), *TerraNova* 2nd Edition (CTB/McGraw-Hill, 2000) and the Wisconsin Knowledge and Concept Exam (WKCE) (CTB/McGraw-Hill, 1997-2001).

CTB used the three-parameter logistic (3PL) model (Lord & Novick, 1968; Lord, 1980) to scale the MC items and the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) to scale the CR items. The 3PL model defines a MC item in terms of three item parameters: the item difficulty or location, the item discrimination, and the level of guessing. The 2PPC model defines a CR item in terms of an item discrimination and a location parameter for each score point. Introductory discussions of IRT can be found in Educational Measurement (Linn, 1989), or Chapter 11 in Introduction to Measurement Theory (Allen & Yen, 1979). More advanced discussions of partial credit models may be found in Muraki (1990, 1992), Yen (1993), and van der Linden and Hambleton (1997).

Item Response Theory Models

Because the characteristics of MC and CR items are different, two item response theory models were used in the analysis of the data. The 3PL model (Lord & Novick, 1968; Lord, 1980) was used in the analysis of MC items. In this model, the probability that a student with ability θ responds correctly to item *i* is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]},$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low-ability student. For analysis of the CR items, the 2PPC model (Muraki, 1992; Yen, 1993) was used. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score at the *k*-th level of the *j*-th item is

$$P_{jk}(\theta) = P(x_j = k-1|\theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \qquad k = 1,...m_j,$$

where

$Z_{jk} = A_{jk}\theta + C_{jk}$.

For the special case of the 2PPC model used here, the following constraints were used:

$A_{jk} = \alpha_j(k-1)$,

and

$C_{jk} = -\sum_{i=0}^{k-1}\gamma_{ji}$, where $\gamma_{j0} = 0$,

where α_j and γ_{ji} are parameters freely estimated from the data. The first constraint implies that higher item scores reflect higher ability levels and that items can vary in their

discriminations. The 2PPC model estimates a total of m_j independent item parameters; for each item there are m_j–1 independent γ_{ji} parameters and one α_j parameter.

Advantages of the 3PL and 2PPC Models

There are a number of advantages to using these two models. First, some students are bound to guess on MC items and the probability of this guessing behavior will vary among items. Students can be expected to guess on MC items that they find difficult. The guessing parameter of the three-parameter model accounts for this guessing behavior on the part of students. By contrast, the Rasch model (Rasch, 1960) assumes that no guessing occurs since there is no guessing parameter present in the model. This is clearly unrealistic.

A second advantage is that item discrimination is allowed to vary among items for the 3PL and the 2PPC models. A single value is used as the discrimination parameter for the Rasch and the one-parameter partial credit model (Masters, 1982). There are several important reasons for allowing discrimination to vary. First, when the Rasch or one-parameter partial credit models are used, all the items by definition must have the same discrimination. As a result, items that do not fit the model must be discarded. The 3PL and 2PPC models place minimal constraints upon the items that can be selected for tests. This allows content editors to select the best items available with minimal constraints on the content.

Another advantage of having items with varying discriminations is that the MC and CR items can be scaled together. MC and CR items will necessarily have varying discriminations. With the Rasch and the 1PPC, all discriminations must be equal. Therefore, the use of these models precludes their being scaled together.

Moreover, a sufficient statistic for ability using the Rasch model is the number correct score. All students who get the same number correct score are assigned the same ability estimate, regardless of the particular items that were answered correctly. Using the 3PL or 2PPC models, item-pattern scores that yield more accurate estimates of ability can be generated. Item-pattern scoring takes advantage of all the information contained in an array of student responses. As teachers typically do when evaluating classroom exams, IRT pattern scoring gives different weights to items according to how each item contributes to the measurement of the subject being tested.

Parameter Estimation

The IRT models were implemented using CTB's PARDUX software (Burket, 1991). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented with the EM algorithm (Bock & Aitkin, 1981; Thissen, 1982). PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. Extensive simulation studies and comparisons between PARDUX and MULTILOG (Thissen, 1990), a program widely used for research purposes, have shown that PARDUX provides precise parameter and ability estimates, and it performs more efficiently than MULTILOG (Fitzpatrick, 1991). Simulation studies have also compared PARDUX with PARSCALE (Muraki & Bock, 1991) and with BIGSTEPS

(Wright & Linacre, 1992). Fitzpatrick and Julian (1996) found that PARDUX provided precise parameter and ability estimates, and performed more efficiently than the other programs. Extensive research with simulation data has also shown that the IRT procedures used here produce accurate vertical scaling (Yen & Burket, 1997). The Stocking and Lord (1983) procedure was used to link parameters from independent calibrations through performance on a set of common items.

The recent versions of PARDUX have the added capability of estimating true score distributions for multiple groups and using them in the parameter estimation process. In the multiple–group case, PARDUX fixes the mean and SD for one of the groups equal to 0 and 1, respectively, and estimates the mean and SD for all other groups. The current WKCE-CRT analyses used one of the multiple-group versions of PARDUX.

Fit Rating

A procedure and a statistic (Q1) described by Yen (1981) were used to measure model-to-data fit. The procedure divides the sample into ten groups according to their ability estimates. For a given item, it then compares the expected and observed proportion of the students for each of the ten ability groups and computes an index (Q1) pooled over the ability groups. The Q1 is distributed approximately as a chi-square with the following degrees of freedom:

$$df = I(m_j - 1) - m_j ,$$

where I is the total number of cells (usually 10) and m_j is the possible number of score levels for item j. To adjust for differences in degrees of freedom among items, Q_1 was transformed to Z_{Q_1} where

$$Z_{Q_1} = (Q_1 - df) / (2\,df)^{1/2} .$$

Based on previous research findings, the Z critical value has been established as a function of sample size (N). Items with $Z_{Q_1} \geq \frac{4N}{1500}$ were given a fit rating of *poor*; other items had a fit rating of *acceptable.*

Description of the Calibration/Scaling Methods Used

Forming a vertical scale spanning Grades 3-8 and Grade 10 is a balancing act between shifts in construct, dimensionality, and the strength of pairwise links. The best calibration results are typically obtained when items delivered to students in a single grade are calibrated separately. This tends to minimize dimensionality concerns within each calibration. However, chaining together a series of individual calibrations does not guarantee a common meaning for the underlying scale that is produced. It is generally assumed that the greatest similarity of the underlying construct is provided when a single concurrent calibration is used. However, this method ignores the fact that the nature of a content area tends to change across grades, and thus a single dimension might be inappropriate.

During the last TAC meeting, several options for balancing these two issues were discussed. There was some concern that the inclusion of Grade 10 would increase the stress on a concurrent calibration. There was also a suggestion that rather than considering the calibrations as having only two options (concurrent, linked single-grade), an option be considered that formed several sets of grades. In light of this discussion, four different vertical scaling methods were performed. They are ordered from the least number of scaling links required to the largest number required.

Vertical Scaling Method 1

A concurrent calibration for all grade levels (Grades 3-8 & Grade 10) was performed. Concurrent estimation establishes the common scale in a single step–the calibration phase– by simultaneously estimating parameters for all items at all grades. The concurrent calibration assumes the presence of a single trait across grades, and concerns have been raised about the assumption, particularly when the scale spans many grades (e.g., Camilli, 1999; Haertel, 1991; Lissitz et al., 2003). Additionally, some research has shown that concurrent estimation runs can sometimes have convergence problems (Kolen & Brennan, 2004).

Vertical Scaling Method 2

In the case of WKCE-CRT, Grade 10 may potentially be very different from the remaining grades. Thus a combination of concurrent calibration of Grades 3-8 with separate calibration of Grade 10 was used. The Grade 10 calibration was then linked to the Grade 3-8 scale.

Vertical Scaling Method 3

Three collections of grades were formed. Specifically, the base scale was formed using a concurrent calibration of Grades 5-7. This calibration was followed by calibrating Grades 3-4 in one concurrent run and then Grades 8-10 in a separate concurrent run. These two concurrent runs were then linked to the scale formed by the Grades 5-7 calibration. This method is thought to provide a reasonably common construct within each calibration run while still minimizing the accumulation of linking errors.

Vertical Scaling Method 4

Each grade was calibrated separately, and then the grades were chained together. The Grade 6 calibration was used to define the base scale. Then the Grade 5 and Grade 7 calibrations were linked to it. This was followed by linking the Grade 4 calibration to the scale through its link with Grade 5, and linking the Grade 8 calibration to the scale through its link with Grade 7. The calibrations for Grade 3 and Grade 10 were similarly linked. This method is thought to put the least strain on the calibration assumptions but introduces the highest probability of the accumulation of linking errors (Kolen & Brennan, 2004).

Horizontal Scaling/Form Equating

Although not reported on herein at this time, once the vertical scaling was completed, the remaining forms/items were placed on the IRT scale using the following two-step process. First, a separate concurrent calibration was performed for each content area/grade level combination that included all items delivered in Forms D, E, and F. Recall from Figure 2 that this calibration was facilitated by including the same set of out-of-level items in each of the three forms. Step 2 was simply to perform a Stocking/Lord linking of these content/grade level calibrations to the vertical scale. This linking was facilitated by inclusion of Form D in the calibration. For Grade 3 and Grade 10, Form D was the vertical scaling form and thus formed a natural link to the scale. For Grades 4-8, recall that the operational items in the vertical scaling form (Form G) were also the operational items in Form D. Thus, again, a strong link was available when performing the Stocking/Lord linking.

Now that all items are on the same scale, a powerful tool is available for future forms to be assembled/scaled. For any future form assembled from this pool, the scaled item parameters have already placed the form on scale. Conceptually, the form is on scale regardless of the absolute difficulty of the set of items chosen. However, in practice, it is prudent to maintain similar levels of difficulty and score precision across forms

administered to students in the same grade. This can be achieved without needing to administer the form in advance of its operational use by simply including statistical characteristics in the specifications for each form (e.g. the test characteristic curve (TCC), information functions, or standard error curves).

As new items are required to assemble future forms, the new items can simply be included as Field Test items at each administration, and placed on scale in the same manner as described for Forms D, E, and F above. That is, for each content/grade level combination, perform a concurrent calibration of the operational items (which are already on the appropriate IRT scale) and the Field Test items. Then perform a Stocking/Lord linking using the operational items as the horizontal link.

Description of Vertical Linking Set Performance

Table 5 and Table 6 below summarize performance on the anchor and operational items for each linkage for mathematics and reading, respectively. If the type of grade-to-grade growth that is assumed to underlie a vertical scale is present, then for each pair of grades it should be observed that the students in the higher grade produce a larger mean score on the linking set than do students in the lower grade. It is also desirable for the correlation between the linking score and the operational score to be similar in the two grades. A similarity in the correlations is evidence that the constructs being assessed for the two groups of items are sufficiently similar for the items to form a reasonable link.

For the mathematics tests, both features are present. Students in the higher grade always produced a higher mean score, and the correlations between the linking score and the operational score were similar. However, the separation between Grade 7 and Grade 8 was not as large as for the other grade pairs.

For reading, there are two exceptions. First, for the link between Grade 7 and Grade 8, the students in Grade 7 produced a slightly higher mean score on the linking items. Second, for the link between Grade 8 and Grade 10, while the means on the linking test followed the expected pattern, the correlation between the linking score and the operational score for Grade 10 was lower than that observed for Grade 8. Moreover, it was the lowest correlation between a linking and operational score observed across the two tables.

Figure 3-Figure 14 expand the across grade comparison to the item level. In each figure, the proportion correct for the lower grade defines the position on the x-axis and the proportion correct for the higher grade defines the position on the y-axis. If growth were uniform for every item, then it would be expected that all item icons in the figures would fall above the 45° line. That is, it is expected that students in the higher grade answer each item correctly more frequently than do students in the lower grade. The item icons distinguish items based on two item features: the response format for the items (MC or CR) and whether the item is on-level or out-of-level for each of the two grades characterized per figure. The two options here are above- grade/operational and operational/below-grade.

A quick visual inspection indicates that while the expected pattern was observed for many of the linking sets, there were also some exceptions. For example, for mathematics G04/G05, at least six items were markedly easier for the Grade 4 students than for the Grade 5 students. Had all of these items been Grade 4 operational items that were out-of-level for Grade 5, there might have been some concern that these items were sensitive to recency of instruction. However, this was not the case. There were as many Grade 5 operational items given out-of-level to Grade 4 students that exhibited the same behavior. It would also have been a concern had all of these been CR items. That could have raised some concerns about the scoring rubric begin tied to some specific set of instruction. However, this does not appear to be the case either. There were just as many MC items exhibiting this behavior as CR items.

Interestingly, these large differences were only exhibited for mathematics items. Although some reading linking items were easier for students in the lower grade, the differences were quite small. The overall linking set mean difference much more closely described difference observed at the item level for reading than for mathematics. Whether this is an indication that the reading construct assessed by the WKCE-CRT is more developmental in nature than the mathematics construct assessed, or is simply an artifact of the specific items chosen to form the linking sets is a matter for further discussion. For the purposes of constructing a vertical scale for the two content areas, this phenomenon is likely an indication that mathematics will introduce some challenges that reading does not. As will be seen in the next section, this was, in fact, the case.

Table 7
Summary of Vertical Linking Set Performance – Mathematics

	Lower Grade			Upper Grade		
	Mean	Std Dev	Corr	Mean	Std Dev	Corr
03 / 04	19.77	5.77	0.85	23.00	5.29	0.89
04 / 05	24.46	7.12	0.92	27.60	6.83	0.89
05 / 06	21.04	7.39	0.90	24.18	7.62	0.93
06 / 07	22.61	8.61	0.91	25.67	8.37	0.92
07 / 08	22.21	7.26	0.92	23.81	7.62	0.90
08 / 10	17.61	8.46	0.90	21.45	8.78	0.92

Note: The correlation columns represent the correlation between the anchor item score and the total operational score.

Table 8
Summary of Vertical Linking Set Performance – Reading

	Lower Grade			Upper Grade		
	Mean	Std Dev	Corr	Mean	Std Dev	Corr
03 / 04	21.83	7.02	0.91	23.32	7.95	0.89
04 / 05	22.54	7.83	0.91	25.21	7.51	0.91
05 / 06	26.34	8.50	0.91	28.83	8.41	0.90
06 / 07	26.07	8.42	0.88	27.98	8.39	0.89
07 / 08	26.97	8.00	0.87	26.83	8.98	0.89
08 / 10	22.78	8.35	0.90	24.40	9.15	0.76

Note: The correlation columns represent the correlation between the anchor item score and the total operational score.

Figure 3

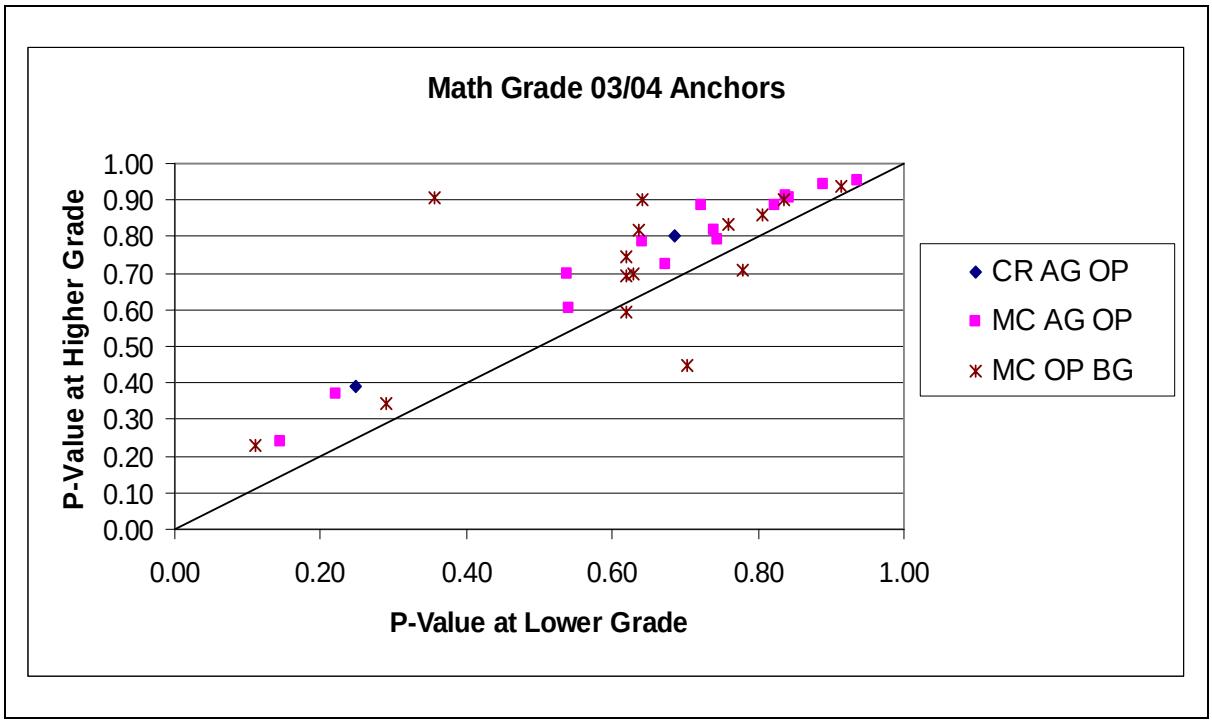


Figure 4

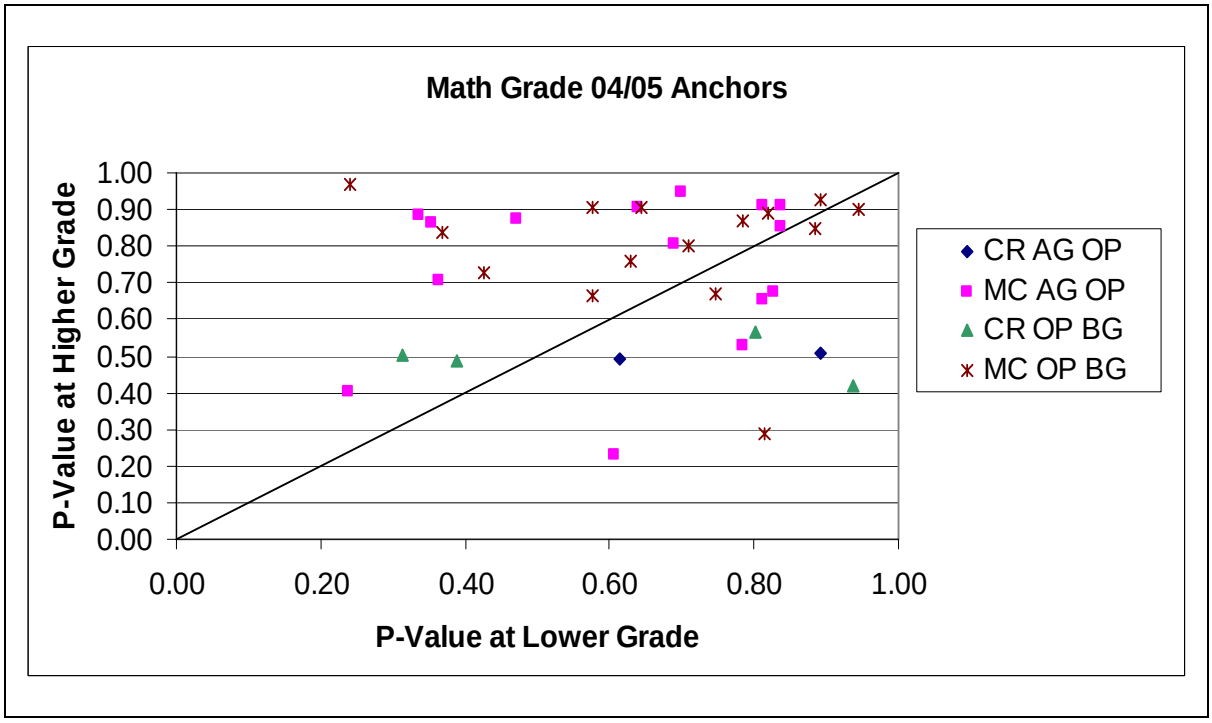


Figure 5

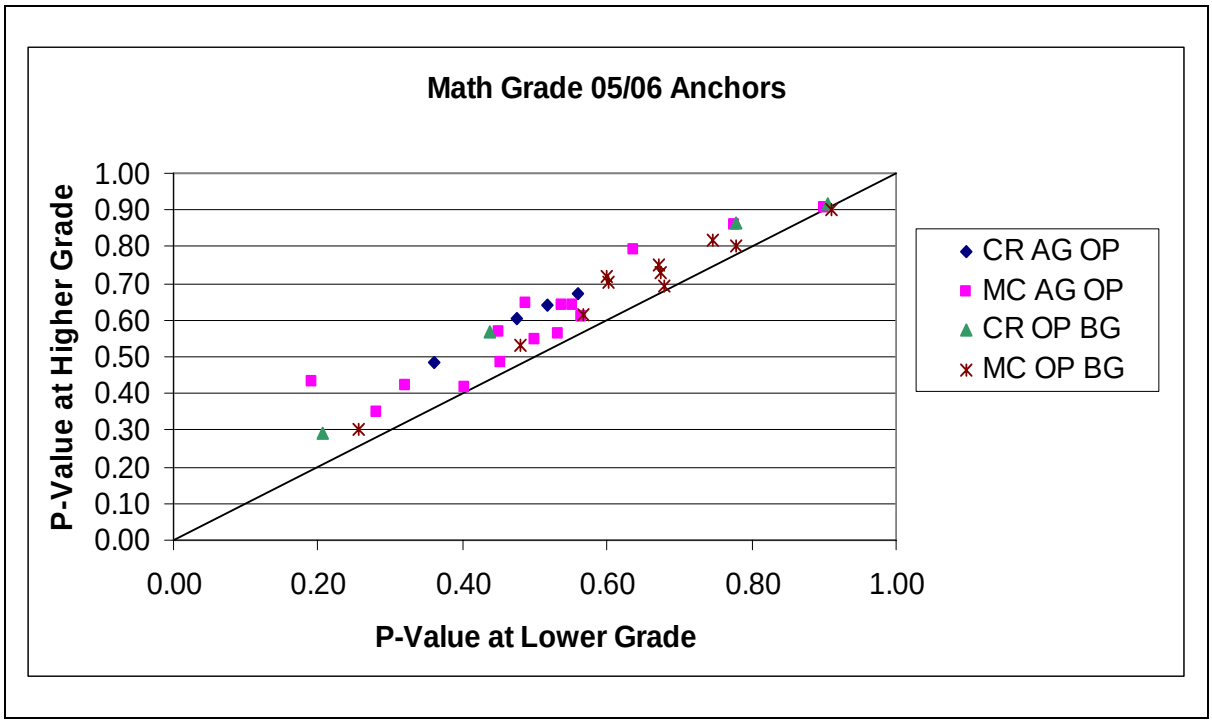


Figure 6

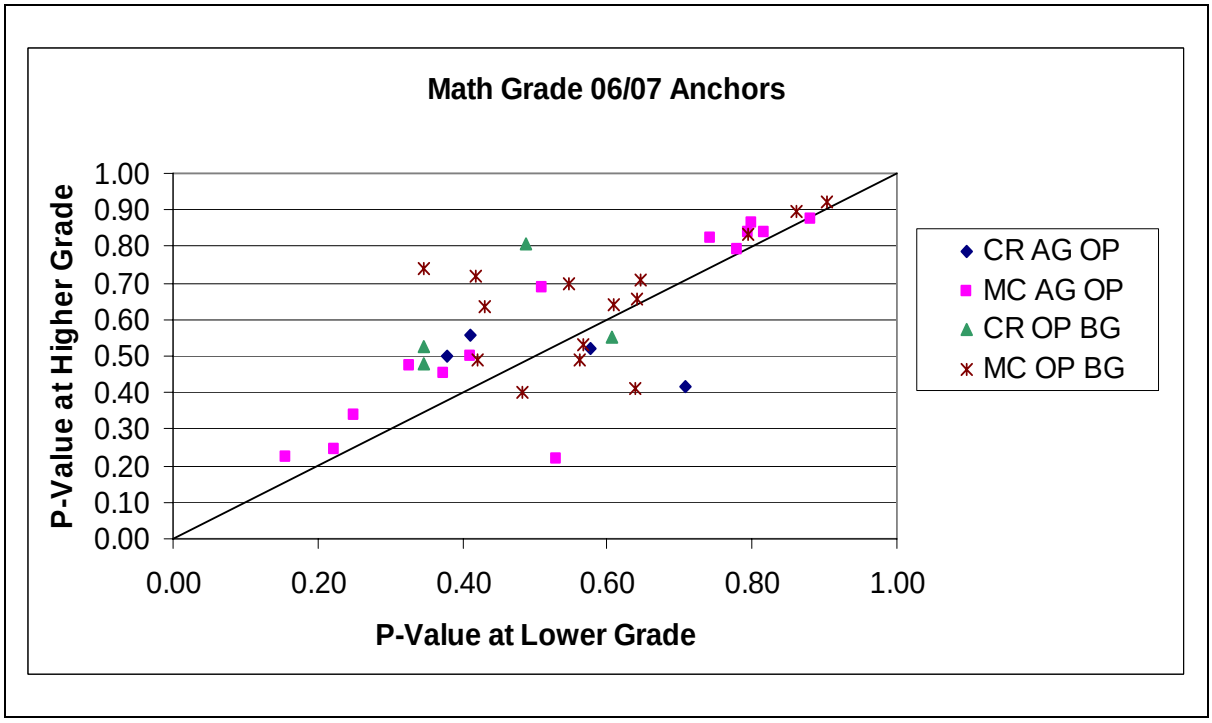


Figure 7

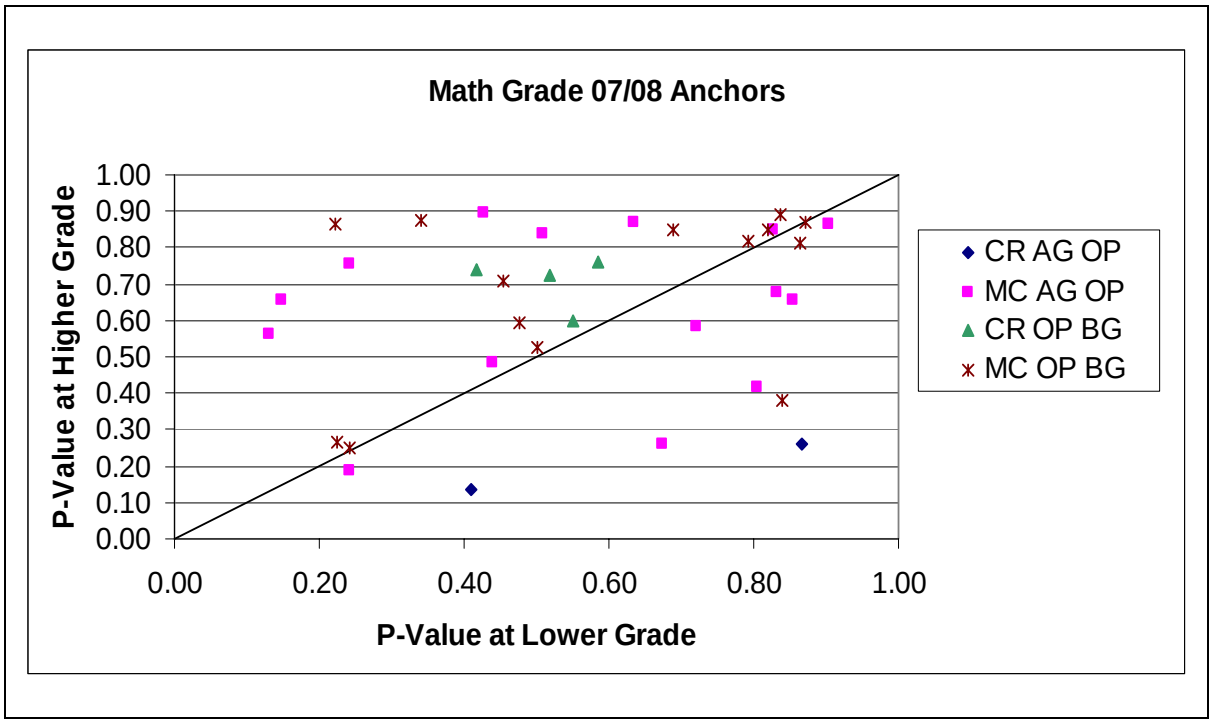


Figure 8

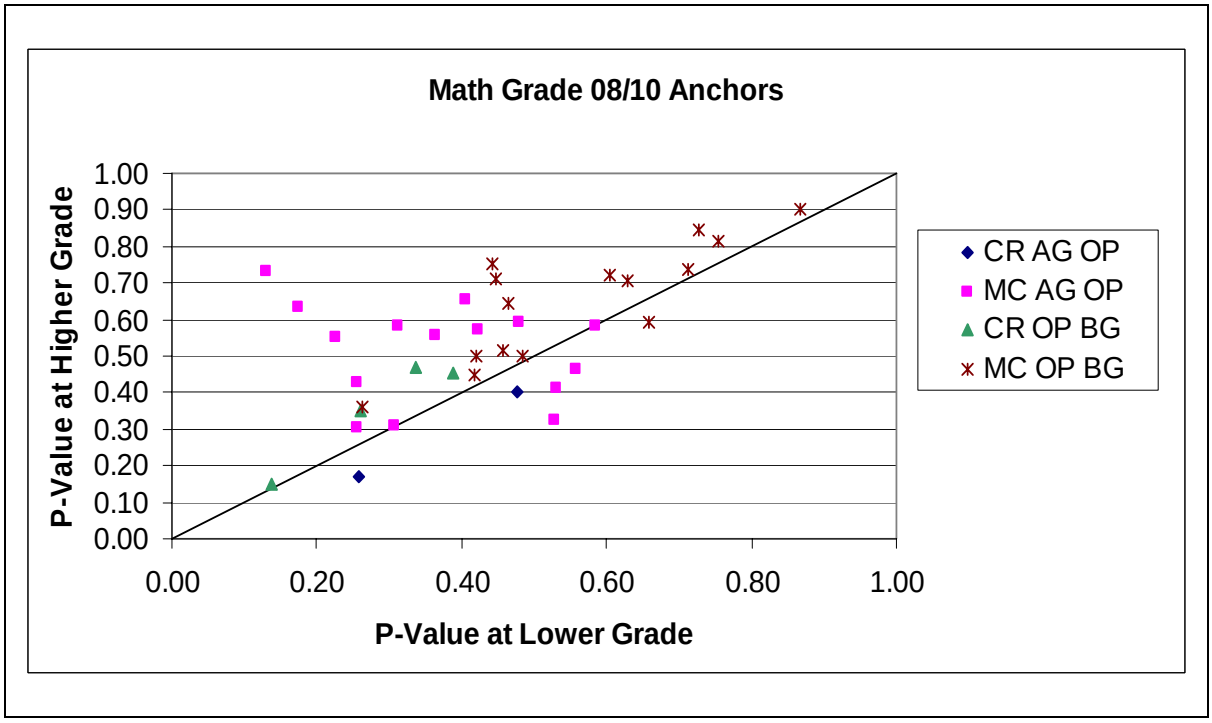


Figure 9

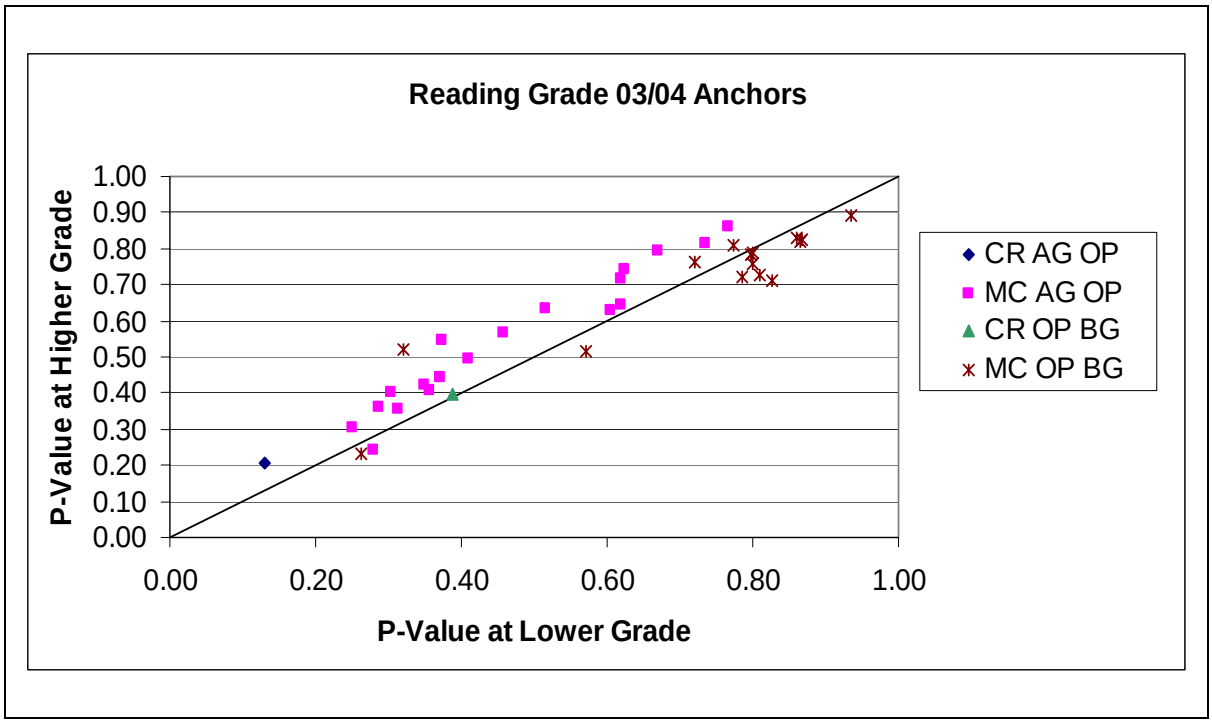


Figure 10

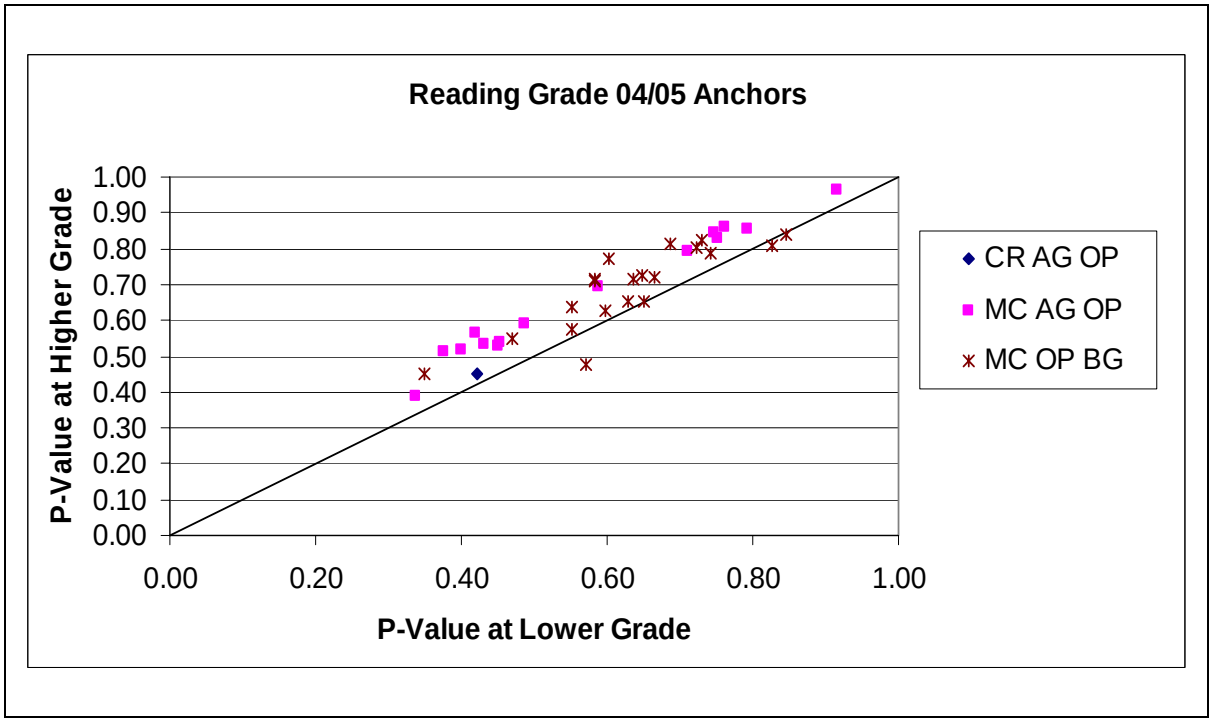


Figure 11

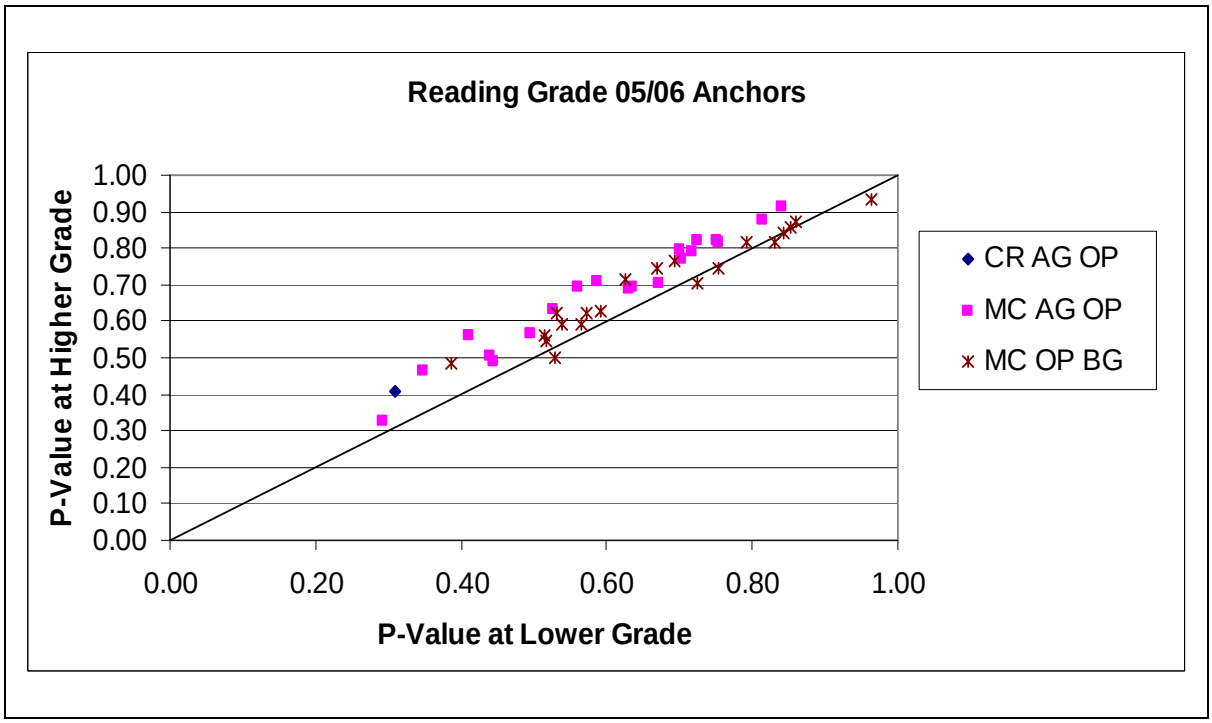


Figure 12

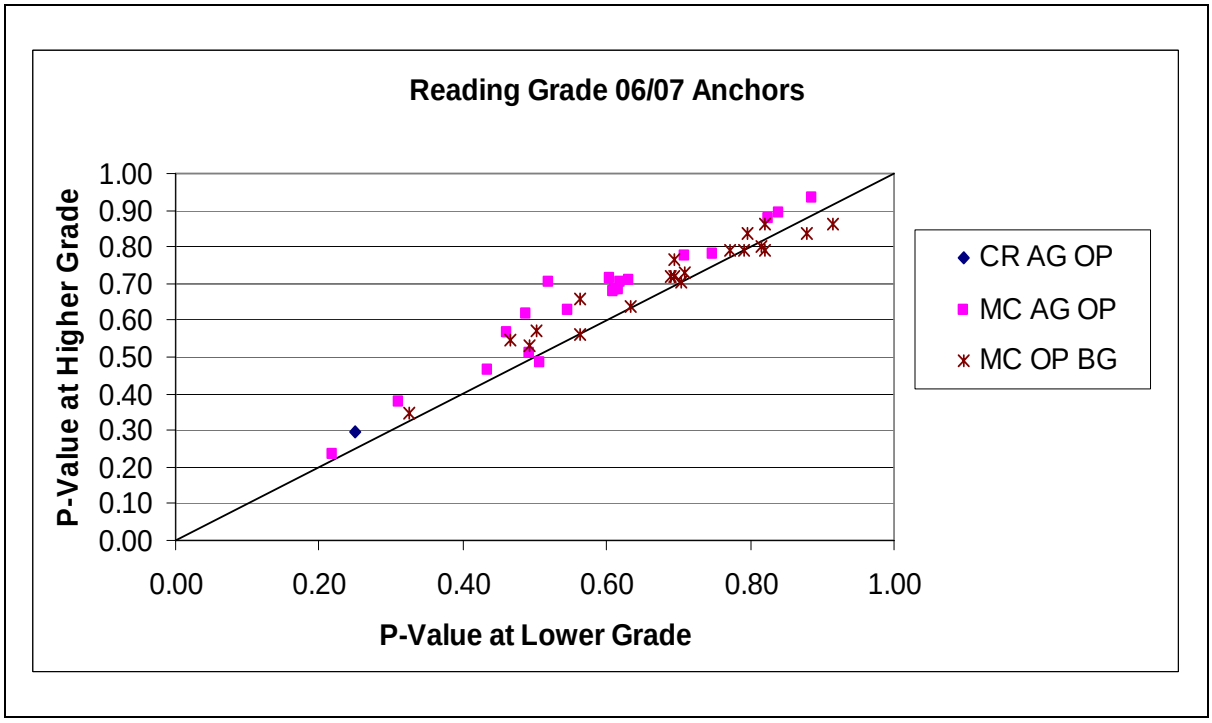


Figure 13

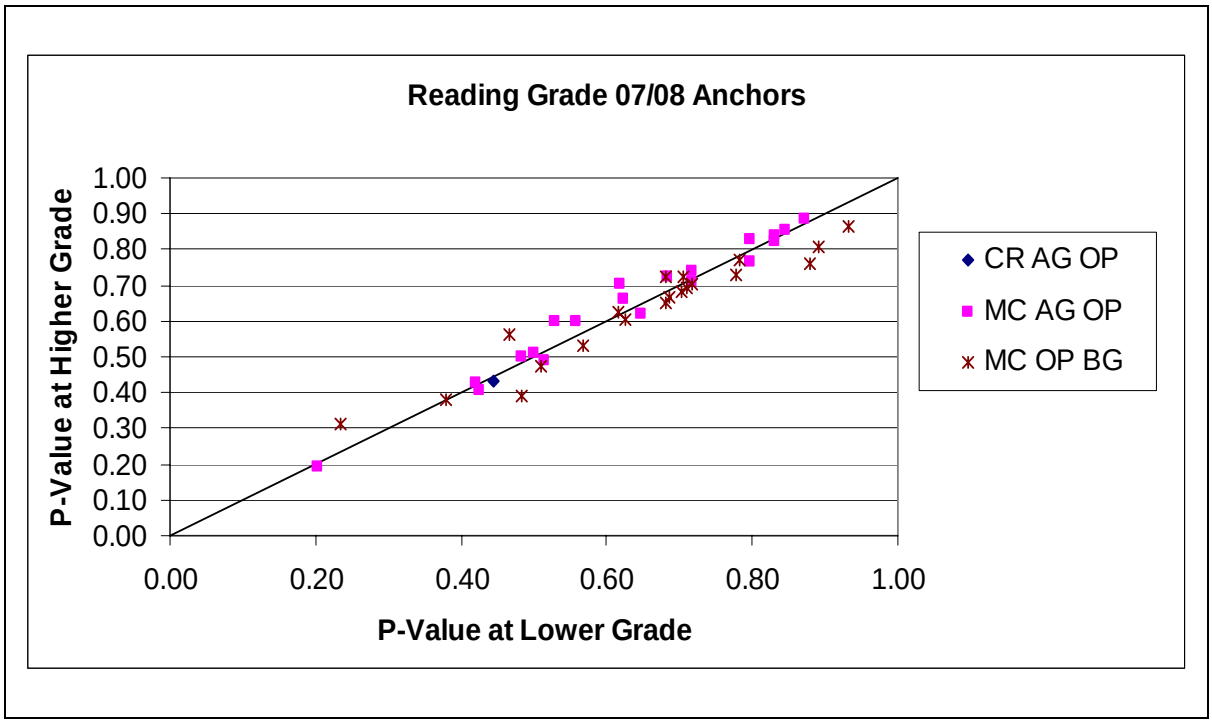
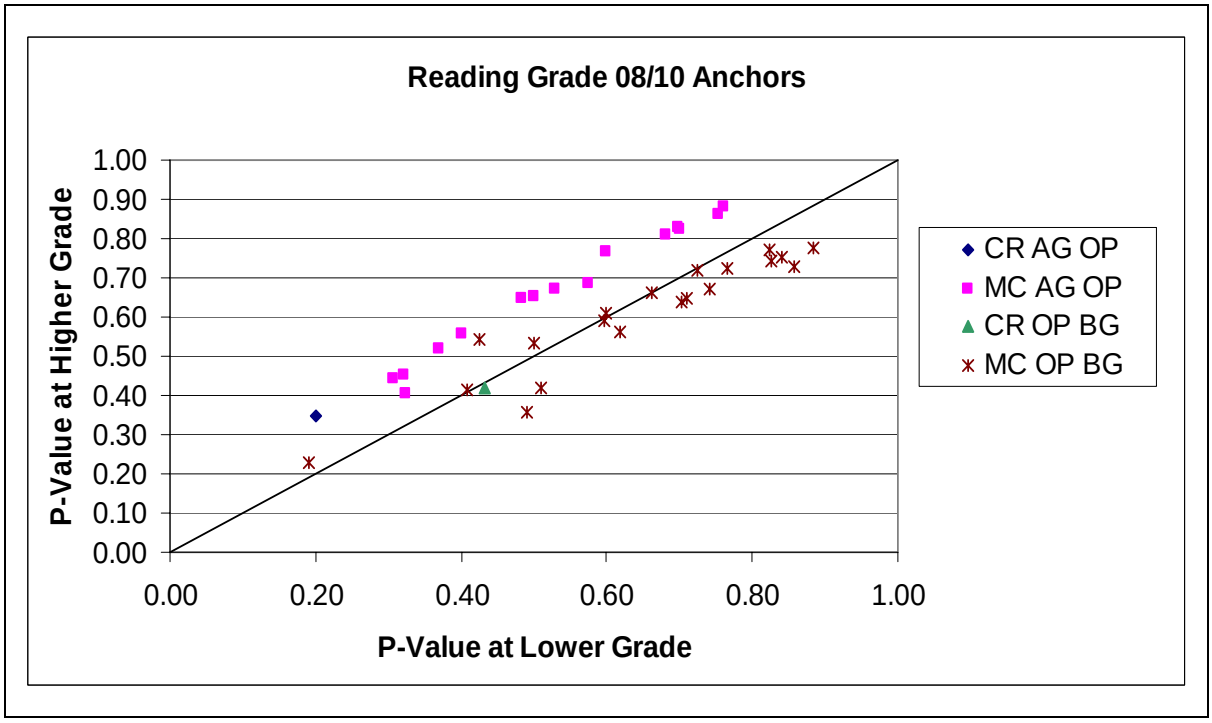


Figure 14



Summary of Results of Alternative Vertical Scaling Options

In this section, the outcomes of the four vertical scaling models will be presented and compared. The comparison will include three components: the number of items with convergence or fit issues, the expected test characteristic curves across grade levels, and the expected score distributions.

Table 9 and Table 10 display raw score summary statistics for mathematics and reading, respectively. The statistics reported are for the operational portions of the vertical scaling forms only. Summary statistics for the above-grade and below-grade anchor items were presented in Table 7 and Table 8

Table 9
Raw Score Descriptive Statistics for Mathematics Vertical Scaling Forms

Grade	N	Items	Mean	SD	SEM	Min	Max	Mean p-value	Alpha	Mean Item Total Corr.
03	1942	56	39.6	11.0	3.0	0	60	0.71	0.92	0.43
04	1836	61	44.6	12.4	3.3	0	68	0.72	0.93	0.43
05	1900	69	45.9	13.5	3.4	0	75	0.67	0.92	0.40
06	2099	69	39.4	13.6	3.4	0	73	0.57	0.92	0.41
07	2066	68	43.0	14.3	3.5	0	76	0.62	0.93	0.43
08	1437	66	33.8	14.6	3.6	0	72	0.51	0.94	0.44
10	1730	61	32.0	14.1	3.5	0	68	0.53	0.93	0.44

Table 10
Raw Score Descriptive Statistics for Reading Vertical Scaling Forms

Grade	N	Items	Mean	SD	SEM	Min	Max	Mean p-value	Alpha	Mean Item Total Corr.
03	1660	62	41.8	12.7	3.3	0	64	0.68	0.93	0.45
04	1546	67	40.6	13.2	3.4	0	68	0.61	0.93	0.42
05	1881	63	45.1	12.4	3.3	0	66	0.72	0.93	0.43
06	1654	63	43.6	11.8	3.1	0	67	0.69	0.91	0.40
07	1611	63	42.4	12.3	3.2	0	66	0.67	0.92	0.41
08	1492	63	38.9	13.0	3.3	0	66	0.62	0.93	0.43
10	1792	59	37.8	13.8	3.5	0	66	0.64	0.93	0.45

During the process of calibration, items that would not converge or that showed exceedingly high levels of misfit were removed in order to provide better estimates for the

remaining items. In several instances, this resulted in a significant loss of items. Once the calibrations and Stocking/Lord linkings were completed, all of the item parameter and ability estimates were in the θ metric. The θ values were transformed to the reporting scale by means of a linear transformation. That transformation was defined by computing linear scaling constants (M1, M2) such that the expected mean and standard deviation for Grade 6 were 500 and 50, respectively.

Table 11 and Table 12 below summarize some of the vertical scaling outcomes. Results are reported for each of the four methods described above. Included in the table are (1) the number of items surviving calibration, (2) the mean scale score, and (3) the standard deviation of the scale scores. The column labeled 'Base' under the block labeled 'Items Retained' defines the number of operational items at the time of assembly. The column labels of 01-04 refer to the vertical scaling method used.

Table 11
Summary of Vertical Scaling Outcomes for Mathematics

		Items Retained					Scale Score Mean				Scale Score S.D.			
Grade	N	Base	01	02	03	04	01	02	03	04	01	02	03	04
03	1942	56	56	56	56	56	431.2	430.9	427.3	427.1	47.1	47.1	50.1	50.8
04	1836	61	61	61	61	61	459.0	459.4	458.8	457.7	48.0	47.8	51.1	52.4
05	1900	69	67	68	68	69	481.0	480.9	480.4	480.0	46.9	46.6	48.1	49.7
06	2099	69	67	67	67	67	500.0	500.0	499.8	499.7	49.8	49.7	49.9	51.4
07	2066	68	67	62	62	68	517.4	518.2	517.7	518.4	47.5	48.3	48.1	51.7
08	1437	66	60	58	66	66	527.6	528.2	529.6	530.0	51.4	52.6	54.3	54.4
10	1730	61	57	61	60	61	552.2	549.7	552.7	552.7	50.4	47.3	49.4	49.9

Table 12
Summary of Vertical Scaling Outcomes for Reading

		Items Retained					Scale Score Mean				Scale Score S.D.			
Grade	N	Base	01	02	03	04	01	02	03	04	01	02	03	04
03	1660	62	62	62	62	62	452.6	451.0	458.2	454.2	40.8	41.6	37.8	46.2
04	1546	67	67	67	67	67	463.8	463.5	470.0	466.1	51.7	50.4	47.0	52.4
05	1881	63	63	63	63	63	484.1	482.4	483.7	485.0	50.3	50.0	48.7	51.2
06	1654	63	63	62	62	63	499.9	499.7	499.8	499.3	50.1	50.5	50.2	52.5
07	1611	63	63	62	62	62	510.3	511.0	510.5	512.6	49.0	48.3	48.0	51.3
08	1492	63	63	63	63	63	510.4	512.2	511.9	512.8	57.0	55.3	55.8	58.0
10	1792	59	59	59	59	59	514.7	523.4	519.8	524.8	70.4	63.6	67.1	66.9

Several general results are worth highlighting. First, in terms of the score resulting from the different vertical scaling methods, the results are all very similar. With only minor exceptions, for all four methods, the means and standard deviations progress in the manner generally deemed desirable when constructing a vertical scale. That is, both the mean scores and their associated standard deviations tend to increase as grade level increases. The distinctions between the methods are most noticeable in the numbers of items that failed to be adequately estimated. While there were almost no calibration issues for the reading tests (within a single grade, at most one item was lost for lack of convergence), the situation was entirely different or the mathematics tests. The four methods encountered dramatically different issues.

- For Method 1, the concurrent calibration of all items from Grades 3-8 & Grade 10, a large number of items demonstrated convergence and model misfit issues. Items for Grade 8 and Grade 10 presented the most difficulties. Six items from the Grade 8 form and four items from the Grade 10 form could not be adequately estimated. These were not the only calibration issues. Several items in each of Grades 5-7 also could not be adequately estimated.
- For Method 2, which consisted of a concurrent calibration of all items from Grades 3-8 followed by an independent calibration of the items from Grade 10 that was then linked to the Grades 3-8 calibration, similar issues were encountered. Interestingly, it was once again the top two grades of the concurrent calibration run that produced a large number of items that could not be adequately estimated. This

was true for six items from Grade 7 and eight items from Grade 8. One item from Grade 5, and two from Grade 6, also could not be adequately estimated.

- For Method 3, separate calibrations for Grades 3 and 4, Grades 5-7, and Grades 8 and 10, almost no calibration issues were encountered. Only one item from each of Grade 6 and Grade 7 could not be adequately estimated.
- For Method 4, separate calibration for each grade, the results were similar to that for Method 3. Only a single item from Grade 7 could not be adequately estimated.

In summary, the issues encountered for mathematics closely reflect the concerns that led to the four calibration methods being used. Items in the higher grades were difficult to calibrate when the calibration included items from the lower grades. However, when the Grade 10 items were calibrated separately or only in conjunction with Grade 8 items, convergence and fit essentially became a non-issue.

The greatest concern about the items lost for Grade 7, and even to some degree for Grade 8, is that the majority of the items with convergence and fit issues were the constructed response items. It is interesting to note that the issues essentially went away when either Grade 10 items were included in the calibration (Method 1) or when Grade 7 items were calibrated in isolation (Method 4).

For mathematics, figures depicting the resulting test characteristic curves and the cumulative frequency distributions for each method are included in Figures 15-18 and 19-22, respectively. For reading, similar figures are included in Figures 23-30. These figures reinforce what can be seen in Table 9 and Table 10. In general, as the grade level increases, so does the scale score mean. The lone exception is that for Method 1 the mean Grade 7 reading score is higher than the mean Grade 8 reading score. However, this is consistent with what might be expected based on performance on the vertical linking items. As was noted in Table 6, the students in Grade 7 outperformed the students in Grade 8.

Final Scale Selection

Taken together, the results indicate that, at least for Mathematics, there is a tension between calibration dimensionality and accumulated linking errors. Including the higher grades in a concurrent calibration caused a non-trivial number of items to exhibit convergence and misfit problems. In terms of the ability to produce stable estimates of items parameters, Method 3 provides reasonable results without introducing the need to link each pair of grades separately. Moreover, in the final evaluation the expected scale score results are very similar. Thus Method 3 was selected to define the new vertical scales for both Mathematics and Reading.

Figure 15
Test Characteristic Curves for Method 1 Mathematics

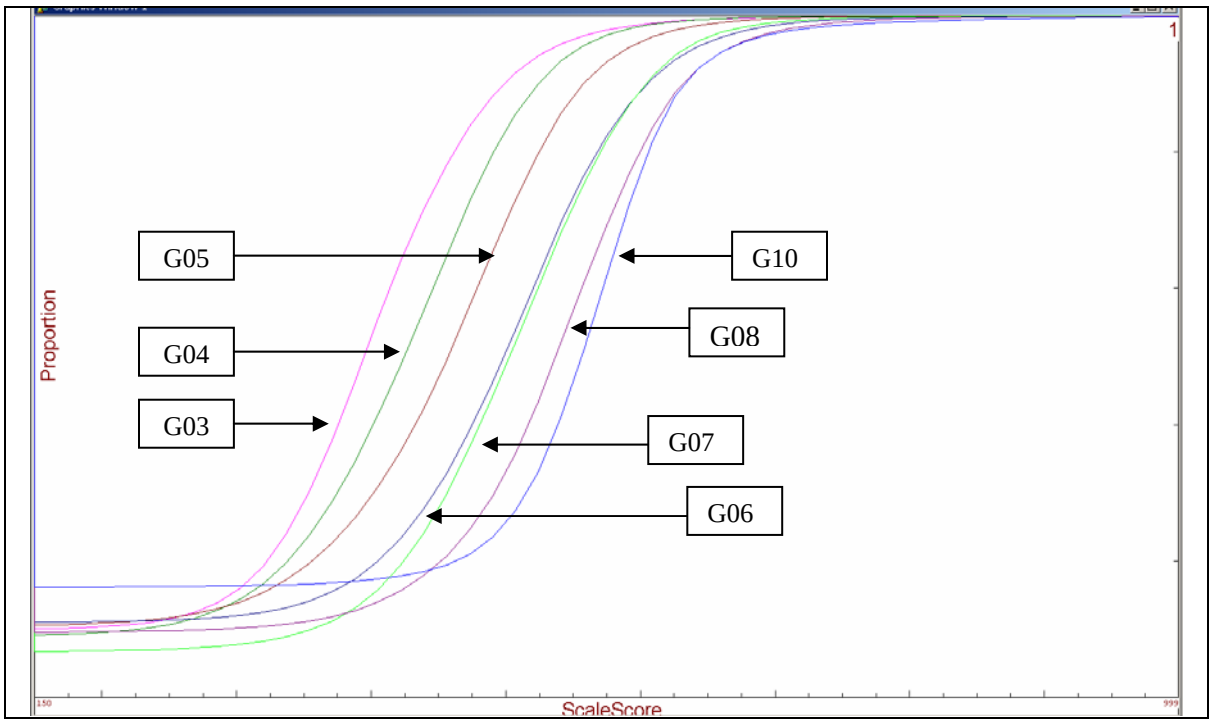


Figure 16
Test Characteristic Curves for Method 2 Mathematics

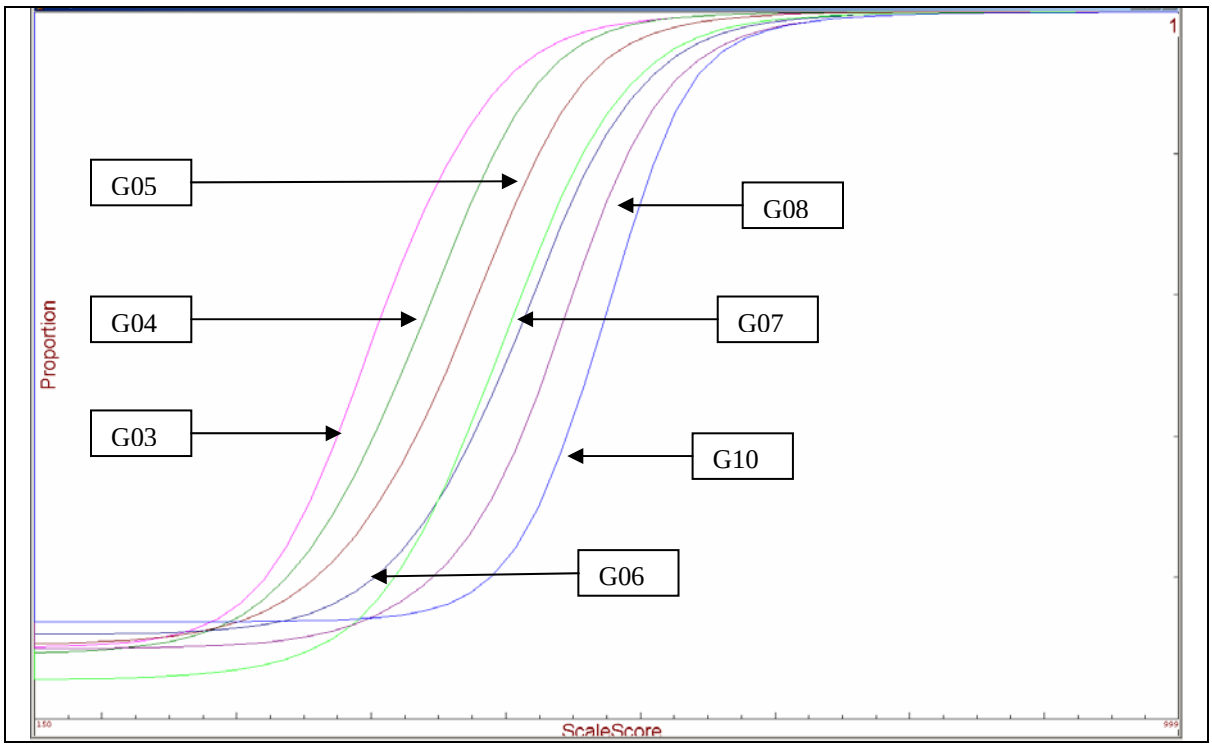


Figure 17
Test Characteristic Curves for Method 3 Mathematics

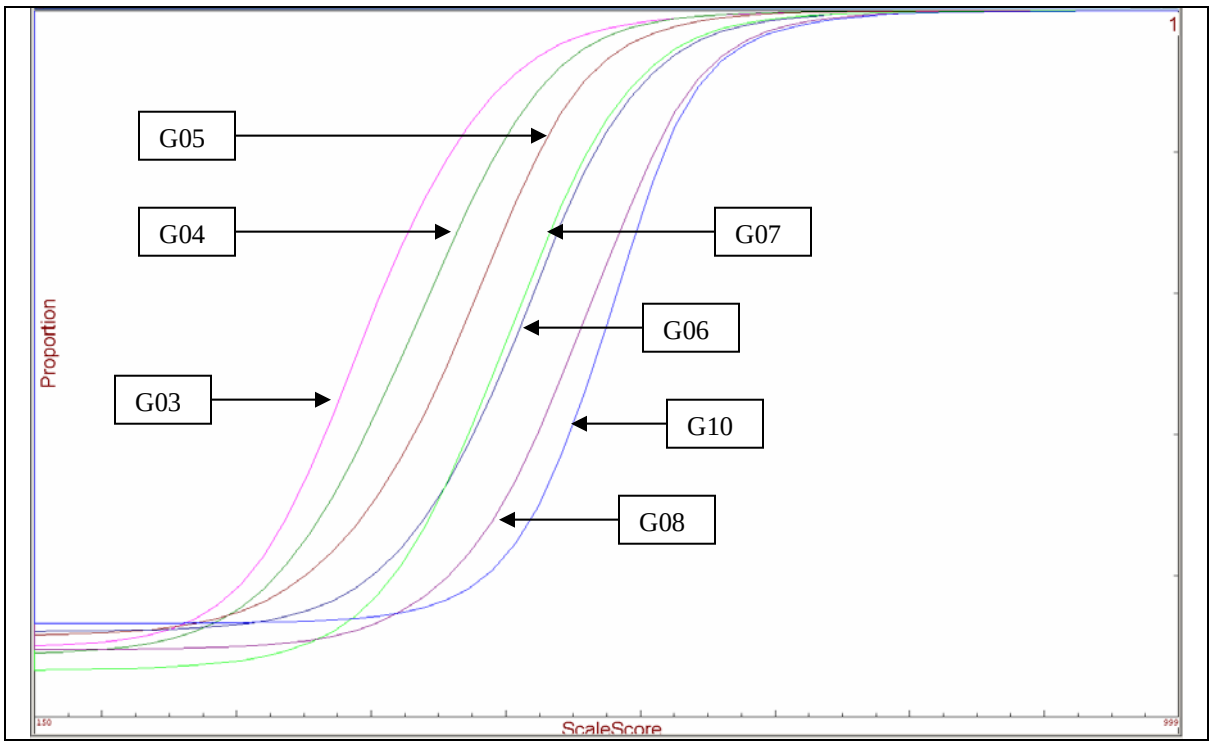


Figure 18
Test Characteristic Curves for Method 4 Mathematics

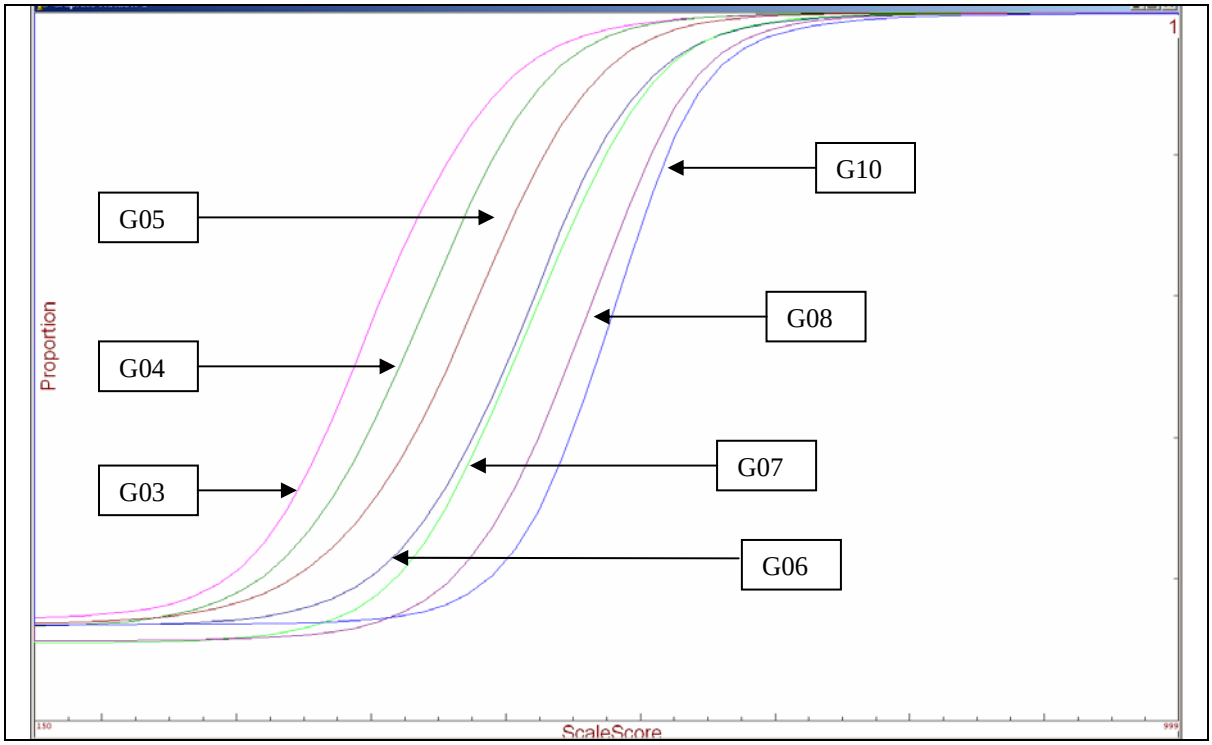


Figure 19
Cumulative Distribution Functions for Method 1 Mathematics

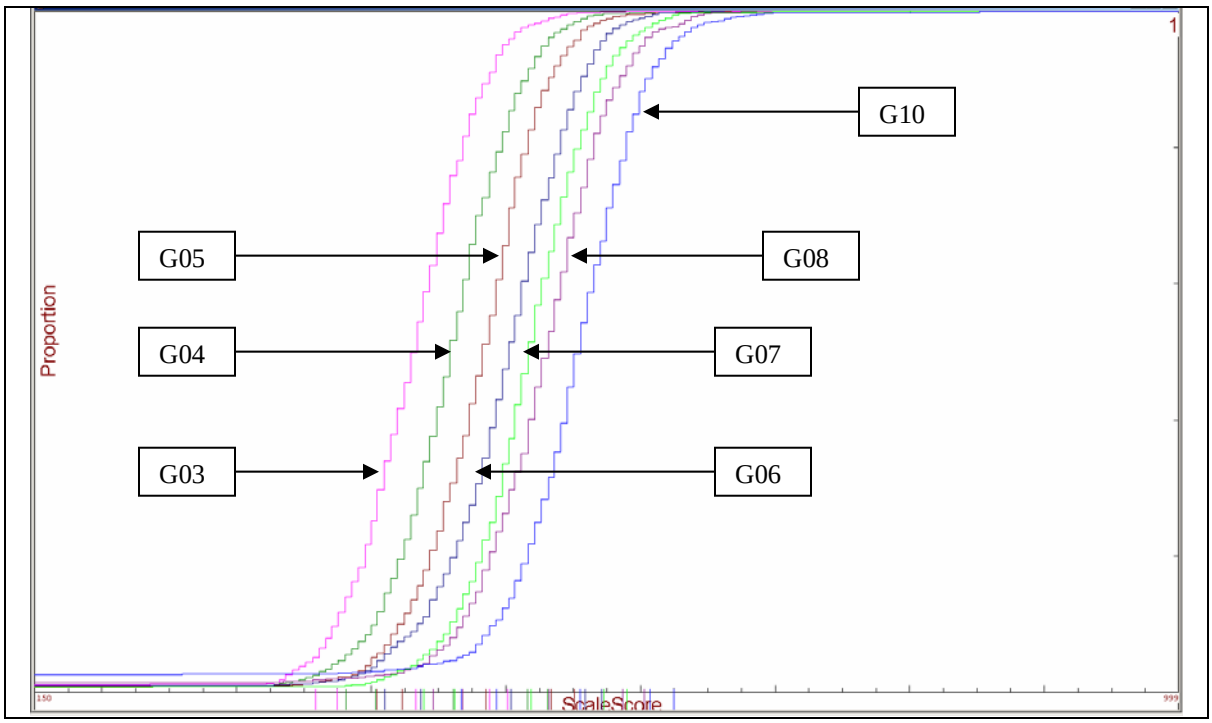


Figure 20
Cumulative Distribution Functions for Method 2 Mathematics

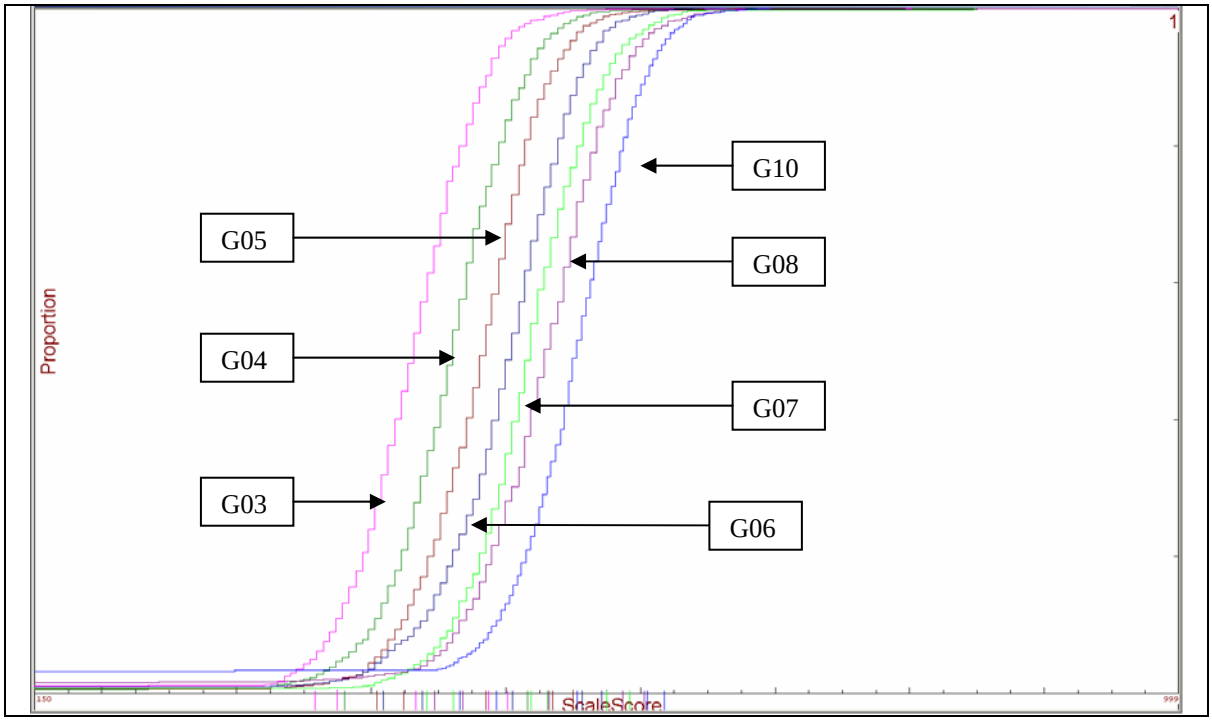


Figure 21
Cumulative Distribution Functions for Method 3 Mathematics

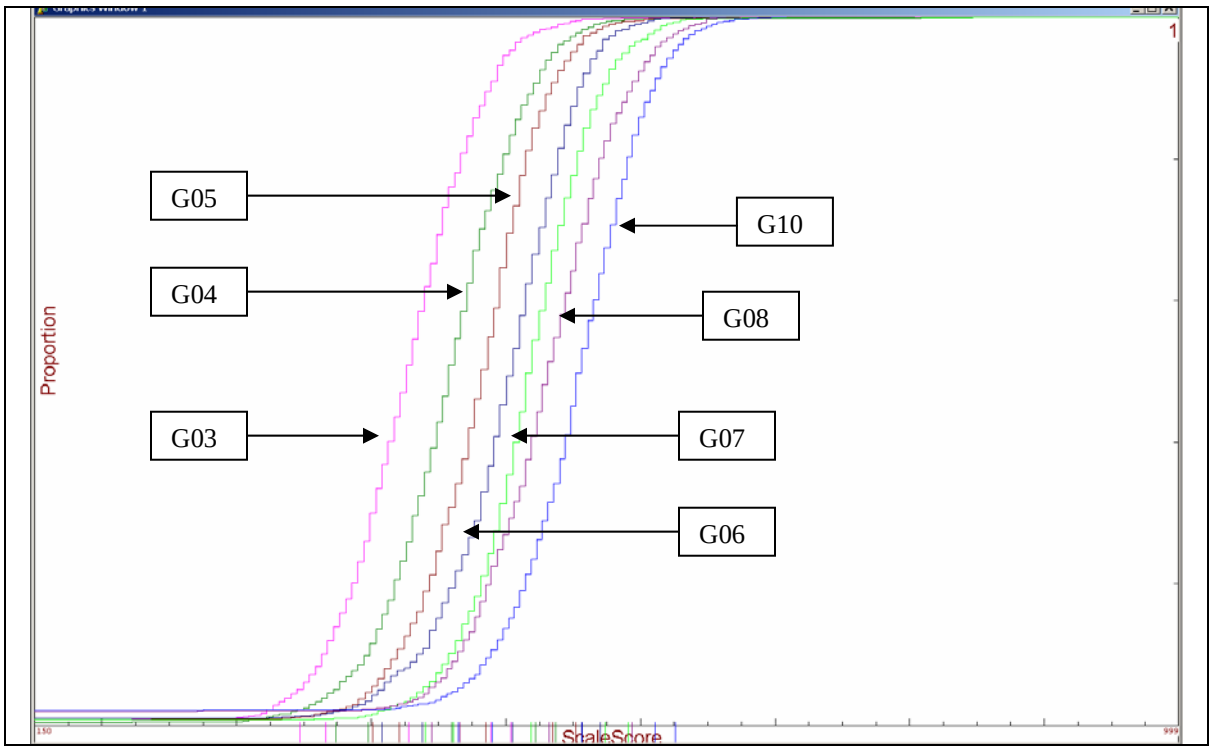


Figure 22
Cumulative Distribution Functions for Method 4 Mathematics

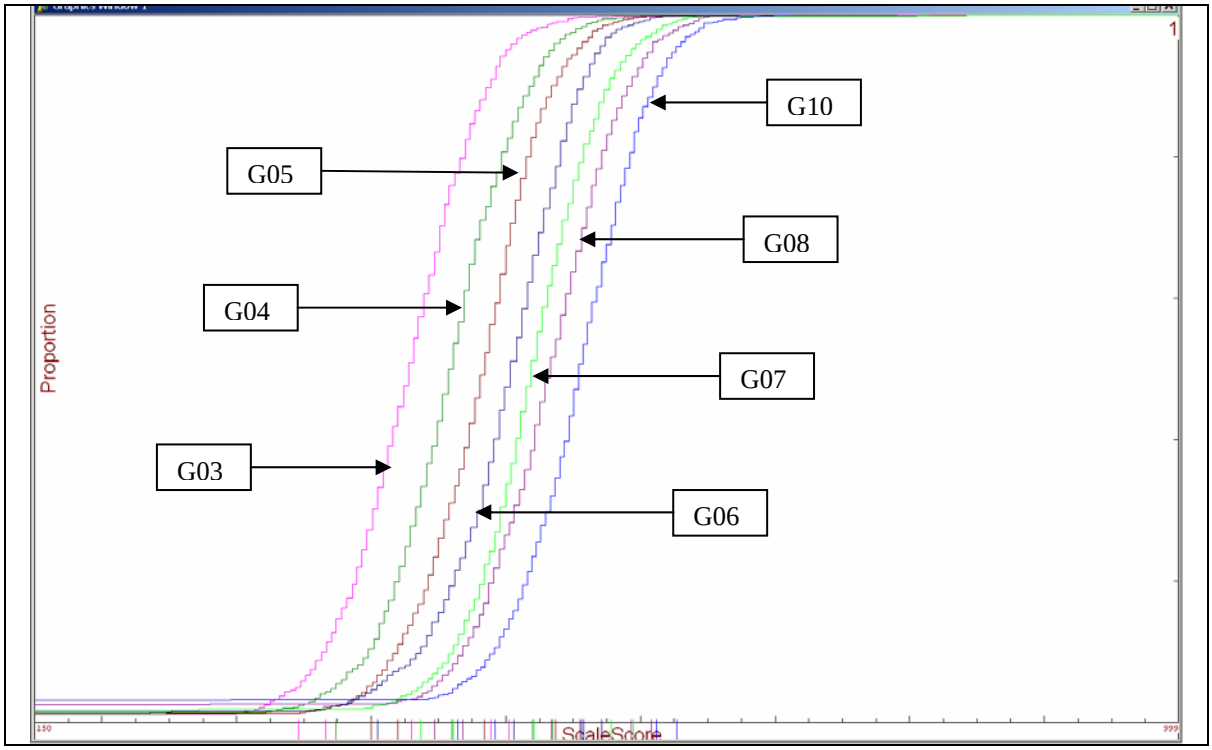


Figure 23
Test Characteristic Curves for Method 1 Reading

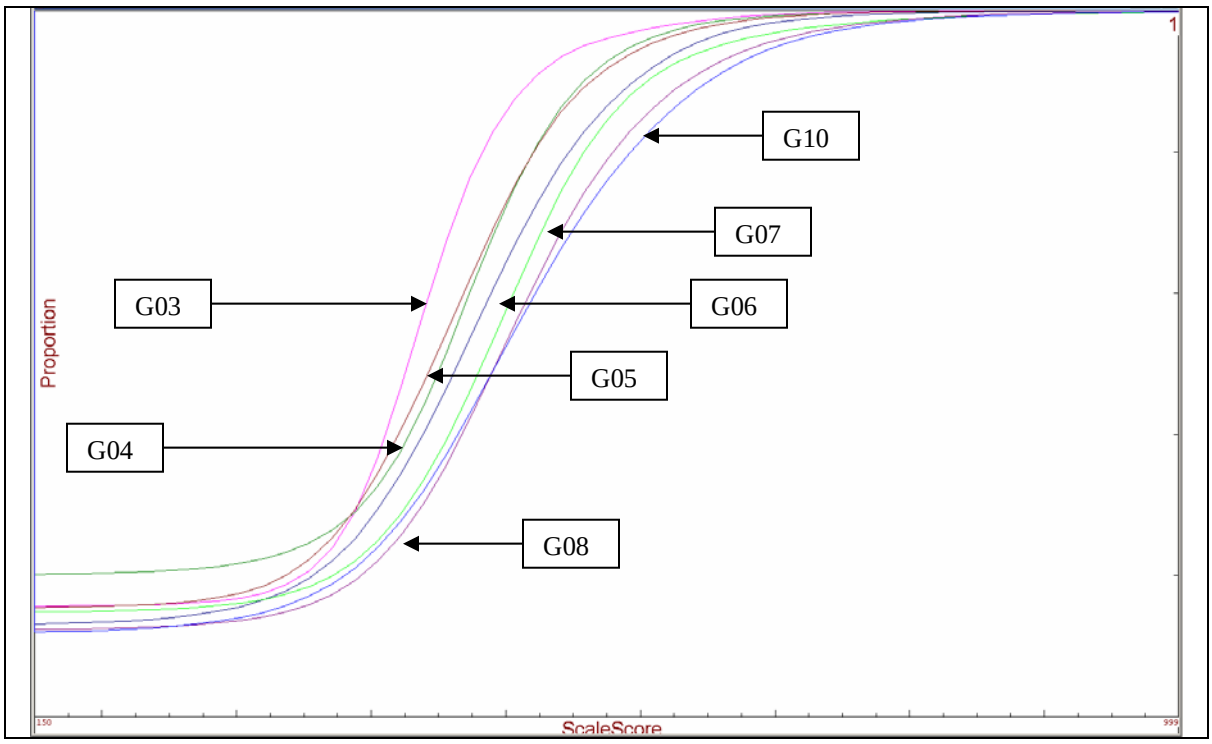


Figure 24
Test Characteristic Curves for Method 2 Reading

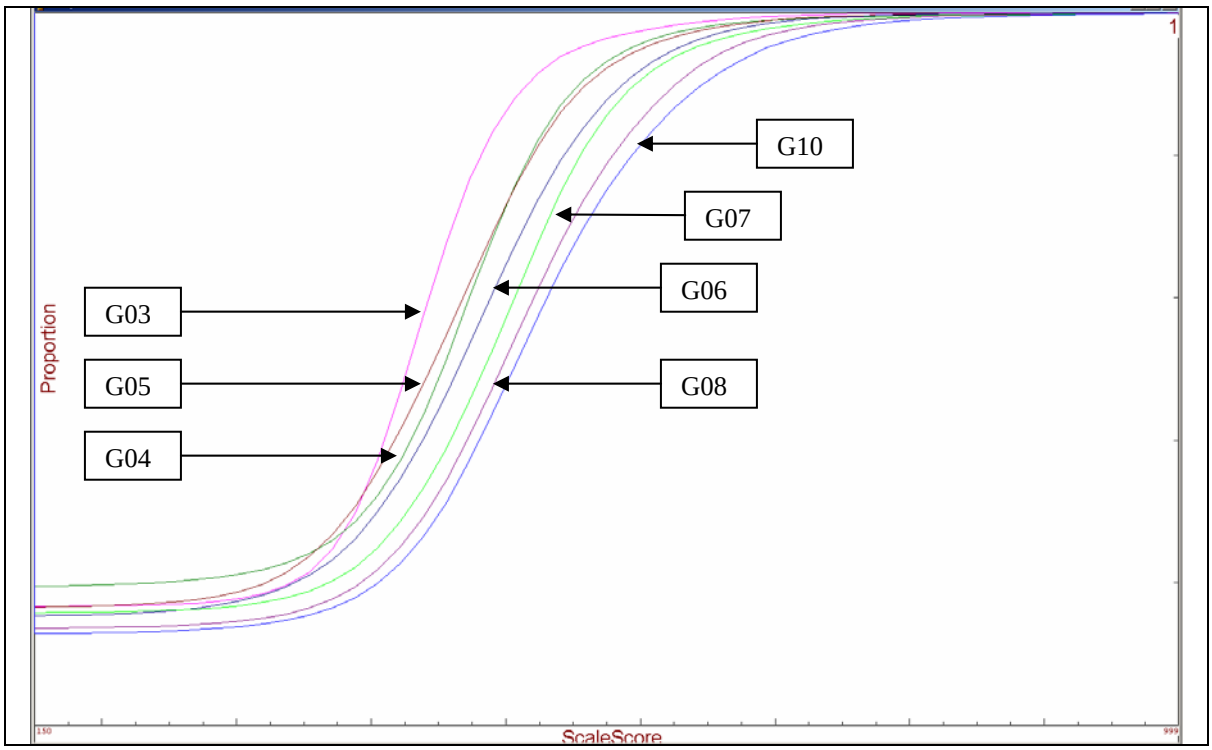


Figure 25
Test Characteristic Curves for Method 3 Reading

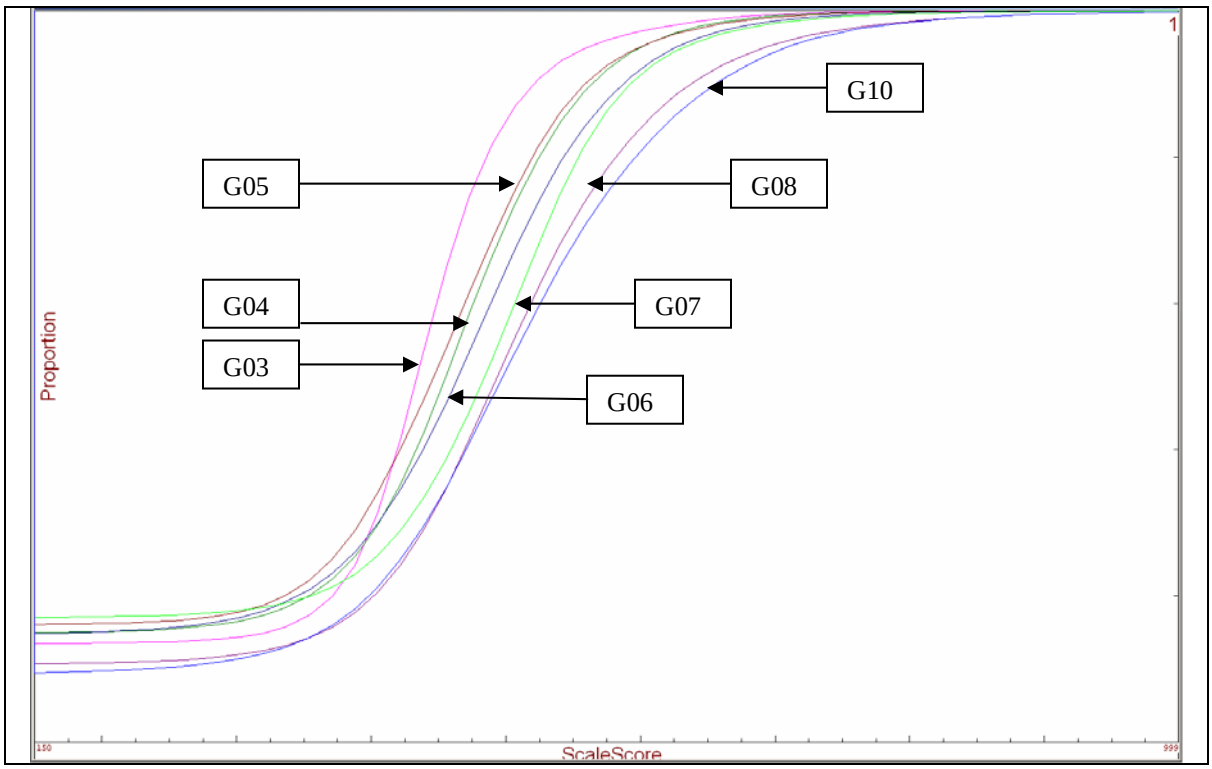


Figure 26
Test Characteristic Curves for Method 4 Reading

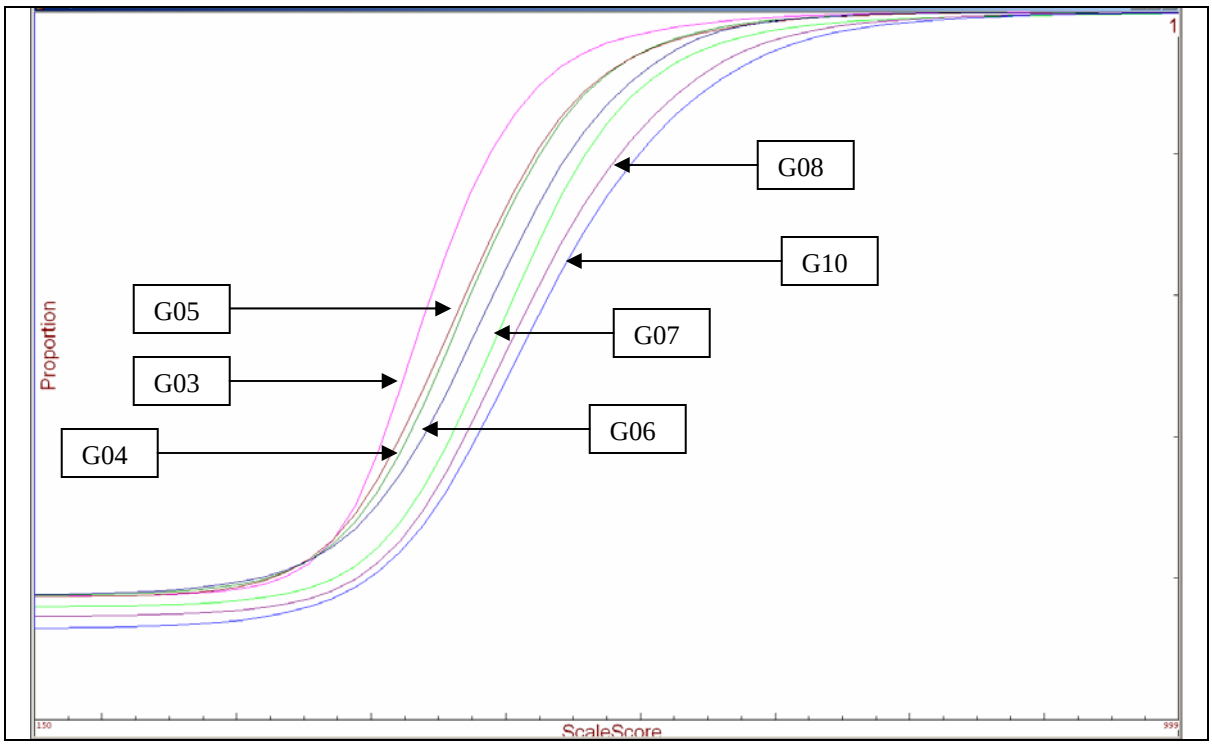


Figure 27
Cumulative Distribution Functions for Method 1 Reading

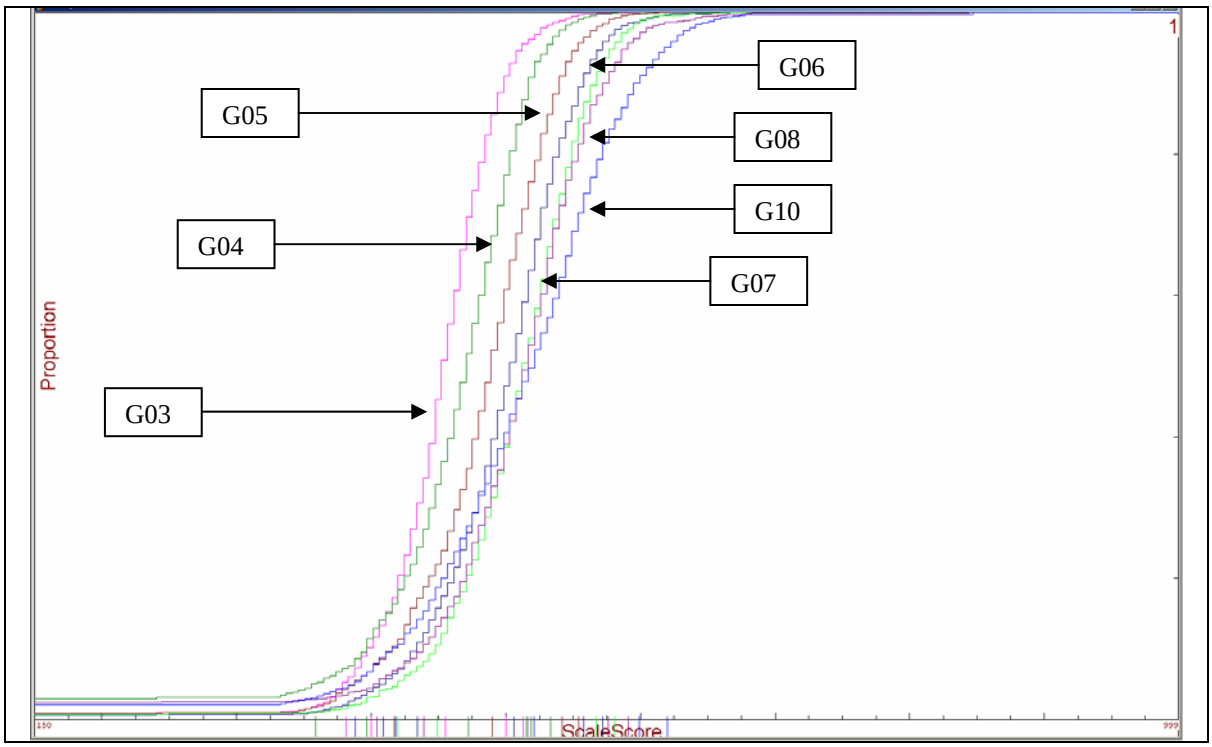


Figure 28
Cumulative Distribution Functions for Method 2 Reading

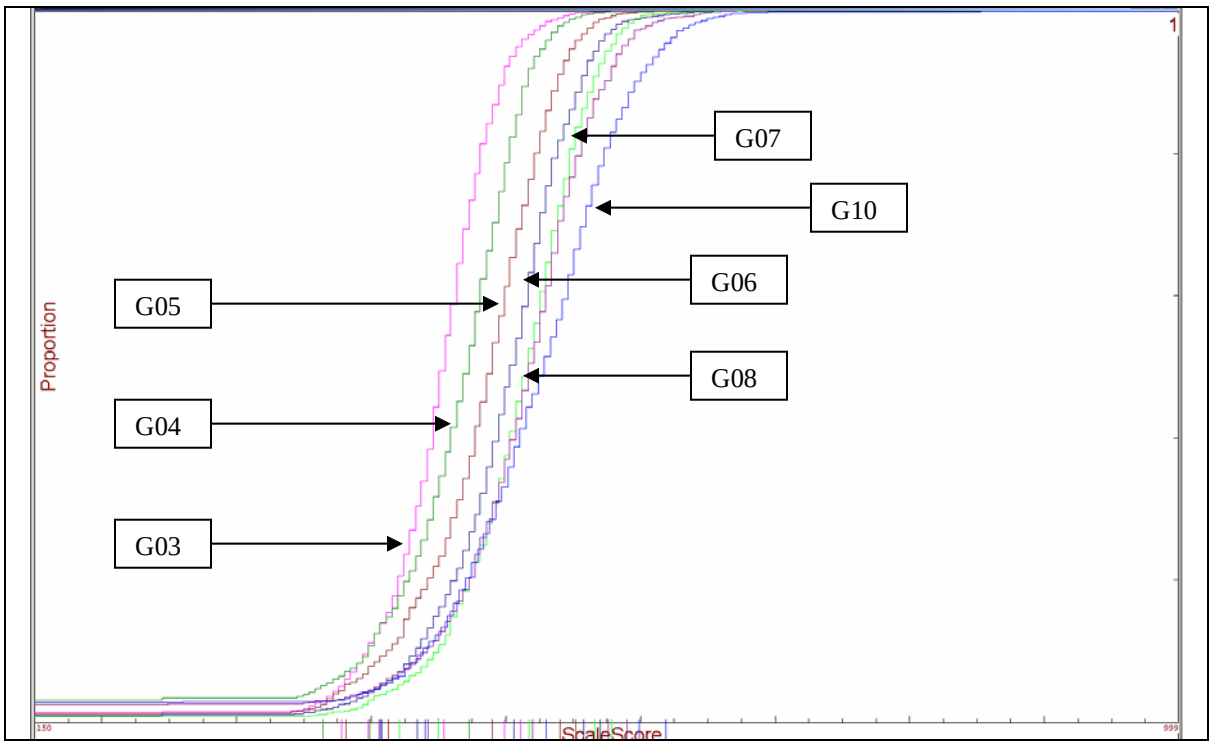


Figure 29
Cumulative Distribution Functions for Method 3 Reading

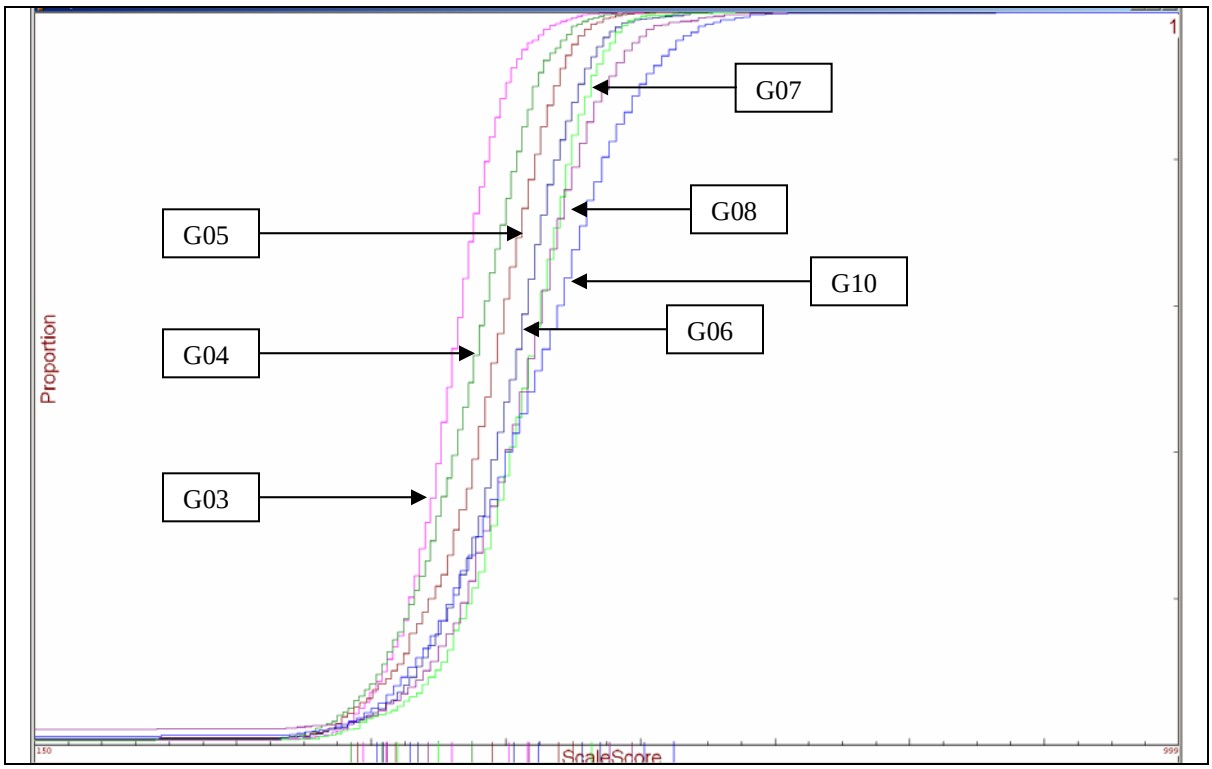
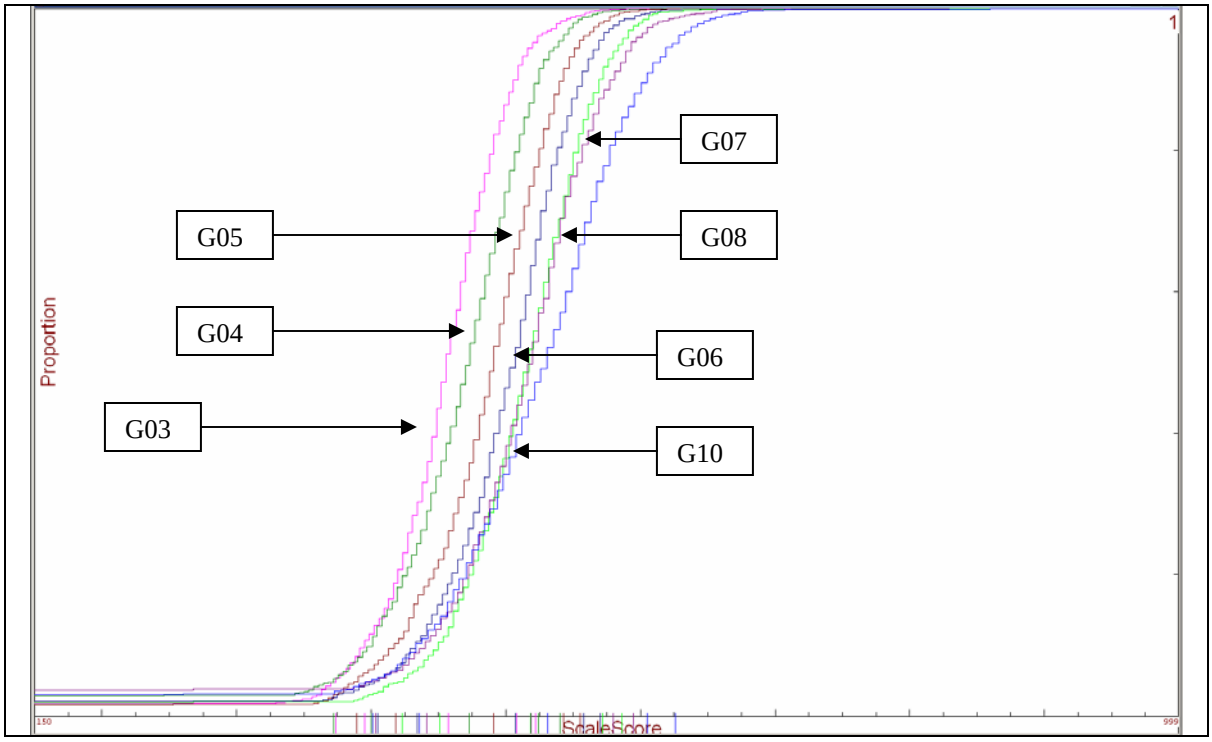


Figure 30
Cumulative Distribution Functions for Method 4 Reading



Part 4: Description of the First Operational Form

Although the vertical scaling forms were assembled in accordance with specifications and produced reasonable psychometric results, not all of the items had been field tested prior to this administration and, for some pairs of grades, the relative difficulties were not as differentiated as they might be. Thus, for some grade/content combinations, the forms were slightly modified. Note that in terms of the general content specifications, the modified forms are indistinguishable from the original vertical scaling forms. However, in several instances the number of items included in the final operational form was changed from the number of items used during the definition of the vertical scale. The purpose of this section is to document the psychometric attributes of the modified forms. Three features of each form are provided. Included are results derived from the December 2004 Field Test that summarize classical statistics, IRT analyses, and differential item functioning (DIF) indices.

Classical Form Summary

Tables 13 and 14 provide summaries of the expected classical statistics for mathematics and reading, respectively. For each grade, the tables report the mean p-value, the mean point biserial correlation, and the reliability. One feature worth noting is that for mathematics the forms generally become more difficult as grade level increases. This can be seen in the fact that the mean p-value for Grade 3 is 0.67 and the mean p-value for Grade 10 is 0.48. The variation in difficulty across grades for reading is notably smaller and has much less of a trend.

Table 13
Classical Statistics for Mathematics Fall 2005 Operational Forms
Grade 3–Grade 8/Grade 10

Grade	Total Number of Items	Total Number of SR Items	Total Number of CR Items	Total Score Points	Mean p-value	Mean Item Total Corr.
03	60	50	10	65	0.67	0.41
04	62	50	12	68	0.67	0.43
05	69	55	14	76	0.61	0.40
06	69	55	14	76	0.57	0.43
07	69	55	14	76	0.59	0.41
08	66	50	16	74	0.47	0.44
10	61	55	6	69	0.48	0.43

Table 14
Classical Statistics for Reading Fall 2005 Operational Forms
Grade 3–Grade 8/Grade 10

Grade	Total Number of Items	Total Number of SR Items	Total Number of CR Items	Total Score Points	Mean p-value	Mean Item Total Corr.
03	62	60	2	66	0.65	0.44
04	62	60	2	66	0.60	0.45
05	63	60	3	69	0.68	0.41
06	63	60	3	69	0.65	0.40
07	63	60	3	69	0.62	0.42
08	63	60	3	69	0.58	0.42
10	59	55	4	67	0.58	0.43

IRT Form Summary

Figures 31, 32, and 33 summarize test characteristic curves, conditional standard errors of measurement, and the information functions for mathematics. The underlying vertical scale defines the horizontal axis of each curve. As is hoped for, there is horizontal separation between the grades, and the TCCs are ordered in terms of difficulty.

Figures 34, 35, and 36 summarize test characteristic curves, conditional standard errors of measurement, and the information functions for reading. Although the scaled score difficulty of each form provides a different perspective than item p-values, a similar pattern can be observed. That is, the horizontal separation for reading is notably less than that observed for mathematics.

A comparison of Figure 31 with Figure 17 for Mathematics and Figure 34 with Figure 25 for Reading provides a perspective on some of the revisions that were made to provide somewhat better separation between adjacent grades than the vertical scale. Two changes are particularly worthy of note. First, for Mathematics, the Grade 7 form was modified slightly to move the TCC more to the right. Second, for Reading, the Grade 4 form was modified in order to move the TCC slightly to the left of the Grade 5 TCC. As is often the case when comparing IRT and classical results, the changes in mean p-values do not give a good sense of the overall impact of changes on each TCC.

Figure 31
Test Characteristics Curves (TCC) for WKCE-CRT Math
2005–06 Operational Form

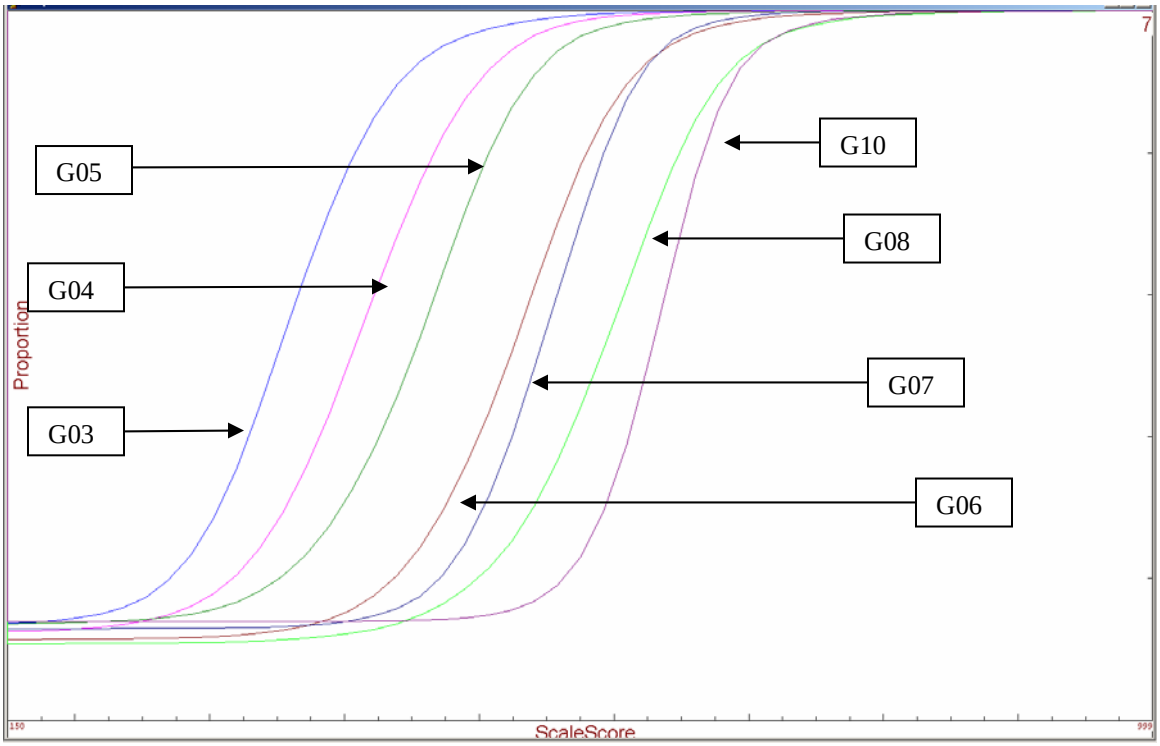


Figure 32
Standard Error Curves for WKCE-CRT Math
2005–06 Operational Form

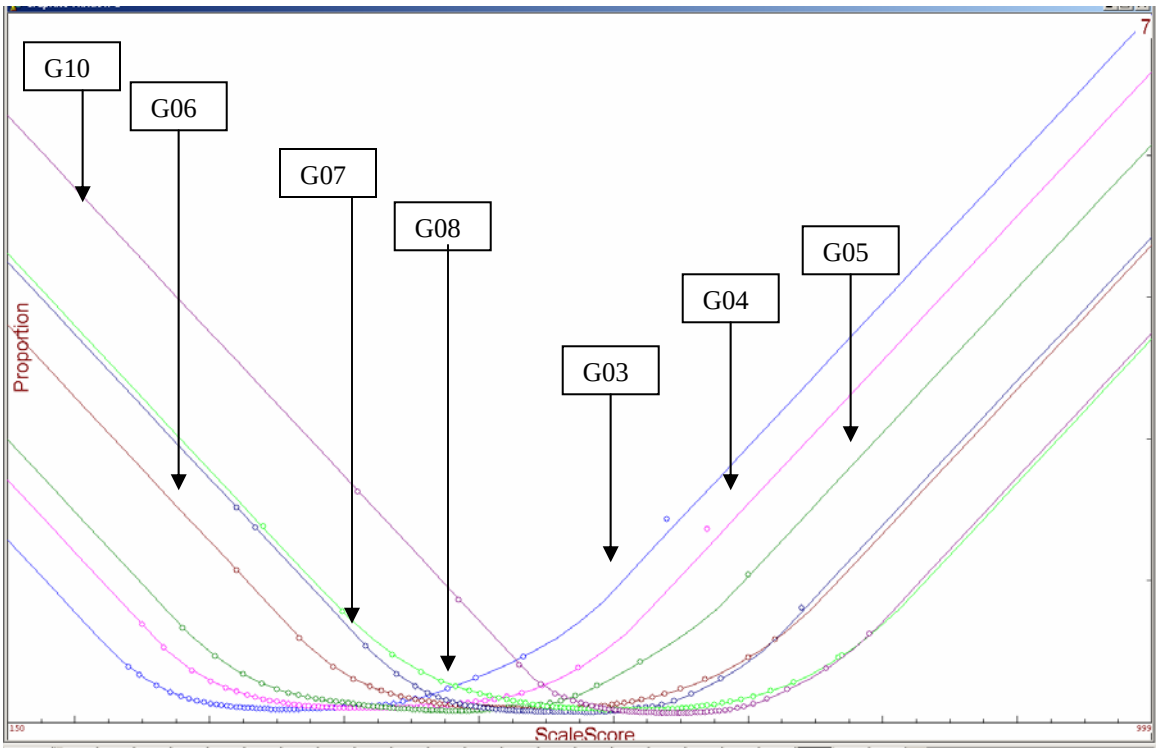


Figure 33
Information Curves for WKCE-CRT Math
2005–06 Operational Form

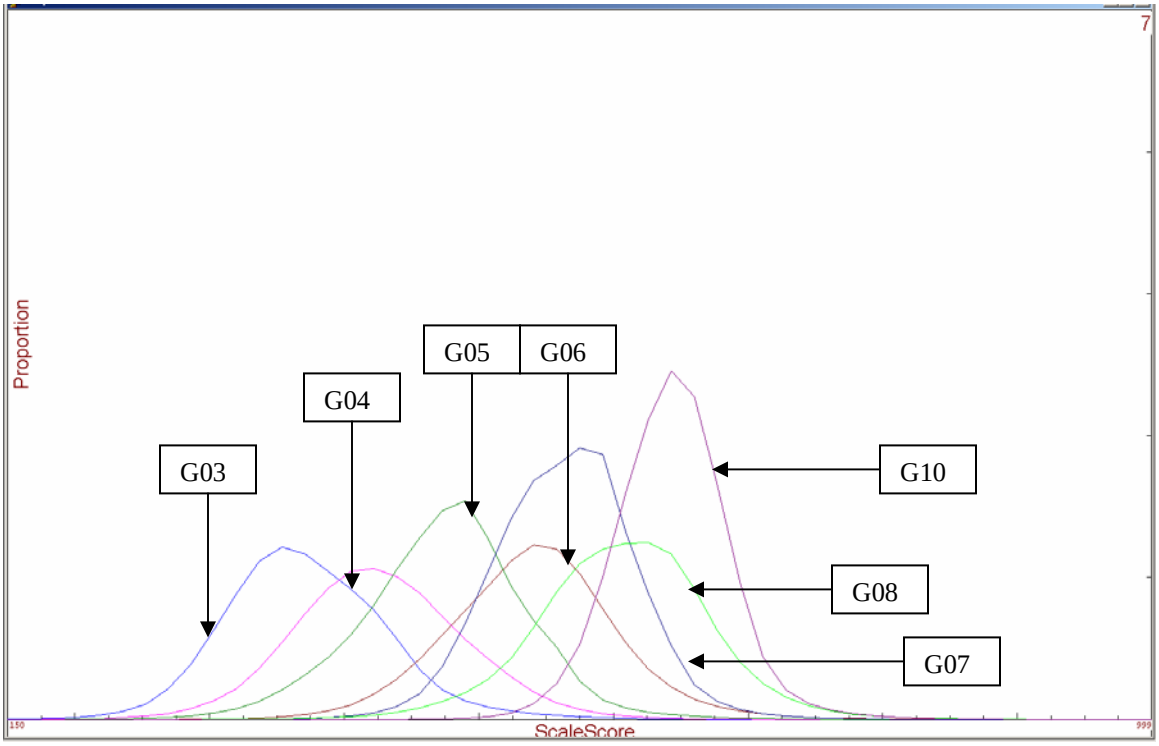


Figure 34
Test Characteristic Curves for WKCE-CRT Reading
2005–06 Operational Form

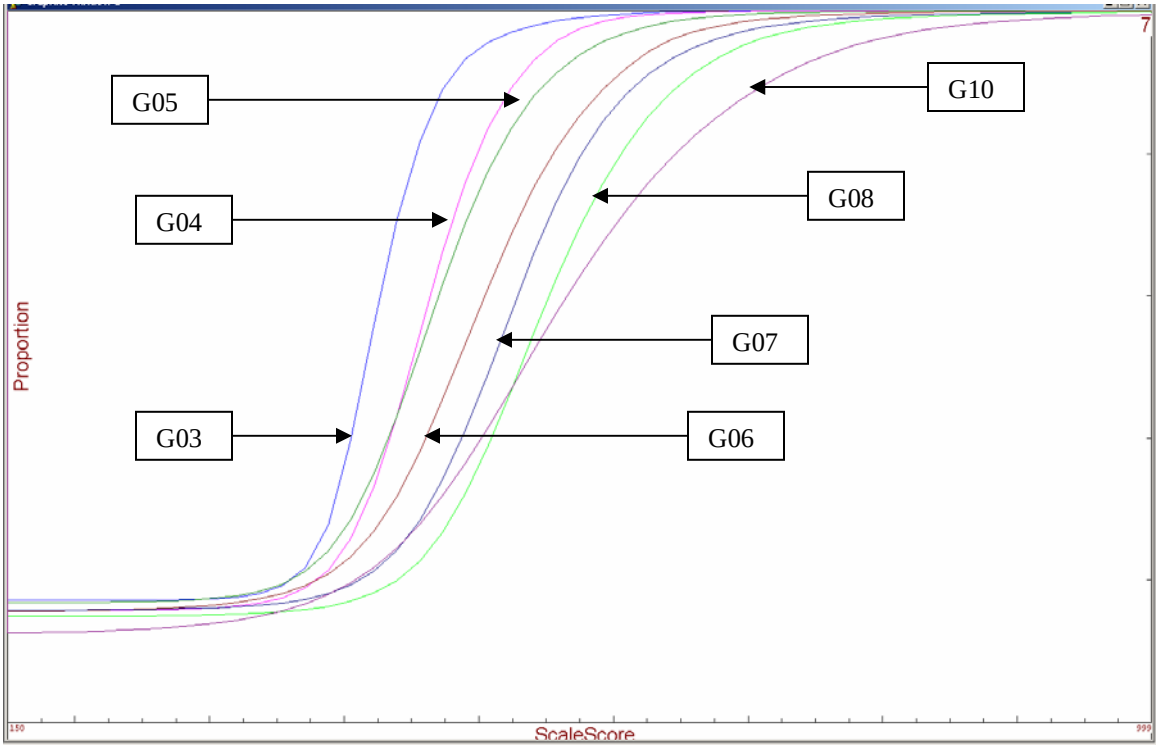


Figure 35
Standard Error Curves for WKCE-CRT Reading
2005–06 Operational Form

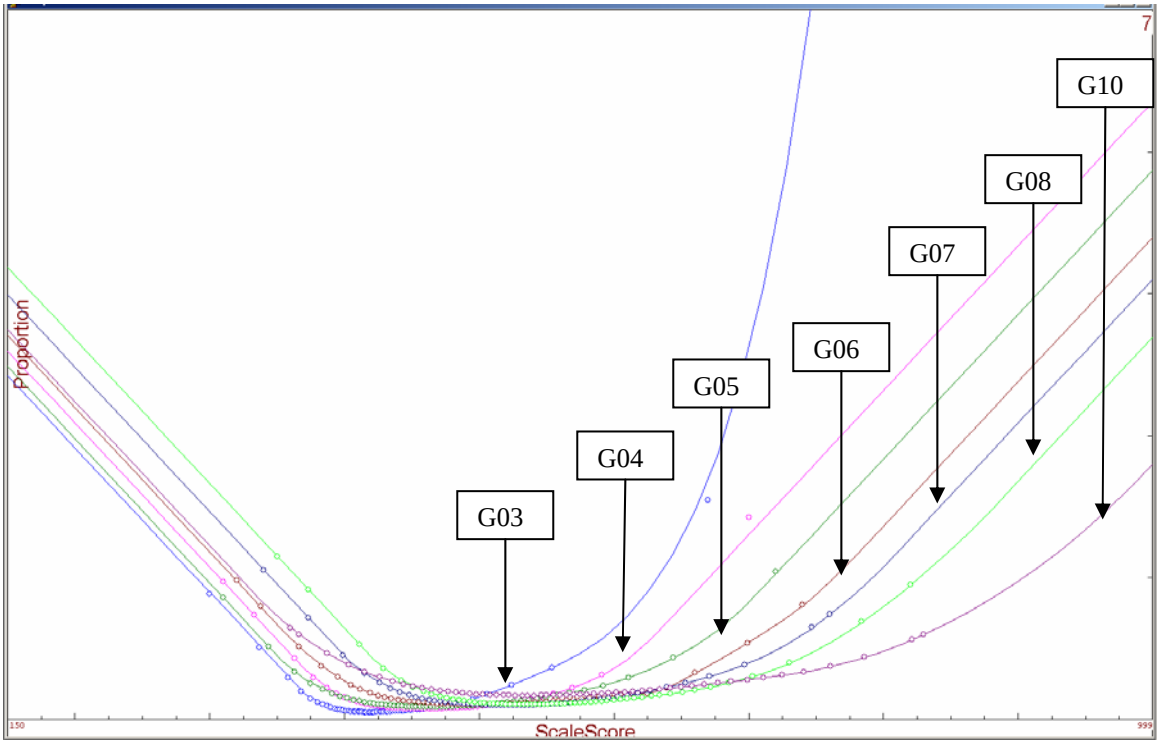
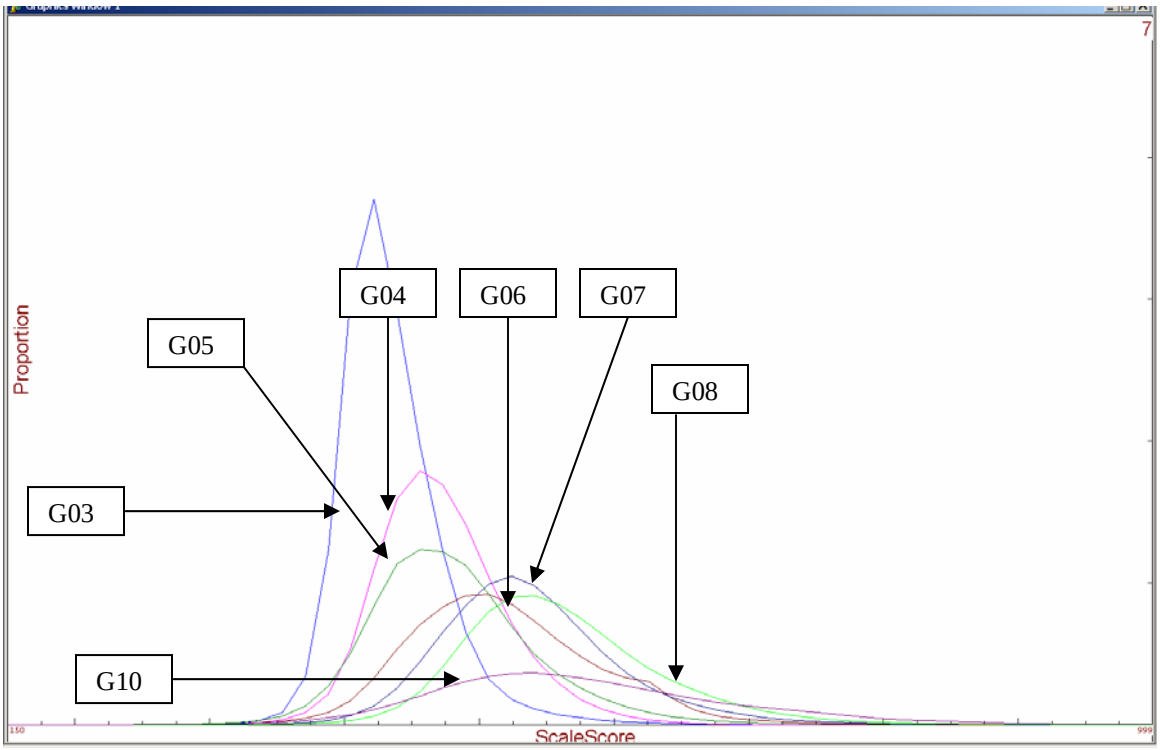


Figure 36
Information Curves for WKCE-CRT Reading
2005–06 Operational Form



DIF Summary

Differential Item Functioning (DIF) indices are known to produce a number of Type I errors (i.e. flag items that have no discernible issues). Thus, three different methods were used to flag items and only items that were consistently flagged were classified as potentially problematic. The three methods used were Linn and Harnisch (1981), Mantel-Haenszel (1959), and Standardized Mean Difference.

Linn and Harnisch (L–H)

The Linn and Harnisch procedure is an IRT based methodology that evaluates the difference between the ICC calibrated using only data from students in the focal group with the ICC calibrated using data from the total group of students. To the degree that the two ICCs are different, DIF is assumed to be the cause.

DIF is defined in terms of the decile group and total target subsample differences, the D_i^-(sum of the negative group differences) and D_i^+(sum of the positive group differences) values, and the corresponding standardized difference (Z_i) for the subsample (see Linn and Harnisch, 1981, p. 112). Items for which $|D_i| \geq 0.010$ and $|Z_i| \geq 2.58$ are flagged for DIF. If D_i is positive, the item is biased in favor of the target subsample. If D_i is negative, the item is biased against the target subsample.

Mantel-Haenszel (M-H) and Mantel

M-H (1959) is a procedure testing the constant odds ratio hypothesis version of DIF. The null DIF hypothesis for the Mantel-Haenszel method is: the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and the reference group, across all levels of the matching variable. The extension of M-H called Mantel (1963) was applied for both MC and CR items. In the case of MC items, this statistic is identical to the M-H statistic without the continuity correction. There is a chi-square test associated with the M-H approach. The M-H approach is the statistical test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds ratio hypothesis. Rejection of null hypothesis suggests that members of the reference and focal groups who are similar in overall proficiency nevertheless tend to differ in their mean performance on the studied item. The criterion used for flagging items was $\chi^2 = 7.99$ ($p=0.005$).

Standardized Mean Difference (SMD)

Unlike the M-H method, the SMD developed by Zwick et al (1993) is an impact type statistic in that the performances compared are not defined for matched groups. The concept of this statistic is that rather than comparing conditional odds-ratios like the M-H method, the conditional means for the two groups are weighted by the frequency distribution for the focal group. Specifically, the statistic is formed by

$$SMD = \sum p_{Fk} m_{Fk} - \sum p_{Fk} m_{Rk},$$

where $p_{Fk} = n_{F+k} / n_{F++}$ is the proportion of focal group members who are at the k-th level of the matching variable, $m_{Fk} = (1/n_{F+k})(\sum y, n_{Rtk})$ is the mean item score for the focal group at the k-th level, and $m_{Rk} = (1/n_{R+k})(\sum y, n_{Rtk})$ is the analogous value for the reference group. A positive value for a SMD reflects DIF in favor of the focal group.

The SMD index expresses results in an item score metric. When item scores are expressed as p-values, the theoretical range for the SMD index is -1.0 to 1.0. Items producing an SMD with an absolute value of .10 and larger are indicative of significant DIF and should be examined carefully.

Tables 16 and 17 summarize, by grade, the DIF statistics for mathematics and reading, respectively. DIF statistics were computed for each gender and ethnicity subgroup with a sample size of at least 200. Any item flagged by at least two of the procedures is included in the table. Items were flagged regardless of whether they favored the focal or reference group. The majority of the differentially functioning items favored the focal group. All items favoring the reference group were MC items.

Table 16
Mathematics Items Flagged for DIF by Focal Group

				Focal Group	
Grade	Max Point	Item #	Type	Female	African-American
03	1	523831	MC	-	+(M), +(S)
03	1	523833	MC	-	+(M), +(S)
03	1	560703	MC	-	+(L), +(M), +(S)
04	1	524031	MC	-	+(M), +(S)
04	1	524037	MC	-	+(M), +(S)
04	1	524147	MC	-	+(M), +(S)
04	2	525318	CR	+(M), +(S)	-
05	1	524229	MC	-	+(M), +(S)
05	1	524369	MC	-	*(L), *(M)
05	1	524417	MC	-	+(M), +(S)
05	2	525331	CR	+(M), +(S)	-
05	2	525339	CR	+(M), +(S)	-
06	1	524599	MC	-	+(M), +(S)
06	1	524745	MC	-	+(L), +(M), +(S)
06	2	525359	CR	+(M), +(S)	-
06	2	525363	CR	+(M), +(S)	-
07	1	524951	MC	*(L), *(M)	-
07	2	525382	CR	+(M), +(S)	-
10	2	261389	CR	+(M), +(S)	-
10	2	264476	CR	+(L), +(M), +(S)	-
10	4	264542	CR	+(M), +(S)	-
10	2	265104	CR	+(M), +(S)	-
10	1	265330	MC	+(M), +(S)	-

Note 1. L–Linn and Harnisch, M–Mantel, S–Standardized Mean Difference.
Note 2. "+" or "*" is used to indicate that the item is in favor of or against the focal group.

Table 17
Reading Items Flagged for DIF by Focal Group

				Focal Group			
Grade	Max Point	Item #	Type	Female	African-American	Hispanic-American	Asian/Pacific Islander
03	1	525790	MC	-	-	-	*(L), *(M)
03	1	525798	MC	-	-	-	+(L), +(M), +(S)
03	3	525969	CR	+(L), +(M), +(S)	-	-	-
04	1	526441	MC		+(M), +(S)	-	-
05	3	526559	CR	+(M), +(S)	-	-	-
05	3	526664	CR	+(M), +(S)	-	-	-
06	3	526896	CR	+(M), +(S)	-	-	-
06	3	526917	CR	+(M), +(S)	-	-	-
06	1	526939	MC	-	-	*(L), *(M)	-
06	3	526991	CR	+(M), +(S)	-	-	-
07	1	527141	MC	+(M), +(S)	-	-	-
07	3	527159	CR	+(L), +(M), +(S)	-	-	-
08	1	527436	MC	*(L), *(M)	*(L), *(M)	-	-
08	1	527446	MC	-	*(L), *(M)	-	-
08	3	527458	CR	+(M), +(S)	-	-	-
08	1	527573	MC	-	+(M), +(S)	-	-
08	1	527630	MC	-	*(L), *(M)	-	-
08	3	527634	CR	+(L), +(M), +(S)	-	-	-
10	3	256318	CR	+(L), +(M), +(S)	-	-	-
10	3	256374	CR	+(L), +(M), +(S)	-	-	-
10	3	256446	CR	+(L), +(M), +(S)	-	-	-

Note 1. L–Linn and Harnisch, M–Mantel, S–Standardized Mean Difference.
Note 2. "+" or "*" is used to indicate that the item is in favor of or against the focal group.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in tow or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443–459.

Burket, G. R. (1991). *PARDUX* [Computer program]. Unpublished.

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25,* 291–312.

CTB/McGraw-Hill. (1968). *Comprehensive Tests of Basic Skills, Form S.* Monterey, CA: Author.

CTB/McGraw-Hill. (1996). *CAT/5 Technical Report*. Monterey, CA: Author.

CTB/McGraw-Hill. (1997). *TerraNova Technical Bulletin 1*. Monterey, CA: Author.

CTB/McGraw-Hill. (1997). TerraNova 1st Edition. Monterey, CA: Author.

CTB/McGraw-Hill. (2000). TerraNova 2nd Edition. Monterey, CA: Author.

CTB/McGraw-Hill. (1997–2001). WKCE. Monterey, CA: Author.

Fitzpatrick, A. R. (1991). *Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE program*. Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A. R., & Julian, M. W. (1996*). Two studies comparing the parameter estimates produced by PARDUX and PARSCALE*. Unpublished manuscript.

Green, D. R. (1975, December). *Procedures for assessing bias in achievement tests*. Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kolen, M.J. & Brennan, R.L. (2004), Test equating methods and practice. New York: Springer-Verlag.

Linn, R. L. (1989). *Educational measurement*. (3rd ed). New York: Macmillan.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*, 109–118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159–173.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247-264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (71, 179–181). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Macmillan/McGraw-Hill. (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*. New York: Author. McGraw-Hill. (1983). Guidelines for bias-free publishing. New York: McGraw-Hill Book Company.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association*, 58, 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Jounral of the National Cancer Institute*, 22, 719-748.

Marzano, Robert J, et al. (1988). Dimensions of Thinking: A Framework for Curriculum and Instruction. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
McGraw-Hill. (1983). *Guidelines for bias-free publishing*. New York: McGraw-Hill Book Company.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.

Rasch, G. (1960). *Probalistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institue for Educational Research.

Sandoval, J. H., & Mille, M. P. (1979, August). *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist, 19*, 219–225.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Thissen, D. (1990). *MULTILOG: Multiple categorical item analysis and test scoring*. Version 6 [Computer program]. Mooresville, IN: Scientific Software.

van der Linden, W.J. & Hambleton, R. (1997). *Handbook of modern item response theory*. New York: Springer.

Wright, B. D., & Linacre, J. M. (1992). *BIGSTEPS Rasch Analysis* [Computer program]. Chicago, IL: MESA Press.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93–111.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Yen, W. M. & Burket, G.R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*, 293–313.

Yen, W. M., Burket, G. R., & Sykes, R.C. (1988*). Non-unique solutions to the likelihood equation for the three-parameter logistic model*. Paper presented at the meeting of the Psychometric Society, Los Angeles.

Yen, W.M. & Candell, G.L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4*, 209–228.

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251