Oregon Department of Education

2006–2007

Technical Report

Oregon's Statewide Assessment System

Reliability and Validity

Volume 4

Last updated on April 25, 2007

Oregon's Statewide Assessment System
Technical Report: Volume 4, Reliability and Validity
Last updated on April 25, 2007

Oregon Department of Education
Office of Assessment and Information Services
255 Capitol Street NE
Salem, OR 97310
503-378-3600
http://www.ode.state.or.us/

Susan Castillo
State Superintendent of Public Instruction

Doug Kosty
Assistant Superintendent

Tony Alpert
Director, Assessment and Evaluation

Steve Slater
Manager, Scoring, Psychometrics and Validity

Chris Minnich
Manager, Test Design and Implementation

Elaine Hultengren
English Language Proficiency Specialist

Ken Hermens
Language Arts Assessment Specialist

Leslie Phillips
Science Assessment Specialist

Leslie Phillips
Social Sciences Assessment Specialist

Sheila Somerville
Electronic Publishing Specialist

Cathy Brown
Mathematics Assessment Specialist

This technical report is one of a series that describes the development of Oregon's Statewide Assessment System. The complete set of volumes provides comprehensive documentation of the development, procedures, technical adequacy, and results of the system:

Volume 1: 2005–2006 Annual Technical Report
Volume 2: Test Development
Volume 3: Standard Setting
Volume 4: Reliability and Validity
Volume 5: Test Administration
Volume 6: Score Interpretation Guide
Volume 7: Alternate Assessment, Program Description
Volume 8: Alternate Assessment, 2005–06 Statistical Summary

All volumes can be found at http://www.ode.state.or.us/search/page/?id=787.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. OVERVIEW

A series of Technical Reports were commissioned in 2006 to provide information about the technical and procedural characteristics of Oregon's Statewide Assessment System (OSAS), created by the Office of Assessment in the Oregon Department of Education with considerable participation and involvement from Oregon educators.

To summarize and inform audiences by compiling existing documentation from a variety of sources into a single easily accessible document, the 2005–06 Technical Reports are the first in a series of reports summarizing and describing the assessment system. Consisting of eight volumes, the reports describe the development, operational procedures, and technical features. The annual report (Volume 1) describes student performance and documents changes to the system and assessment-related activities undertaken during the year. Volume 8 describes the results of the alternate assessment administered to students with disabilities and is also updated every year. The Department updates Volumes 2 through 6 as new information becomes available or as new procedures are implemented. Together, the reports describe the progress toward meeting the academic achievement standards of Oregon's public school students and the process and technical adequacy through which this progress is measured.

Updated as new evidence is collected demonstrating the states assessment system, Volume 4: Validity and Reliability describes the reliability, validity, and comparability evidence relating to Oregon's tests.

2. PURPOSE OF OREGON'S STATEWIDE ASSESSMENT

Students in Oregon public schools take assessments created by the Office of Assessment in the Oregon Department of Education. Oregon's assessments need to demonstrate that they measure what they are supposed to measure (how well students have mastered state content standard) and that scores are consistent in that results mean the same thing regardless of when, where, or who takes the assessment.

The intended purpose and interpretation of the data from the OSAS as called for under federal and state legislation is to monitor student progress toward mastery of the state content standards (see the Oregon Act for the 21st Century at http://landru.leg.state.or.us/ors/329.html or the Oregon Revised Statutes at http://www.leg.state.or.us/ors/). By aggregating across groups of students, the test scores also are designed to provide information about the relative success of schools, districts, and the state as a whole in helping students master the state content standards. The goals of the system are as follows:

- Provide instructionally useful evaluation of individual evaluation student progress toward mastery of the Academic Content Standards
- Guide instructional program improvement
- Ensure that the state is progressing toward the state and federal goals for high standards for all
- Inform the public

The validity and use of data from the tests depend on evidence that the tests measure the content that is required by the Academic Content Standards and do so reliably.

3. Introduction to Technical Adequacy

OSAS tests are rigorously examined in reference to the guidelines provided in the Standards for Educational and Psychological Testing (1985, 1999), which provide criteria for the evaluation of tests, testing practices, and effects of test use for a broad set of assessments, including alternate assessments. The Critical Elements identified in NCLB legislation further describe the evidence based on these standards that is necessary to validate the tests for the intended purposes.

Validation is a process of developing a scientifically based argument supporting the intended use of test scores and their relevance to their intended use. According to the standards, validity refers to the degree to which evidence and theory support the interpretations of test scores as described by the intended uses of tests. The appropriateness and usefulness of the entire assessment system rests on the tests meeting the relevant standards of validity. Demonstration is required to show that Oregon's tests

1. measure what they are supposed to measure. In this context, assessments must measure how well students have mastered state content standards.

2. scores or ratings are consistent. Results must mean the same thing regardless of when the assessment was taken, where it was administered, and what group is being assessed.

These standards include the following components of technical quality:

Reliability
Content validity
Concurrent (Construct) validity
Criterion validity
Comparability of scores
Fairness and accessibility

The sections below describe evidence of how Oregon's assessment system satisfies these standards.

4. Reliability

Reliability refers to the consistency, stability, and accuracy expected from test scores. Oregon has conducted several studies of reliability describing the standard errors of measurement and classification accuracy overall and for subgroups and for the TESA (long and short) and paper-and-pencil versions of the assessments.

This evidenced is summarized below and described in the following numbered sections:

4.1 Analysis of the standard errors of measurement (using operational data from 2003–04 suggests that the system of assessments provides similarly reliable test scores across the range of ability, except for the extreme ends of the distribution. Standard errors of measurement are similar across TESA and paper and pencil and by subgroup (ethnicity, LEP, and Special Education), showing that the proficiency of students with the same overall proficiency level is measured with the same reliability regardless of demographic subgroup.

4.2 The results of another study indicate reliability of achievement classification through high classification accuracy for all paper-and-pencil, Plain Language, and TESA forms for Math and Reading, using 2003–2004 operational data across the range of performance levels ("Not Yet Meets," "Meets," "Exceeds").

4.3 A study of classification accuracy (using 2005–06 data) provides evidence of high accuracy in classification for the paper and pencil and TESA (long and short) across the range of student proficiency.

4.4 To support the state in reporting scores in a credible, defensible, and useful manner, another study (2005–06) describes the strand reliability and the precision of cut scores.

4.1 Test Information Curves and Standard Error of Measurement

Source: Northwest Evaluation Association. (2005). Oregon Statewide Assessment System. Standard errors of measurement, reading and mathematics knowledge and skills tests.

Item response theory addresses reliability by decomposing the information in a score from the "noise" or error in a score. Test information functions (introduced in Volume 1) are the inverse of standard errors of measurement and describe the extent to which tests provide consistent information across the range of student abilities. The information function for an individual dichotomous item for any ability (any value of theta) is:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

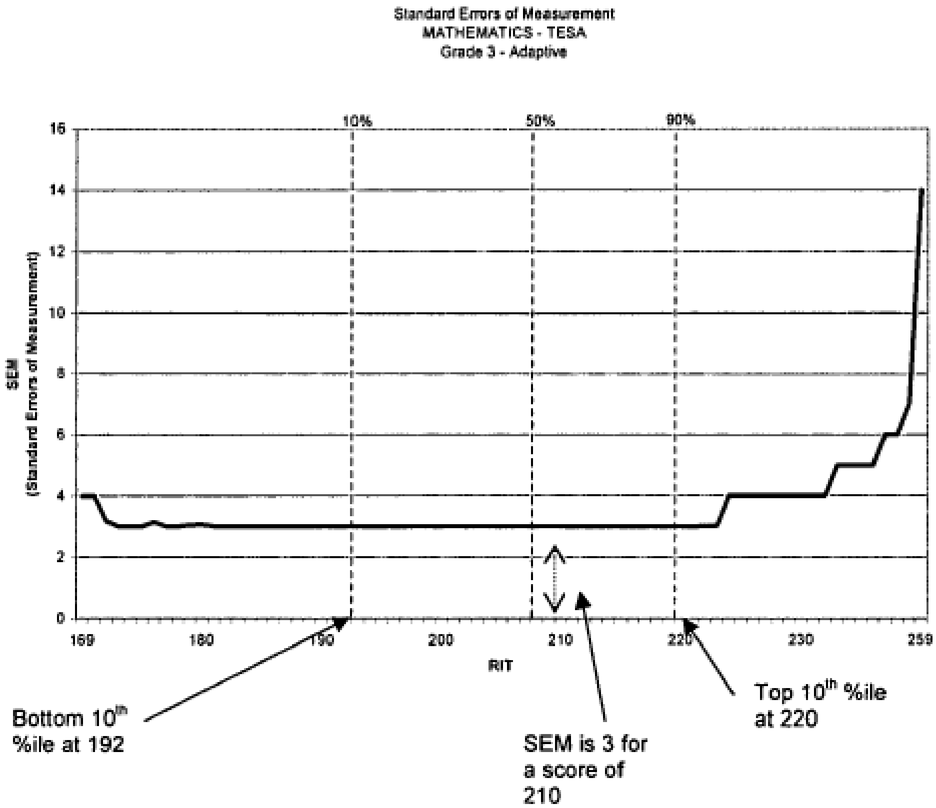
where P(θ) is the probability of getting an item correct given θ and Q(θ) = 1-P(θ).

For each score, the standard error of measurement is the root of the inverse information function

$$SE_i(\theta) = \sqrt{\frac{1}{I_i(\theta)}}$$

The example standard error curve below is for the adaptive Math test for grade 3. This curve demonstrates that for the 90% of students earning scores between 192 and 220, the standard error of measurement was approximately 3 RIT points and provided similar information across the range of proficiency, except for the students in the 99th percentile, where the standard error increases.

Figure 1.
Standard Error of Measurement Curve



The 2003–04 analysis of the standard errors of measurement suggests that Oregon's system of assessments provides similar and consistent information across the range of ability.

The graphs in Appendix A provide the standard error curves for TESA and for each paper-and-pencil test form of the OSAS for Math and Reading, grades 3–10. Results are presented by test (TESA and paper and pencil) and by subgroup (ethnicity, LEP, and Special Education); all are similar to the curve described above, across TESA and paper and pencil and by subgroup—showing that students with the same overall proficiency level have the same amount of error regardless of demographic subgroup.

Appendix A describes results for Plain Language forms as well as for targeted paper-and-pencil forms; both were operational in 2003–04, although neither is used operationally in the current system.

4.2 Reliability of Achievement Classification, NWEA Classification Analysis 2003–04

Source: McCall, M.S. Technical review of the Oregon Department of Education assessment system. Northwest Evaluation Association.

Current school and district evaluation models use achievement categories rather than scores. The reliability of placement into these categories is a type of decision consistency measure. In general, when a category is represented by a range of scores, a student with proficiency in the middle of the range is more likely to be placed accurately than one at or near either end. For IRT tests, an estimate of the proportion of students correctly classified, given their item responses, can be made using the SEMs around the cutpoints (2.14). Rudner (2006) used this approach to evaluate the classification accuracy from Oregon's tests.

The results, in Appendix F, provide evidence of reliability of achievement classification through high classification accuracy for all student proficiencies and high accuracy for all forms. Overall reliabilities ranged from 84–99%, with most falling above 90%.

4.3 Reliability of Achievement Classification, 2005–2006 Classification Analysis

Source: Doran, H., and Cohen, J. (2006) Oregon Technical Report I. Technical Report for the Oregon Department of Education.

Another study of classification accuracy used a different method to evaluate the consistency of classification. Classification analysis describes the probability of misclassification for each RIT scale score using an IRT-based method for assessing classification accuracy for students scoring at each RIT score. For students above the cut score for scoring in the Meets category, the study estimated the probability of a "false positive"—labeling the examinee proficient when he or she is not. Conversely, for those students scoring below the cut score required for the Meets category, the study estimated the probability of a "false negative"—labeling the examinee not proficient when he or she is.

The analysis used three test forms (the TESA [full and short] and paper and pencil) for grades 3 and 8 for Reading and grades 5 and 10 for Mathematics.

The probability that someone with a below-proficient score of x is really proficient is

$$p(x < c \mid \theta > c^*) = \frac{\int_c^{\infty} p(x \mid \theta) f(\theta \mid \mu, \sigma) d\theta}{\int_{-\infty}^{\infty} p(x \mid \theta) f(\theta \mid \mu, \sigma) d\theta},$$

where c is the cut score required for passing in the same, assigned metric; θ is true ability in the true-score metric, μ is the mean, and σ is the standard deviation of the population distribution; and c^*

is the cut score in the true-score metric. We assume that the transformation from c^* establishes c as a fixed point. The function $p(x|\theta)$ is the probability of the particular pattern of responses given the item parameters and θ, and $f(\theta)$ is the density of the proficiency θ in the population. These cases are the false negatives—examinees who are wrongly deemed not proficient.

The corresponding formula for false positives—examinees who are wrongly deemed proficient—is

$$p(x > c \mid \theta < c^*) = \frac{\int_{-\infty}^{c} p(x \mid \theta) f(\theta \mid \mu, \sigma) d\theta}{\int_{-\infty}^{\infty} p(x \mid \theta) f(\theta \mid \mu, \sigma) d\theta}.$$

Each observed score was regarded as a random sample from a population distribution, assumed normal with a location and scale parameter. However, because the parameters of the population distribution (μ, σ) are unknown *a priori*, they were estimated from the observed data using the following:

$$L_i(\mu, \sigma \mid z_i) \propto \int p(z_i \mid \theta) f(\theta \mid \mu, \sigma) d\theta$$

where θ represents the proficiency score, z_i is a vector of item responses for subject i, μ is the population mean of the proficiency distribution, and σ is the standard deviation of that distribution. The estimates of μ and σ were obtained by maximizing the log-likelihood function summed over observations.

Table 1:
True Score Means for Each Grade by Subject by Form Assessment

	Full TESA		Synthetic Short TESA		Paper-and-pencil	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Grade 3 Reading	0.592	1.13	0.663	1.1	1.811	1.101
Grade 8 Reading	2.919	0.926	3.026	0.887	3.568	.921
Grade 5 Mathematics	1.793	1.013	1.845	1.001	2.688	.951
Grade 10 Mathematics	3.202	1.037	3.137	1.037	3.947	0.942

Because Oregon uses the Rasch model for the paper-and-pencil test, the total number correct is a sufficient statistic for a particular test form and characterizes student proficiency. In other words, the likelihood function for equivalent number-correct patterns is maximized in the same location irrespective of which items were answered correctly. Therefore, for the paper-and-pencil test, it is necessary to compute only the classification accuracy for each possible number-correct score and not for each student individually on the paper-and-pencil test.

Because TESA presents each student with a set of items most accurately aligned with his or her ability level, the item parameters vary for each of the N students tested. Equivalent number-correct scores do not yield a likelihood function that is maximized in the same location. Consequently, even though the item parameters are based on the Rasch model, it is necessary to estimate the classification accuracy for each student conditional on the items presented to that student. To facilitate the reporting of the results, classification accuracy estimates were averaged at each similar RIT score as follows for the false positives and negatives, respectively:

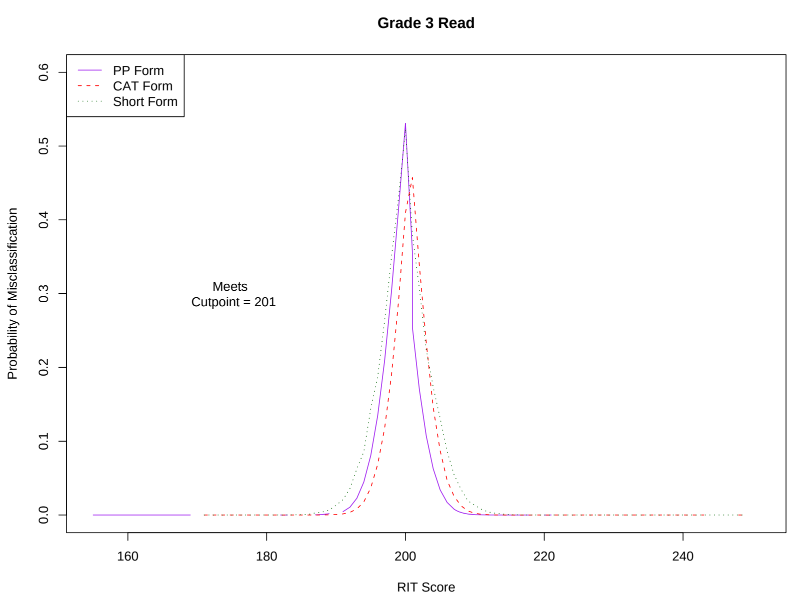
$$n^{-1}\sum_{i=1}^{n} p_i(x > c \mid \theta < c^*)$$

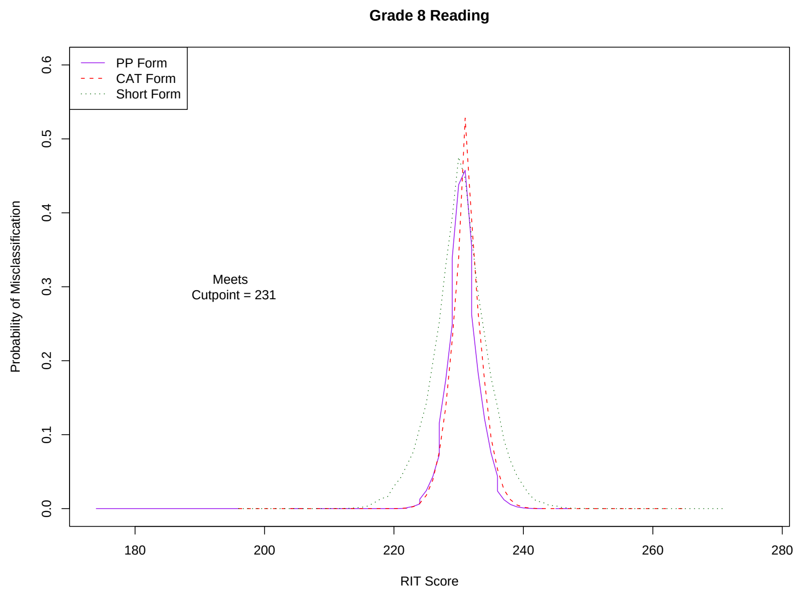
$$n^{-1}\sum_{i=1}^{n} p_i(x < c \mid \theta > c^*)$$

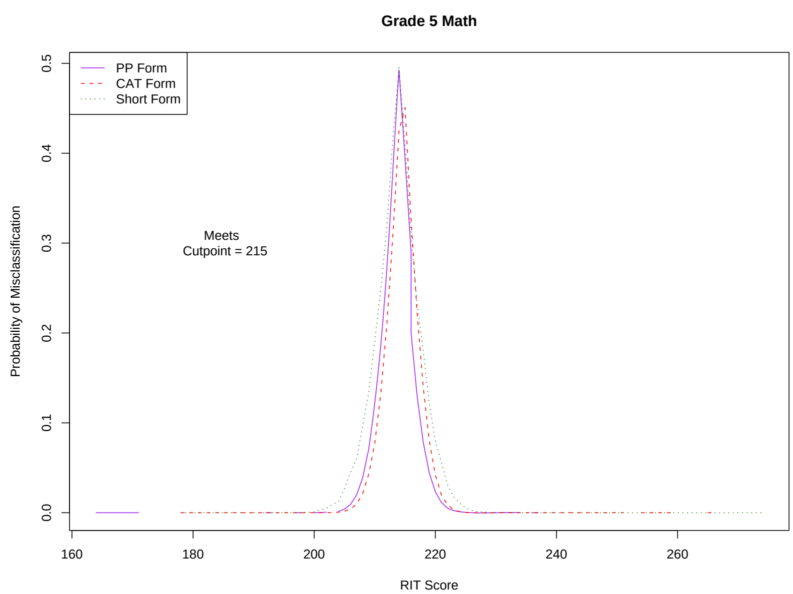
The tables in Appendix C describe the results of the analysis; the graphs below summarize the results.

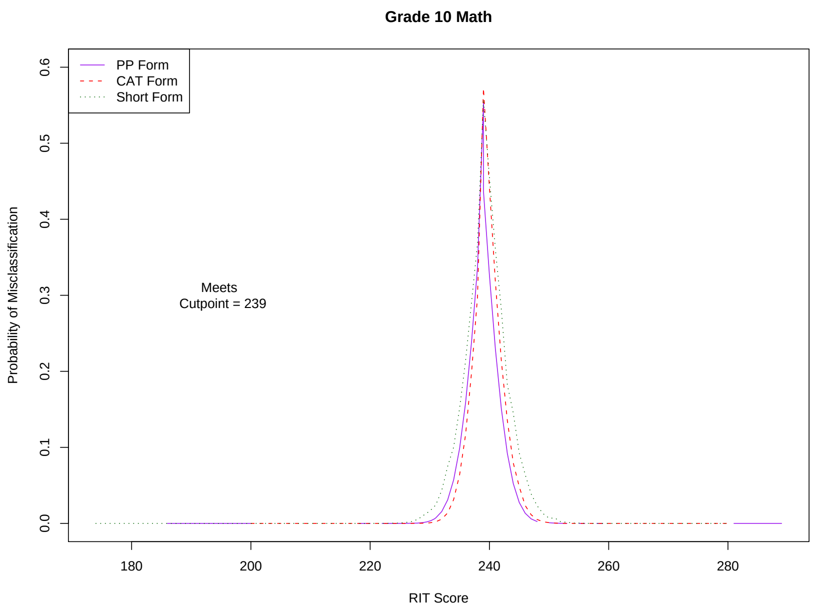
Misclassification rates to the left of the Meets cutpoint on the *x*-axis are the false negatives, and those to the right are the false positives. In all cases, the closer a student is to the Meets cutpoint, the higher the probability that the student will be misclassified because the distribution always spikes nearest this cutpoint. Conversely, the further a student scores from the cutpoint, the lower the probability that he or she will be misclassified. Note that the density curves asymptote near zero and in some cases are extremely small. However, the probability of misclassification is never truly zero.

Figure 2.
Classification Analysis Summary




Grade 8 Reading
PP Form
CAT Form
Short Form
Meets
Cutpoint = 231
Probability of Misclassification
0.0
0.1
0.2
0.3
0.4
0.5
0.6
180
200
220
240
260
280
RIT Score


Grade 5 Math
PP Form
CAT Form
Short Form
Meets
Cutpoint = 215
Probability of Misclassification
0.0
0.1
0.2
0.3
0.4
0.5
160
180
200
220
240
260
RIT Score



The figures illustrate that each test form, including the TESA, the paper-and-pencil form, and the short form (created for this study), is similar in its classification accuracy across the range of ability levels. This is evident because the plots for the false positives and negatives are superimposed in the figure across all levels of ability; providing some evidence of comparability across the TESA (long and short) and paper-and-pencil test forms.

The classification accuracy analyses also highlighted a consequence of the rounding rules used by ODE. For example, according to the grade 10 paper-and-pencil conversion tables provided by ODE, a raw score of 29 out of 54 items correct yields a RIT score of 239, which places this student at the proficient cutpoint. However, directly maximizing the likelihood for a raw score of 29 items correct results in a RIT score of 238.5, which ODE rounds to 239. As such, the maximum likelihood estimate (MLE) for this number-correct score is actually lower than the cutpoint required for scoring at proficient. Consequently, the misclassification for this number-correct score is greater than 50% as more mass in the posterior density lies below the proficient cutpoint, given that the MLE is lower than the cut score.

4.4 Strand Reliability and Precision at Cut Scores, 2005–06

Source: Doran, H., and Cohen, J. (2006) Oregon Technical Report I. Technical Report for the Oregon Department of Education.

To evaluate the precision with which the content standards are measured using the strand scores, the authors calculated a band of "indeterminacy" around the proficient cut score within each of the content strands. This may assist the state with Critical Element 7.1, which requires the state to report scores to parents in a "credible and defensible" fashion.

Although the proficiency standards were set using the total score from the entire test, the same cut score can be projected onto each content strand to evaluate a student's performance on that content strand. This analysis used the same RIT cut score that was used to determine whether a student is scoring within the Meets category for the total score.

The band of indeterminacy around the proficient score for the content strand was created using the content strand standard errors and not the standard errors derived from the total test score. These standard errors are, in general, larger than the standard errors for the overall test because they are derived on the basis of a small subset of items.

Consequently, the standard error around the strand score was used to subdivide the scale into three ranges: below, near, and above proficient. The band of indeterminacy surrounding the proficient score to create the near classification is

$$near = \theta(j)c \pm se\ \theta(j)c$$

where $\theta(j)c$ is the RIT score on the jth content strand, and se($\theta(j)c$) is the standard error associated with the content strand scaled score. When a content strand scaled score plus its standard error overlaps with the proficient cut score, it was included in the near category. Scores falling below this band were labeled below, and scores above this band were labeled above.

This analysis included scores from the TESA for grades 3 and 8 in Reading and 5 and 10 in Mathematics. Strands for Reading include Vocabulary, Read to perform a task, Demonstrate general understanding, Develop an interpretation, Examine content and structure: informational text, and Examine content and structure: literary text. Math strands include Calculation and estimation, Measurement, Statistics and probability, Algebraic relationships, and Geometry.

Appendix B provides the complete strand reliability study tables for each grade and subject. Table 2 presents the range of scores encompassed within the band of indeterminacy for Reading and Math, respectively. For each content strand, the band of indeterminacy is illustrated in gray (in the appendix) to highlight those scores that overlap with the proficient classification and are therefore indeterminate. The scores within this band are indistinguishable from proficient—we can only determine that the score is near the proficient cut score and not that it is above or below proficient.

Table 2:
Reading and Math Strand Reliability Score Range, 2005–06

Grade/Subject	Strand	Width of RIT band
Reading		
3	Word Meaning	192–209 (17)
3	Locating Information	190–212 (22)
3	Literal Comprehension	195–208 (13)
3	Inferential Comprehension	193–208 (15)
3	Evaluative Comprehension	193–209 (16)
3	Literary Forms	190–211 (22)
3	Literary Elements and Devices	191–213 (22)
8	Word Meaning	224–239 (15)
8	Locating Information	224–242 (18)
8	Literal Comprehension	224–238 (14)
8	Inferential Comprehension	224–239 (15)

Grade/Subject	Strand	Width of RIT band
8	Evaluative Comprehension	223–239 (16)
8	Literary Forms	222–242 (20)
8	Literary Elements and Devices	224–238 (14)
Math		
5	Calculations and Estimation	209–221 (12)
5	Measurement	208–222 (14)
5	Statistics and Probability	207–223 (16)
5	Algebraic Relationships	208–222 (14)
5	Geometry	208–222 (14)
10	Calculations and Estimation	228–250 (22)
10	Measurement	228–250 (22)
10	Statistics and Probability	232–246 (14)
10	Algebraic Relationships	234–244 (10)
10	Geometry	233–245 (12)

The band of indeterminacy may be useful for the state in its system of reporting subscores or Score Reporting Categories (SRCs) to parents and other levels. In particular, the federal review process asks that states report scores in a credible and defensible manner.

Currently, the state generates score reports with scores for SRCs, and the point estimates are surrounded by confidence intervals derived from the content strand standard errors.

5. Content Validity

Content validity is the degree to which an assessment measures the knowledge and skills it was designed to measure; content validity is typically determined by expert judgment. Oregon's content standards describe what student should know and be able to do. Oregon's tests were specifically developed to match Oregon's content standards. Each item is written to measure a specific category within each subject. Tests are designed so that items sample from each major category across a difficulty range that matches the full ability range of students in the state; items follow detailed test specifications to sample the domain adequately.

Evidence of content validity includes the following:

5.1 Rigorous content standards identifying what students should know and be able to do that were developed and revised with comprehensive review by Oregon educators, parents, and other citizens

5.2 Test specifications that provide a clear link between the test content and the content standards and their corresponding performance levels

5.3 A consensus-driven test item development process, using panels of educators from around the state to make judgments about the content relevance and representativeness of potential items and tasks that ensure test item faithfulness

5.4 Ongoing studies to evaluate and increase the extent that instruction, assessments, and the Academic Content Standards are aligned

The sections below describe the processes through which the state strives to meet each of these points.

5.1 Content Standards

Source: Roeber, E. D. (1996). Review of the Oregon Content and Performance Standards. Report for the Oregon Department of Education by the National Standards Review Team.

Source: Oregon State curriculum standards and areas of alignment with the PSAT/NMSQT, SAT I and SAT II. (2002). Report for the Chancellor's Office of the Oregon State University System by Educational Testing System and The College Board.

As defined in the Oregon Educational Act for the 21st Century, "academic content standards" or "academic standards" mean expectations of student knowledge and skills in identified content areas adopted by the State Board of Education under ORS 329.045.

As called for in the law, the Academic Content Standards went through a comprehensive review by Oregon educators, parents, and other citizens. They represent a broad-based consensus about the important academic goals for K–12 education in the state and were adopted by the State Board of Education in 1996 after an extensive public review process. Over 60,000 copies of the draft content standards were sent to every school in the state. They were reprinted in many local newspapers and discussed at 39 public hearings conducted across the state. More than 1,100 written responses were received, and these were used to revise the draft content standards again. Over 100,000 copies of the draft approved for first reading by the State Board of Education were circulated across the state for review.

The standards delineate the content of both instruction and assessment. They can be found at

http://www.ode.state.or.us/teachlearn/real/standards/.

Before State Board adoption, a National Standards Review Team, composed of experts in curriculum and assessment, was invited in June 1996 to comment on the content and performance standards under development. The review provided an external check on (1) the extent to which the Oregon standards are comparable to standards under development elsewhere; (2) the extent to which they can be (and should be) taught to all students; and (3) the manner in which they can be assessed. The panel praised Oregon on several aspects of the standards, commending their extensive development process, the coherence of the standards, and the collaboration among educators. The National Standards Review Team presented its findings to the State Board of Education, and these suggestions were incorporated into the standards.

- The Reading benchmarks were aligned to content on the PSAT/NMSQT and SAT I verbal tests. The Literature benchmarks were covered by the SAT II Literature subject test.

- The Writing benchmarks aligned extensively to the SAT II Writing test. The single topic omitted from the benchmarks (appropriate use of resources) was assessed through classroom work sample requirements.

- The Math benchmarks and strands are comprehensively aligned with the PSAT/NMSQT, SAT I, and SAT II (Levels IC and IIC) Math tests.

- The content specified by the Science benchmarks is almost completely covered in the SAT II Science subject tests (Biology, Chemistry, and Physics), with the exception of Earth and Space Science, which were present in the content standards but not assessed by the College Board tests.

The authors recognized that although the College Board tests covered advanced content outside of the scope of the content standards, there was good to excellent alignment between the standards (10th grade benchmarks) and the College Board tests.

5.2 Test Specifications

Test specifications define how the content standards are to be assessed (e.g., multiple choice, state performance assessment, local work sample), provide further specificity to the skills and knowledge expected of students, and convey to teachers what they can expect on state assessments. Test specifications provide a clear link between the test content and the content standards and their corresponding performance levels.

A general overview of test specifications for each test is found on the Department's Web site and contains the number and percent of items in each Score Reporting Category (SRC). Each SRC has items with approximately the same range of difficulty across reporting categories. Detailed test specifications for the Oregon Statewide Assessment are available on the same page at

http://www.ode.state.or.us/search/page/?=496.

By serving as the foundation for test development, the specifications place boundaries around the domains for which score inferences are desired. For example, scores in inferential reading comprehension are intended to represent the ability to draw conclusions from a reading passage.

The test specifications for all fixed forms and item pools are derived from Oregon's Academic Content Standards. Content specifications are determined for each standard, benchmark, and indicator relevant to the Academic Content Standards for each of the assessments covered in the test. These specifications include the appropriate number and types of items relating to the academic standards. Later, the specifications are linked to Performance Level Descriptors (PLDs), developed to help educators and test developers understand how the academic standards manifest in student performance at different levels of achievement.

TESA item pools and fixed-form paper-and-pencil tests are developed according to the test specifications. Item pool and fixed-form descriptions are available in the Annual Technical Report, Appendix Tables G.1 and H.1–56 available at http://www.ode.state.or.us/search/page/?id=787. A study evaluating the match between the TESA tests and the test specifications is summarized below in Section 8.6.

5.3 Test Development

One particularly powerful source of support for these intended interpretations of test scores is documentation that each test item addresses knowledge or skill required to achieve the content standards. Items are developed to measure these academic standards, per the content specifications. The Joint Standards (AERA, 1999; see pages 11–12 in particular) underscore the importance of this type of content evidence of validity.

Test item faithfulness to the Academic Content Standards is strengthened by using subject matter experts to design the test frameworks, blueprints, and specifications and to write the test items. Expert opinion is an important tool in assuring the educational community that the Academic Content Standards implemented in the classroom are represented faithfully on the test and that the entire student population and all subgroups of the population have had the opportunity to learn what the tests require of them (AERA et al., 1999).

A large team of qualified item writers, including current and former teachers, educators, professional test developers, and experts in educational measurement, create these items. The Oregon Department of Education uses panels of educators from around the state to make judgments about the content relevance and representativeness of potential items and tasks. Judgments by subject matter experts about the content of test questions are a vital part of the test development process and are gathered at multiple stages of the test development process.

Each subject area has a content panel, made up of about 36 teachers and curriculum specialists, representative of the state geographically and demographically. Each panel meets two or three times per year to review items written by professional test developers under contract with the Department. Items are accepted, rejected, or modified by the panel to make sure that they represent the constructs embodied in the content standards and test specifications.

The expert judgment provides evidence that test items adequately sample the desired content domain. Moreover, that judgment also provides information about potential threats to validity, including construct-irrelevant elements in the test items.

A Fairness and Sensitivity Committee is also empanelled to provide input on potential bias in the test content. The fairness and content panels revise or delete items as they deem necessary. Any items that survive this rigorous examination become part of the field-test item pool. Field testing provides data on the psychometric characteristics of the new items (e.g., item discrimination, item difficulty) in order to make the final selection of items for the operational assessment based on their ability to elicit the skills called for in the test specifications.

Following field-test administration, field-test items and forms are analyzed and submitted to a second round of reviews prior to selection for the operational test item bank. Items typically are rejected at the field-test stage if they have a point-biserial correlation of less than 0.20 or if the percent of students answering correctly is too low or too high for the targeted grade (difficulty) level or if they detract from internal consistency reliability (e.g., due to low item-test correlation).

Psychometricians review all flagged items. Finally, the expert judgment provides the means of interpreting the statistical results from field tests. In addition to judgments of content relevance, the panels appraise the technical quality of items, looking for items that are free from such flaws as (1) inappropriate readability level, (2) ambiguity, (3) inappropriateness of keyed answers and/or distractors, (4) unclear instructions, and (5) factual inaccuracy.

When necessary, the content and fairness and sensitivity review committees reevaluate the field-test items in the context of each item's statistical performance.

An assessment specialist from ODE selects items from this pool to meet the requirements described in the test specifications. For each form or item bank, bookmaps are developed to document appropriate content coverage across forms.

When coupled with expert judgments of content representativeness based on clear, explicit test specifications, empirical evidence of response consistency provides strong validity information. Volume 2: Test Development (available at http://www.ode.state.or.us/search/page/?id=787) describes the item and test development process. It shows that items address Oregon content standards, have been approved by subject matter experts, and have been vetted with representative committees.

5.4 Alignment of TESA Item Banks and Fixed-Test Forms to Content and Performance Standards

Source: Measuring up: A report on education standards and assessments for Oregon. (2000). Technical Report Prepared for the Oregon Department of Education by Achieve, Inc.

Source: North Central Regional Educational Laboratory, Surveys of enacted curriculum. http://www.secsupport.org

In 2000 Oregon was one of the first states to take advantage of the Achieve, Inc. Benchmarking Initiative. This process provided an external review of Oregon's standards and assessments, including the alignment between the two. The final report *Measuring Up: A Report on Education Standards and Assessments for Oregon* described in detail the results of the study, which included the following:

- Oregon's standards were measurable, clearly written, and jargon-free. They did not attempt to dictate pedagogy, and test specifications for Reading, Writing, and Mathematics clarified expectations and sufficiently linked the standards and assessments.

- The organizational structure of the standards could be simplified. Despite a helpful side-by-side format, the progression of knowledge and skills was not always clear, and the standards tended to repeat content across several grades. When compared to the benchmark standards, the Math standards could be made even more rigorous, as could the expectations for early literacy, algebra, and geometry.

- The English and Math standards measured important content and skills and overall; Oregon's assessments were praised as doing a good job of measuring the content and performances described by the standards.

- The assessments measured some of the objectives in the standards better than others and were not rigorous enough for some grade levels. The tests adequately measured the foundation skills although they did not always assess more cognitively demanding content. Suggested improvements included increasing the rigor of the tests and ensuring that they grow progressively more demanding as students grow older.

The results of this review influenced the revisions of the English and Mathematics content standards adopted in 2002 and 2003, as well as the methods the state used to ensure alignment of the assessments. As the assessments evolved, they were developed and revised to address the issues this study uncovered, including the following:

- Increasing the correspondence between test items and the specific content knowledge and skills laid out in the revised content standards

- Adding additional items assessing more cognitively demanding skills and concepts in addition to foundation skills

- Writing items that were comparable to the rigor and depth of the benchmark standards and that assessed more rigorous skills as children advanced grade

Additionally, Oregon uses CCSSO's Surveys of Enacted Curriculum (SEC, see http://www.ccsso.org/projects/Surveys_of_Enacted_Curriculum/ and http://www.ode.state.or.us/search/results/?id=235) to monitor and improve the state's alignment of instruction to content standards and assessments (paper and pencil and adaptive) on an ongoing basis.

The SEC offers a practical method for collecting, reporting, and using consistent data on the alignment between standards, assessments, and instruction. Surveys are available for English Language Arts and Reading, Mathematics, and Science (K–12). The survey instruments and reporting tools provide an objective approach for schools, districts, and states to analyze instruction in relation to content standards.

The SEC provides ongoing reports describing the match between content standards and assessments in both content coverage and cognitive processes, allowing for the continual monitoring and strengthening of the alignment between the instruction, standards, and assessments.

6. Concurrent Validity

A basic concept of validity is that persons who score high on a test should score high on other measures of the same construct. To the extent that two measures address the same latent construct, scores for the same individuals should agree. Conversely, a lack of relationship with theoretically unrelated measures helps substantiate the meaning of the test score. The extent to which related measures are correlated with the test scores and support, or contradict, state assessment scores validate the measure of academic achievement for the intended purposes.

Construct validity is evidenced by relating the student scores on Oregon's tests with scores on other tests measuring the same construct. Construct validity includes the following:

6.1 Oregon's test scores are highly correlated with test scores from a nationally normed test—the California Achievement Tests.

6.2 Oregon's test scores are highly correlated with scores from the Iowa Test of Basic Skills.

6.3 Oregon's tests scores are highly correlated with scores from the Northwest Evaluation Association (NWEA) subject tests.

6.4 Oregon's Reading tests are highly correlated with the Lexile Scale.

To the extent the scores on the tests measure similar constructs, scores for the students should agree. All studies show that results generalize beyond the state's assessments. Table 3 summarizes the state's evidence of concurrent validity, and the sections below provide details of the studies.

Table 3.
Concurrent Validity Summary: Correlation of Scores on Oregon Statewide Assessment and California Achievement Test (1992), Iowa Test of Basic Skills (1998), NWEA Subject Tests (2003–04), and Lexile Scale (2004)

	Reading				Math		
Grade	**California Achievement Test**	**ITBS**	**NWEA**	**Lexile**	**California Achievement Test**	**ITBS**	**NWEA**
3	.75	.78	.73	.77	.74	.76	.66
4	.77		.78		.75		.83
5	.80	.84	.81	.76	.75	.85	.76
6	.77		.71		.75		.83
7	.80		.82		.75		.84
8	.77	.82	.80	.73	.80	.84	.80
10	.76		.76	.77	.78		.82

All correlations are significant at $p<.001$.

6.1 Correlation with California Achievement Tests

Source: Oregon Department of Education. (2001) *Technical Digest—Validity—Oregon Statewide Assessment Program*.

In 1992, the Oregon Department of Education was granted approval by the U.S. Department of Education to substitute its own state tests for nationally normed achievement tests for the purpose of evaluating Chapter 1 (now Title 1) programs. The criteria for such approval included the demonstration of a strong correlation between the state test and a nationally normed achievement test.

The Department's application for use of state tests in Chapter 1 evaluation included documentation of concurrent validity of the state assessment with the California Achievement Test. Across grades 3 through 9, the average correlation between the two measures was 0.77 (ranging from 0.74 to 0.80).

6.2 Correlation with Iowa Test of Basic Skills

Source: Oregon Department of Education. (2001) *Technical Digest—Validity—Oregon Statewide Assessment Program.*

A similar study was conducted in 1998 to determine the degree of correlation between the Iowa Test of Basic Skills (ITBS), Form K, and the state Reading and Mathematics tests. A random sample of 200 students at each grade level (3, 5, and 8) who took both the ITBS and the state assessment was selected. Correlations were high, ranging from .76 and .85.

6.3 Correlation with NWEA Tests

Source: McCall, M.S. (2005). Technical review of the Oregon Department of Education assessment system. Northwest Evaluation Association.

A third study using 2003–04 data by the Northwest Evaluation Association examined the relationship between the OSAS Math, Reading, Writing, and Science tests and NWEA-developed Math and Reading tests in use by some Oregon districts as an additional assessment tool. The study found high correlations in scores for the ODE and NWEA tests in the same subject; the correlations were consistently high but tended to increase with grade. Table 3 above shows the results for Math and Reading.

Appendix G provides the correlations across all subject tests for grades 5, 8, and 10 for Oregon's Math, Reading, Writing, and Science subject tests. In general, between subject correlations are lower than correlations between tests within the same subject, supporting the discriminate validity of the tests.

Results indicate that science has a strong relationship to both Reading tests, even stronger than the relationship between writing and reading. This may indicate a need to scrutinize science items for language complexity. That said, science reading is demanding, with a difficult vocabulary. Based on these results, ODE is evaluating the language of items used in the Science tests.

Writing and Reading tests are highly correlated at all grades, showing that language arts skills are linked to one another. Except for the science pattern noted above, moderate correlation with other subjects indicates that literacy is part of every subject. Overall, these patterns show that ODE tests are functioning as expected. The subject constructs generalize to external tests, and performance does not appear to be dependent on the ODE platform.

6.4 Correlation with Lexile Scale

Source: Oregon Department of Education. (2001) *Technical Digest—Validity—Oregon Statewide Assessment Program.*

Another recent study (2004) equated the Oregon Reading test to the Lexile scale, a measure of reading ability (or text difficulty), and involved over 3000 students in four grades who took both tests within a two-week period. The high correlations demonstrate that the state tests measure content and skills similar to other widely accepted achievement tests measuring the same construct, providing additional evidence that scores on the state test correlate to other measures of similar constructs.

7. CRITERION VALIDITY

The Oregon Department of Education has developed a set of performance standards in the form of minimum test scores for meeting and exceeding Oregon content standards. The purpose of the criteria is to qualify students for the Certificate of Initial Mastery. The Oregon Education Act for the 21st Century assumes that meeting standards or earning the CIM indicates that students are ready for or likely to be successful in coursework or jobs.

To investigate the relationship between performance on state tests and in real-life settings, the following evidence is provided:

7.1 Oregon state assessment scores should relate to performance in the first year of state four-year and community colleges, providing evidence of a clear association between performance on state tests and mastery of the CIM standards and performance in the first year of higher education.

7.2 Students who meet proficiency criteria should be significantly more likely to meet criteria for employment eligibility, as determined by positive relationships between proficiency on the state tests (and mastery of the CIM standards) and pre-employment and pre-apprenticeship tests.

7.3 Test scores should not be highly associated with constructs or content that they are not intended to measure. Evidence of this is provided in the concurrent validity section.

7.1 Relationship Between Performance on CIM Benchmarks and First-Year College GPA

Source: Oregon Department of Education. (2003). The first year: Student performance on the 10th grade benchmark standards and subsequent performance in the first year of college. Oregon Department of Community Colleges and Workforce Development, The Oregon University System.

In 2003, the Oregon Department of Education collaborated with the Chancellor's Office of the Oregon University System (OUS) and the Oregon Department of Community Colleges and Workforce Development to explore the relationship between student performance on the statewide tests and subsequent performance in the first year of college.

Using the 1999–2000 10th-grade benchmark tests in Math, Reading, and Writing, researchers compared performance on the tests with first-year GPA overall, first-year Math GPA, high school GPA, SAT I scores, and grades in individual college courses. Participants included students taking the OSAS tests in 1999–2000 who subsequently enrolled in an OUS institution or Oregon community college as a freshman in 2001. The group included 18,601 students (6,082 from OUS and 12,519 from community colleges).

Results provide evidence that performance at the 10th-grade benchmark is associated with a student's freshman year college performance two years later. Students who met or exceeded the standards on the 10th-grade Math, Reading, and Writing OSAS tests had higher GPAs, higher SAT scores, and higher retention rates than students not meeting the standards (see Tables 4–6).

Table 4.
GPA (High School, First-Year College, First-Year College Math) of OUS Students by Performance Level on the Oregon Math Test at 10th Grade (CIM) Benchmark

OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	3.21	3.44	3.60	3.16	3.31
1st Year College GPA	2.52	2.87	3.17	2.80	2.82
1st Year Math GPA	2.37	2.65	2.97	2.50	2.63
Returned 2nd Fall	74%	80%	85%	71%	72%
SAT I	917	1067	1215	1047	1073

Table 5.
GPA (High School, First-Year College, First-Year College Arts and Letters) of OUS Students by Performance Level on the Oregon Reading Test at 10th Grade (CIM) Benchmark

OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	3.17	3.38	3.56	3.17	3.31
1st Year College GPA	2.45	2.77	3.14	2.80	2.82
1st Year Arts/Letters GPA	2.36	2.74	3.11	2.64	2.64
Returned 2nd Fall	74%	80%	83%	70%	72%
SAT I	881	1032	1186	1044	1073

Table 6.
GPA (High School, First-Year College, First-Year College Writing) of OUS Students by Performance Level on the Oregon Writing Test at 10th grade (CIM) Benchmark

OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	3.25	3.46	3.64	3.18	3.31
1st Year College GPA	2.60	2.91	3.22	2.81	2.82
1st Year Writing GPA	2.81	3.07	3.34	3.03	3.02
Returned 2nd Fall	76%	80%	87%	72%	72%
SAT I	977	1078	1169	1052	1073

Table 7.
Correlations between Oregon State Assessments and GPA in OUS Subject-specific College Courses

	Math	Reading	Writing	Problem Solving	All Tests
All Courses	.32	.32	.25	.18	.37
Math Courses	.28	.20	.14	.14	.27
Arts and Letters	.25	.25	.21	.15	.30
Foreign Language	.24	.22	.17	.13	.27
Science	.37	.36	.22	.19	.40
Social Studies	.33	.37	.24	.17	.39
Calculus	.26	.24	.17	.08	.28
English Comp	.12	.17	.18	.11	.21

Mathematics problem solving was included in the study but is no longer operational in Oregon's current system. Similar results were obtained by identical analyses conducted on the same cohort of students enrolled in community colleges.

Appendix E provides supporting tables for the study.

7.2 Relationship Between Performance on CIM Benchmarks and Pre-employment Exams

Source: Relationships between student's performance on Oregon's 10th grade Certificate of Initial Mastery (CIM) Assessments and Selected Oregon Business Pre-Employment Tests. (1999). The Oregon Business Council, The Oregon Department of Education, and The Oregon University System.

The Certificate of Initial Mastery (CIM) is intended to represent a solid core of foundational knowledge and skills. Students who are skilled in the academic foundations required by the CIM standards are expected to be well-educated and contribute to the success of Oregon's economy. Students with these skills should do well in Oregon's business pre-employment exams focusing on academic knowledge and skills.

In 1999, 250 juniors were randomly selected from three metropolitan Oregon high schools to take (1) the CIM (10th grade) Math, Writing, and Reading assessments; (2) a pre-employment exam from a large Oregon employer; and (3) a pre-apprentice exam from another large Oregon employer. The three schools were representative of the student population in Oregon.

The pre-employment exam tests reading and math knowledge, in line with those required by the CIM standards and necessary for success on the job. The pre-apprentice exam tests writing and advanced math skills, beyond those identified by the CIM standards and necessary for success at a higher paying job. Because both tests were in use at the time of the study, their exact contents could not be described in detail.

Student records of performance on the state tests were provided in binary form as passing (meeting or exceeding the standard) or not passing (not yet meeting) the state test. Students received feedback as to whether they passed the employment tests or not.

Crosstabs were used to compare student performance on the CIM to the pre-employment tests. Table 8 below presents the results.

Table 8.
Number of Students Passing Pre-employment and Pre-apprentice Exam by CIM Standards Met

	Number of students	Passed pre-employment exam	Passed pre-apprentice exam
Met CIM Standards:			
All	47	74.5	78.8
Writing	145	56.6	76.6
Math	120	63.3	66.7
Reading	152	56.6	73.0
Did NOT meet CIM standards:			
All	222	39.2	56.3
Writing	124	32.3	41.1
Math	139	30.9	54.0
Reading	115	30.4	42.6

Results indicate that students who met all of the CIM standards passed the employer's pre-employment exam more than twice as often as students who did not achieve the CIM standards.

- Students who met the CIM reading standard passed the employer's pre-employment exam nearly twice as often as students who did not achieve the CIM reading standard.
- Students who met the CIM math standard passed the employer's pre-employment exam more than twice as often as students who did not achieve the CIM math standard.
- Students who met the CIM writing standard passed the employer's pre-employment exam nearly twice as often as students who did not achieve the CIM writing standard.

Results also indicate that students who met all of the CIM standards passed the employer's pre-apprentice exam about one-third more often than students who did not achieve the CIM standards.

- Students who met the CIM reading standard passed the employer's pre-apprentice exam almost twice as often as students who did not achieve the CIM reading standard.
- Students who met the CIM math standard passed the employer's pre-apprentice exam about one-fifth more often than students who did not achieve the CIM math standard.
- Students who met the CIM writing standard passed the employer's pre-apprentice exam almost twice as often as students who did not achieve the CIM writing standard.

Performance on a test of mathematics is less closely related to performance on a test of writing. Therefore the lower probability of passing the pre-apprentice writing test is understandable. Meeting the state standards is associated with success on pre-employment exams, providing evidence of validity.

8. Comparability of Scores

Student scores should not depend on the mode of administration or type of test form. Because TESA is the required mode of testing for the statewide assessments system, comparability of scores obtained via alternate means of administration must be established as comparable to scores obtained through TESA. Evidence of comparability across test types and administration modes suggests that (1) TESA and paper-and-pencil tests are similar to each other and provide consistent results; (2) TESA long and short forms are similar to each other and provide consistent results; (3) operational item parameters (difficulty) are consistent across year in TESA; (4) TESA item pools are similar to each other and provide consistent results across year; and (5) side-by-side tests perform similarly to English tests.

The studies providing supporting evidence of these findings include the following:

Comparability of paper-and-pencil and TESA test scores:

8.1 The adaptive tests provided more information at every level of achievement than fixed-form tests. In all studied conditions the adaptive test had no more than 1% of the students imprecisely measured, compared to the fixed-form tests that had at least 6% of the students imprecisely measured.

8.2 Scores from computer-based tests are similar to scores from paper-and-pencil tests, although young students or those with limited computer experience may be at a slight disadvantage with computer-based tests, especially for Reading tests.

Comparability of long and short TESA test scores:

8.3 Long forms provide additional precision in estimates but show some evidence of test fatigue.

8.4 Standard errors for the long and short forms are similar across grade for Math and Reading. Across grade and subject, the average standard errors of the long forms of the TESA are less than 1 RIT scale point smaller than the averages for the short forms.

Comparability of TESA test scores over time:

8.5 The within-grade linkages for each of the three TESA item pools are consistent across year.

8.6 Operational item parameters (difficulty) are consistent across year.

Comparability of TESA-administered tests to the test specifications:

8.7 Tests administered via TESA are compared to the test specifications; many tests deviate from the test specifications by one or two items per content strand (SRC); a few of the tests administered by three or more items.

Comparability of alternate testing options (additional options available for standard administration) and TESA:

8.8 Evidence suggests that the scores from side-by-side tests are comparable to English language tests. Plain Language was discontinued as a separate form when Oregon implemented Universal Design in all test development and, as of 2006–07, Juried Assessments are no longer used for accountability.

The sections below describe each of the studies providing evidence of comparability.

8.1 Comparability of Fixed Form and TESA Tests

Source: Kingsbury, G.C., and Hauser, C. (2004). Computerized adaptive testing and No Child Left Behind. Northwest Evaluation Association. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

A concern involving the use of adaptive testing for NCLB purposes is the requirement that all students be tested with material that specifically addresses content standards for the grade in which they are enrolled. Oregon's item selection algorithms include a variety of constraints (outlined in the test specifications and described in detail in Volume 1, Annual Technical Report) to ensure that all items selected in an adaptive test meet the requirements of the test specifications.

The test information function serves as a valuable tool for test design, enabling developers to understand the measurement properties of a test during development. Comparing the test information for adaptive versus fixed-form tests provides information on the comparability of the forms. Ideally, the adaptive form should provide at least as much information at each score point as the paper form.

To make this comparison, four sets of fixed-form tests (two grade levels, grade 4 and grade 8, for Reading and Mathematics) and two adaptive tests were used. Each fixed-form set had the same general characteristics but consisted of two tests of different difficulty, one centered at the 35th percentile and one centered at the 70th percentile. For Reading, 40 items were selected to match specific grade-level content standards. For Mathematics, 50 items were selected to match specific grade-level content standards. Item difficulties were selected to correspond to the classic design of a wide range fixed-form test, with 36% of the item difficulties between the mean and 1 standard deviation, 9% between 1 and 2 standard deviations, and 5% between 2 and 3 standard deviations.

The adaptive tests included 424,328 and 251,399 TESA Reading test records for grades 4 and 8, respectively, and 428,661 and 368,441 TESA Math tests for grades 4 and 8, respectively, from the spring 2003 testing season. For each RIT score on each test, the mean of the standard error of measurement was calculated.

The following analyses compared the comparability of the tests:

Information Analysis. To evaluate the level of information yielded by each test across the range of student performance, test information (the reciprocal of the squared measurement error for a particular scale score) was computed for the adaptive and fixed-form tests. The calculation of the test information was done in the same manner for both tests; however, because the items in an adaptive test differ from student to student, the test information values for the adaptive tests were averaged across all students in the sample for each final test score.

Impact Analysis. The information functions provide an excellent method to describe the difference between tests, although population independence prevents fully describing the impact that differences in information might have on a specific group of students. To measure this impact, a simple statistic was developed using the relationship between the standard deviation of achievement in the population and the standard error of individual test scores. The development of the statistic proceeded as follows:

- Before testing, the best estimate of the student's achievement is the mean achievement level in the population. The standard error of this estimate is equal to the standard deviation of achievement in the population.

- Testing provides an improved estimate of a student's achievement level. The standard error of the score is reduced in a fixed relationship to the amount of information gained.

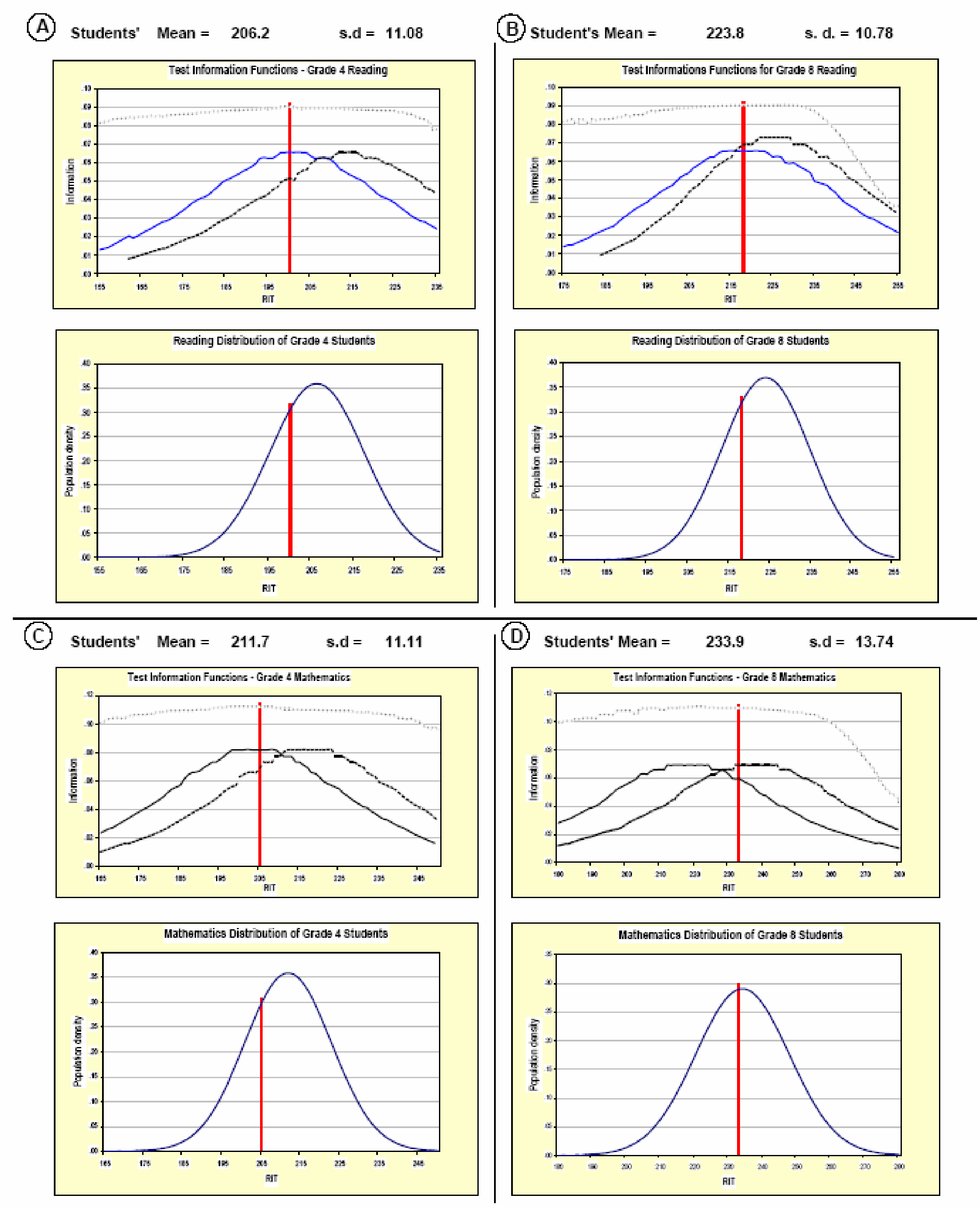
- The ratio of the standard error to the standard deviation (the standard ratio) indicates the degree of reduced uncertainty about the student's achievement level. The ratio shrinks from 1 (when the standard error equals the standard deviation, prior to testing) as we add information about the student.

- A ratio of .30 or less indicates that the test provides substantial information about the student that can be used to make meaningful instructional decisions. A ratio above .30 indicates that the test was less useful for making fine distinctions among students and their educational needs.

Calculating the percentage of students for which each test did not meet the ratio of .30 or better indicated the percentage of students for whom instructionally effective information was not available.

Figure 3 presents the information functions for all the tests examined and is divided into four panels, A–D, with two charts per panel. The top chart presents the test information functions for the particular grade level and content area, and the lower chart presents a distribution of student achievement. The achievement distributions provide context for the test information functions.

Figure 3.
Test Information Functions for Adaptive and Fixed-Form Tests in Reading and Mathematics in Grades 4 and 8



Adaptive Test
Easy fixed-form Test
Difficult fixed-form Test

The top chart in each set provides three test information functions, one for each test type. The lower two information functions, appearing as somewhat normally distributed, are for the two fixed-form tests. The function depicted in a solid line is for the test centered at the 35th percentile; the one depicted as a dashed line is for the test centered at the 70th percentile. The information function running mostly across the top of the chart is for the adaptive test. In each of the information function charts, the lowest and highest RIT scores for the subject are set approximately at percentiles 1 and 99. The vertical line in the center of the charts represents the cut score for proficiency.

The results in Figure 3 show that in each comparison the adaptive test provided the most information at all achievement levels. For students in the lower end of the achievement range (below 170 on the measurement scale), the adaptive test provided more than three times the information provided by the fixed-form test.

Table 9 presents the results of the impact (standard ratio analysis). This table shows the percentage of student scores expected to exceed a ratio of .30, based on population standard deviations from a norming study with over 70,000 students per grade level. For each subject and grade-level combination, portions of the expected student scores that exceed the criterion (hence yielding minimal information) using fixed-form tests were large enough to question the use of these tests for these populations.

Table 9.
Approximate Percentages of Students Whose Test Scores Would Be Associated with Minimal Information*

	READING			MATHEMATICS		
		Fixed Form			Fixed Form	
Grade	Adaptive	Moderate	Hard	Adaptive	Moderate	Hard
4	0.0	25.7	31.9	0.0	20.4	27.3
8	0.0	29.3	26.0	0.8	15.7	6.6

* Minimal information was defined as the standard error of measurement >= .3 standard deviations from the NWEA 2002 grade level norms.

Note that if the criterion level for the standard ratio had been changed from .3 to .25, the results for the adaptive tests would remain virtually unchanged, while close to 100% of the expected student scores on the fixed-form tests would fail to reach the desired impact ratio.

These results suggest that adaptive testing provides an information function that more closely approximates the information needed to meet all requirements of NCLB. The adaptive test provided more information at every level of achievement than either fixed-form test and more precisely measured student scores. In no condition did the adaptive test have more than 1% of the students imprecisely measured. In no condition did either fixed-form test have less than 6% of the students imprecisely measured. This difference in the impact on students' scores may become even more important as the AYP targets become more rigorous for schools.

8.2 Comparability of Computer-Based and Paper-and-Pencil Tests (2002)

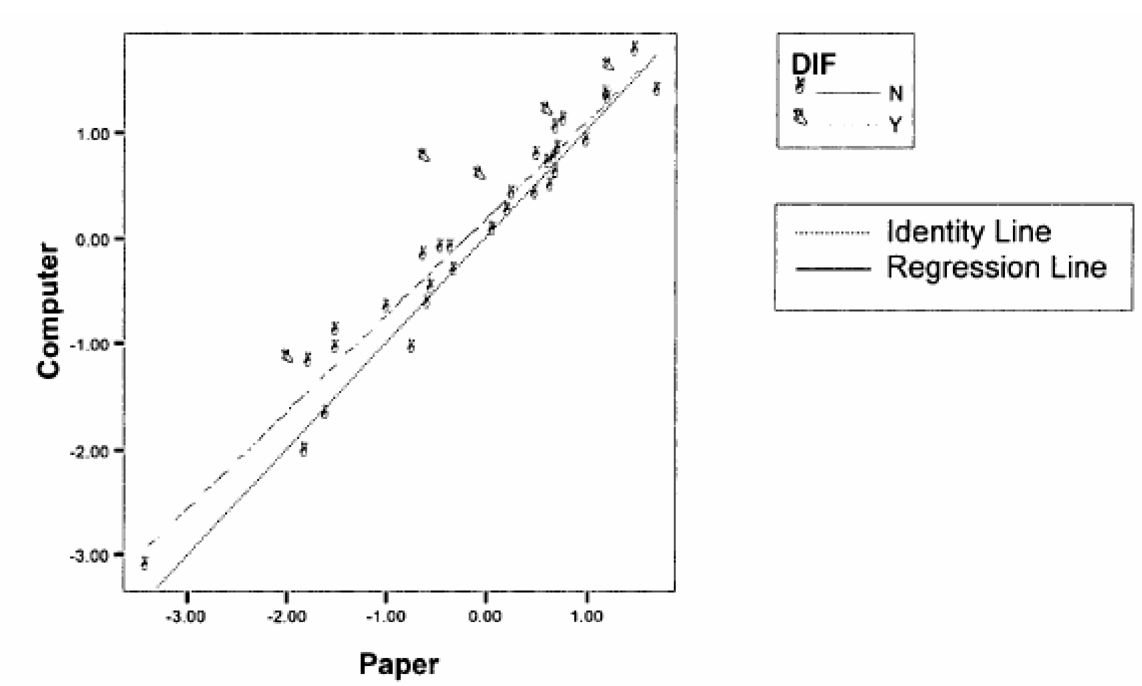
Source: Choi, S. W., and Tinkler, T. (2002). Evaluating comparability of paper and pencil and computer based assessment in a K-12 setting. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Prior to implementing TESA, Oregon established the comparability of paper-and-pencil with computer-based tests. Comparability of computer-based and paper-and-pencil tests was evaluated by administering the Math and Reading tests from the 2001 spring testing window to a sample of students. Half the items were presented as a paper-and-pencil test, and the other half were computer-based in fixed blocks of items. The order of presentation was counterbalanced at the classroom level and included approximately 800 students per subject and grade level, including third and 10th grade mathematics and reading classrooms in 14 elementary schools and 14 high schools.

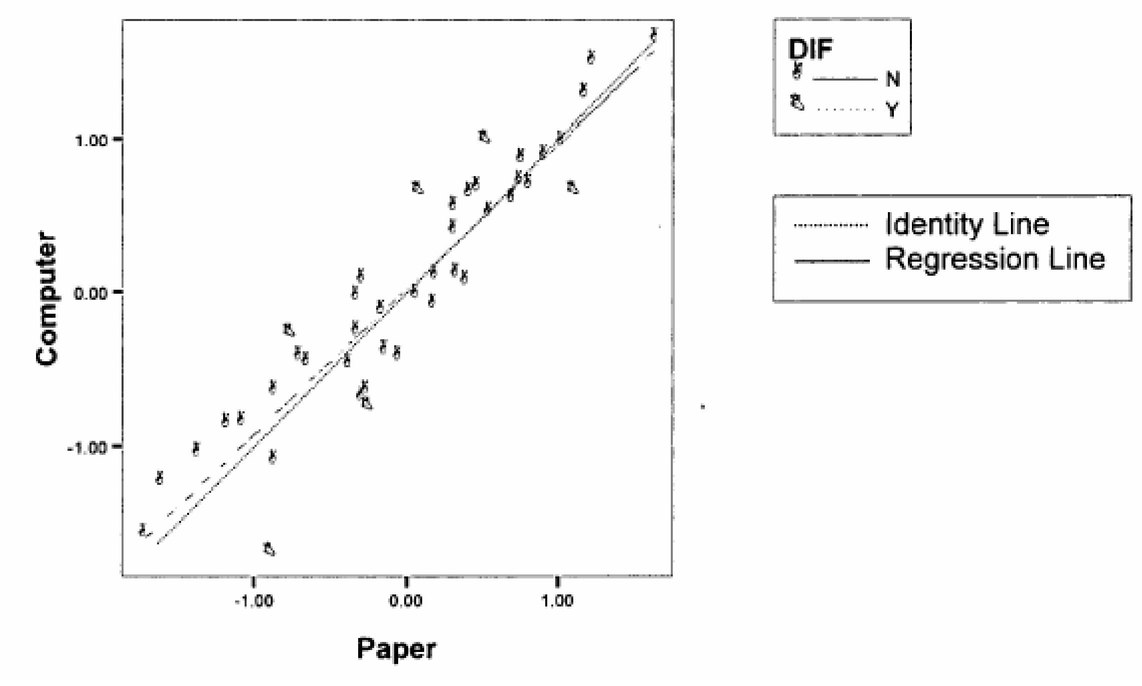
Graphical relationships between paper-and-pencil and computer item difficulty estimates are compact and linear. Figure 4 displays the scatter plots of computer and paper item difficulty estimates.

Figure 4.
Comparison of Paper and Computer Calibrations, Math and Reading, Grades 3 and 10

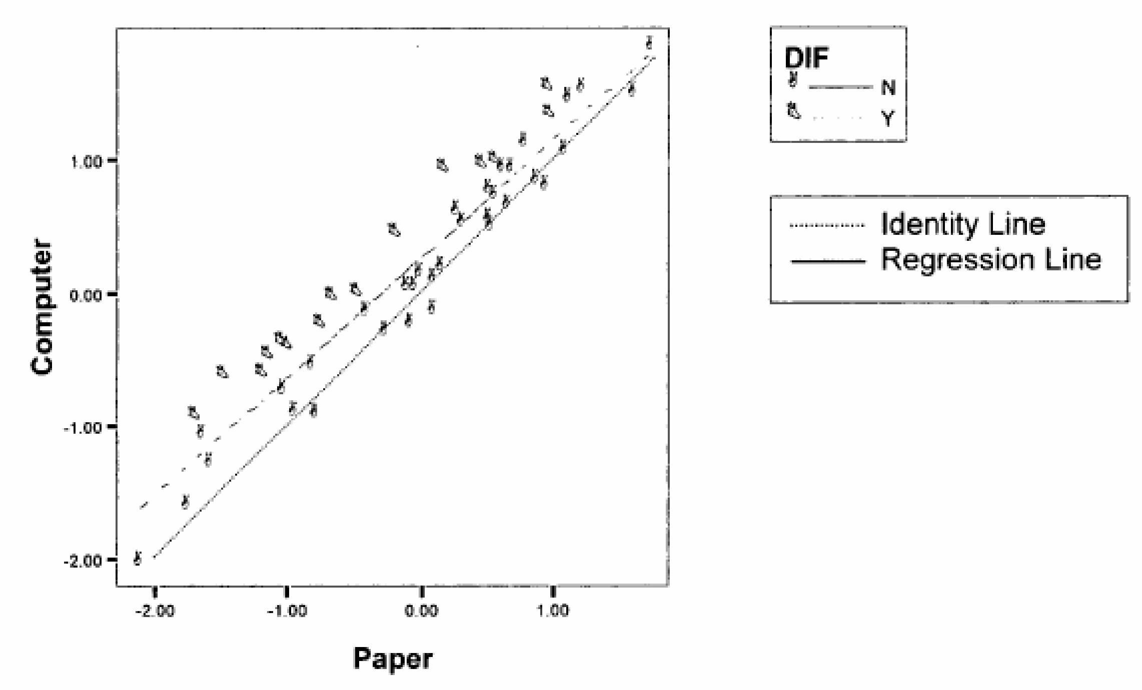
Mathematics, Grade 3



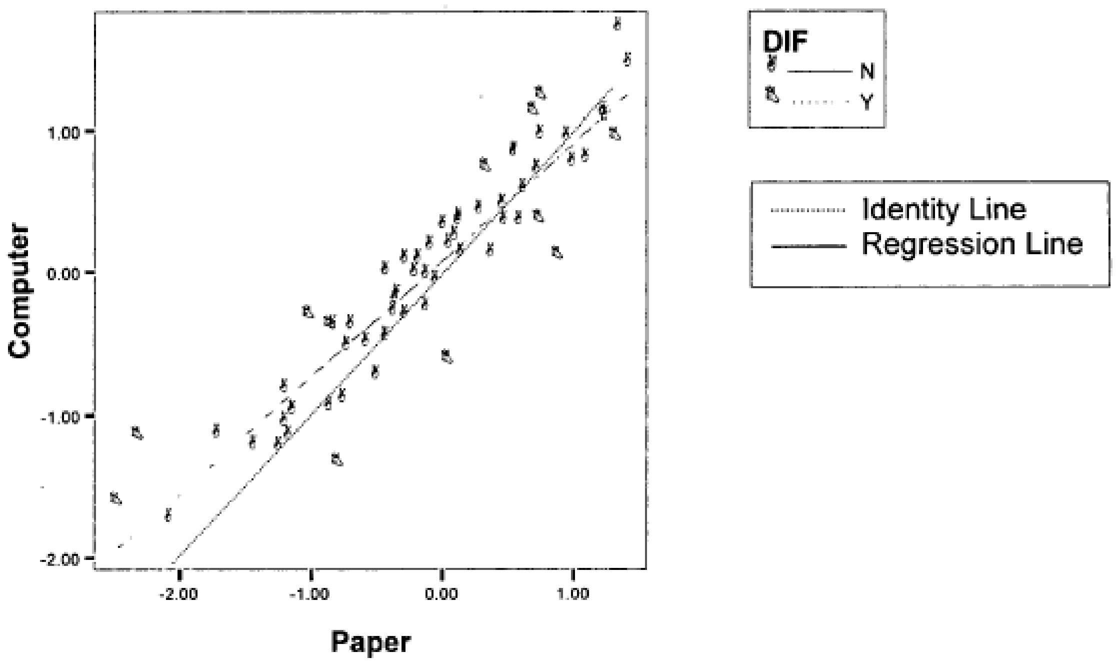
Mathematics, Grade 10



Reading, Grade 3



Reading, grade 10



The dotted line indicates the line along which identical estimates would lie; a solid line indicates the linear regression of computer-based item difficulties on paper-and-pencil item difficulties. A larger number of items were flagged for Reading compared to Mathematics for the standardized difference in item difficulties, primarily attributed to the larger offset between calibrations for the Reading items compared to the Math items.

Table 10 below shows the correlation coefficients between paper-and-pencil and computer-based item calibrations and the average item difficulties for each. The correlation coefficients were high, ranging from 0.93 for the 10th-grade Reading test to 0.96 for the third-grade Math test.

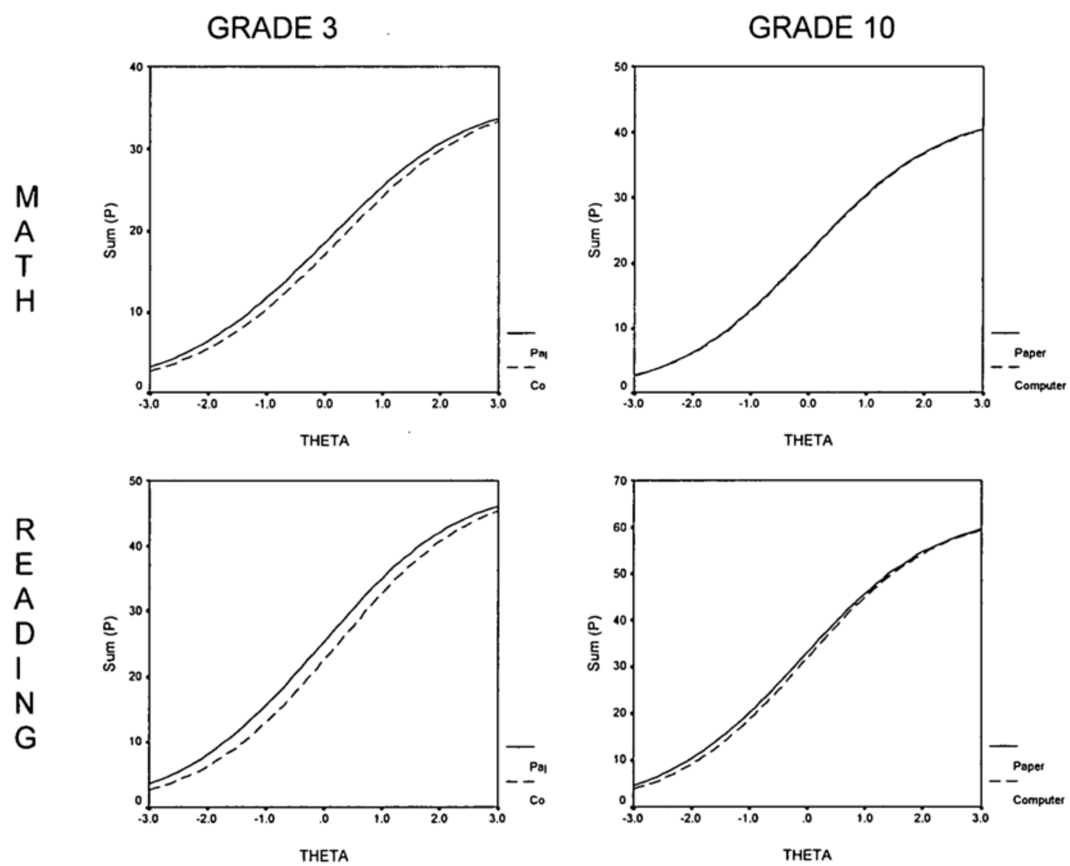
Table 10.
Comparison of Paper (P&P) and Computer (CB) Calibrations

Subject	Grade	Number of items	Correlation between P&P and CB	Mean item difficulty		
				P&P	CB	Difference
Math	3	36	0.954	-0.13167	0.07306	0.20472
	10	43	0.931	-0.01791	-0.000233	0.01767
Reading	3	49	0.956	-0.09837	0.17816	0.27653
	10	63	0.932	-0.13540	-0.03079	0.10460

There were, however, varying degrees of mode effects as measured by a mean difference in item difficulty estimates; the mean differences in item difficulty estimates were larger for third graders than for 10th graders. Furthermore, the mean differences were larger for Reading tests than Math tests, suggesting that, for third graders, most items became slightly harder when presented on computer by an average of a little over a quarter logit. There was practically no difference in mean item difficulties for the 10th-grade Math test.

The same pattern can be observed in Figure 5, which displays test characteristic curves (TCC), i.e., expected number-correct scores as a function of ability. The offset between two TCCs was smallest for the 10th grade Math test and largest for the third grade Reading test.

Figure 5.
Test Characteristic Curves



This study suggested that scores are comparable when the administration mode of the test changes, although the youngest test users who have less experience with computers may experience more difficulty with computer-based tests than with paper-and-pencil tests. Frequent exposure to online exams are likely to eventually eliminate the novelty and mode effect as will providing accommodations and tools simulating standard paper-and-pencil practices (such as allowing students to highlight selections of text in a passage).

8.3 Consistency of Item Performance Across Long and Short TESA Tests (2004-05)

Source: Doran, H., and Cohen, J. (2006). Oregon Technical Report I. Technical Report for the Oregon Department of Education.

Investigating differential item functioning (DIF) on the basis of item location generates some evidence for considering whether a fatigue effect appears to be present in the data. Significant differences between items conditional on location (if the items appear more difficult at the end of the test) may suggest that the longer form of the test pays for precision with increased bias due to test fatigue.

To investigate the extent to which items appearing at the end of the long test function differently from when they are presented earlier, a DIF analysis of the items was conducted on two "blocks" of test items. Block A comprised items appearing in positions 1 through 35; Block B comprised items appearing after position 35. Two Rasch parameter estimates were generated for each item: the first based on students who were presented with the item within Block A and the second based on students presented with the same item in Block B. Some items were sparsely presented in either Block A or B, so each block included only the items with 100 or more responses.

Data was from the 2005-06 administration and included grades 5 and 10 for math and grades 3 and 8 for reading.

The software program used to generate parameter estimates was AM, a program developed by AIR (Cohen and the American Institutes for Research, 2002) for estimating parameter estimates and standard errors in complex samples. After the parameter estimates for each item were generated, the following was used to assess DIF across the two blocks:

$$DIF = \frac{\hat{b}_{ia} - \hat{b}_{ib}}{\sqrt{se(\hat{b}_{ia})^2 + se(\hat{b}_{ib})^2}}$$

where *i* indexes item and *a* and *b* index the block from which the item was categorized based on its position. The result is a *t*-statistic and, given the large sample sizes per item, the critical value of 1.96 was used to assess whether a significant difference existed between the item parameters.

Gauging the practical significance of the statistical result is a task often reserved for a subsequent statistic referred to as an effect size measure. For this analysis, the following expression was used to determine the extent to which scores on the RIT scale would change as a function of the *b*-values obtained from the DIF analysis:

$$ES = [(\bar{b}_a - \bar{b}_b) * p_b] * 10$$

$$se(ES) = \frac{\sum_{i=1}^{K} (b_{ai} - b_{bi})^2 * p^2{}_b}{k(k-1)}$$

where $\bar{b}_a$ is the mean of the item difficulties for the items in Block A, $\bar{b}_b$ the mean of the item difficulties for the items in Block B, p_b is the proportion of items from the total test in Block B, and k is the number of test items. In the presence of no DIF, ES = 0, and no difference would be expected for student scores on the RIT scale. However, ES ≠ 0 indicates that scores differ as a function of the changing difficulty of the item parameters. If items appearing in the second block are affected by the fatigue effect and appear more difficult than those in the first block, the ES measure would be negative. An ES = –10 indicates that a student's score decreases by approximately 10 units on the RIT scale.

The results of the DIF evaluation are provided in Table 11 below by grade and subject and seem to indicate more DIF for the grade 8 Reading items than for other grades and subjects. Of the items used for this analysis, 53% (79 of 149) of the grade 3 items exhibit DIF when presented at the end of the test, whereas 65% (110 of 170) of the grade 8 Reading items were found to be more difficult when presented at the end of the test. In Mathematics, 42% (62 of 147) of the items included in this analysis were found to exhibit DIF, whereas only 34% (51 of 148) of the grade 10 Mathematics items indicated DIF as a function of item location. As with Reading, we see larger DIF at the lower grade than at the higher grade. Appendix D presents statistics for these items.

Table 11 also presents the effect sizes as approximate bias in RIT scores due to fatigue from items past position 35. In all cases, the results indicated that the current test length may be inducing a small fatigue effect on student scores. Although the results all hover near 1 RIT point, this is approximately one-third the size of the standard error of measurement (which is about 3 RIT points on average).

Table 11.
Estimated Impact of Fatigue on RIT Scores for Tests Over 35 Items

Grade/Subject	Expected Change in RIT Score	Number of items with DIF (number of items)
Grade 3 Reading	–.72 (.011)	79 (149)
Grade 8 Reading	–.98 (.011)	110 (170)
Grade 5 Mathematics	–.35 (.008)	62 (147)
Grade 10 Mathematics	-.26 (.009)	51 (148)

8.4 Comparability of Standard Errors Across TESA Long and Short Tests (2004-05)

Source: Doran, H., and Cohen, J. (2006). Oregon Technical Report I. Technical Report for the Oregon Department of Education.

To evaluate the comparability of scores obtained by long and short TESA tests, as well as to evaluate the precision gained by longer tests against the costs in terms of student fatigue, the conditional standard errors between the full TESA form and a synthetically created short test form were compared across the entire range of ability. If the short form provides the same amount of information of examinee ability as the long form, then the standard errors from the two tests will overlap at all ranges of ability.

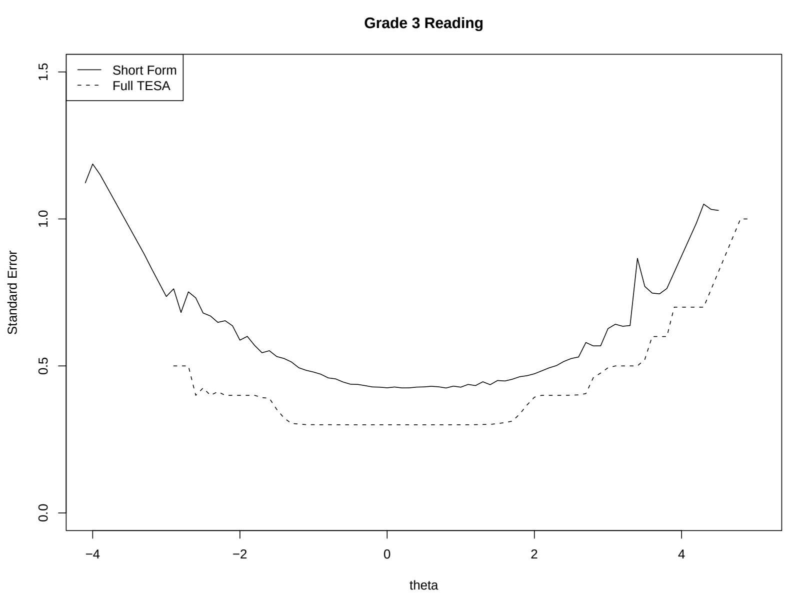
A short form was created by selecting the first 35 operational items presented to each student. Using these items, an overall RIT score and its precision were computed. The standard errors were derived from the inverse of the test information function (TIF) as follows:

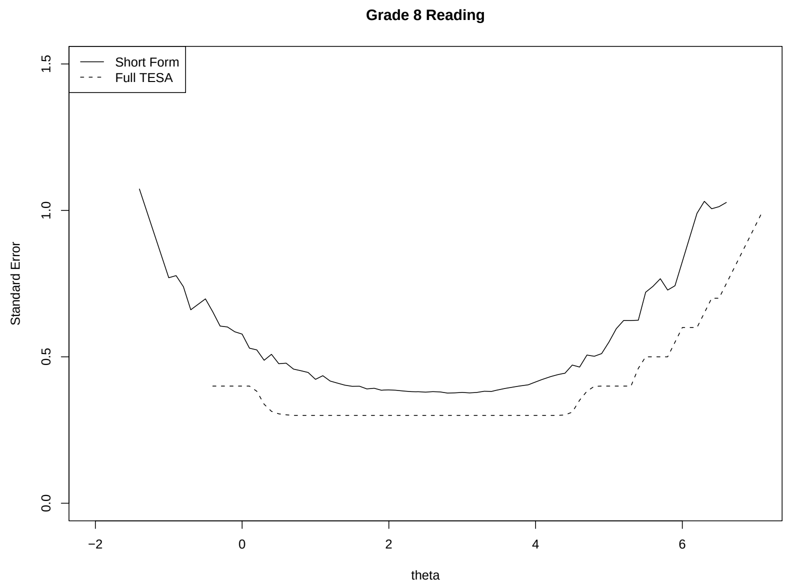
$$TIF(\theta) = \sum_{i=1}^{k} I_i(\theta)$$
$$SE(\theta) = TIF(\theta)^{-1/2}$$

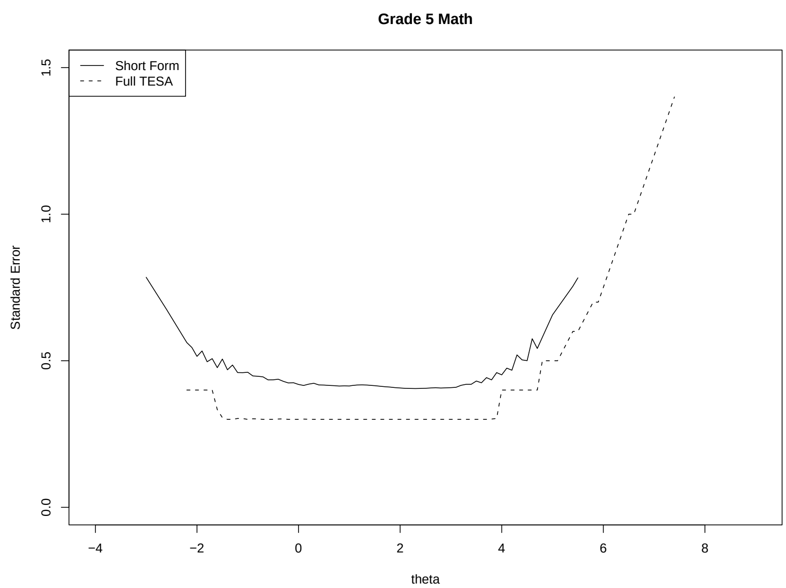
where Ii(θ) is the information function for the *i*th item. Because the TESA is adaptive, each student receives a different set of items, with each new item conditional on his or her response to the prior item; it is possible for students to obtain the same RIT score even though they are exposed to a different set of items. Consequently, the same RIT is measured with a slightly different precision and requires that we average the standard errors across equivalent RIT scores to facilitate reporting of the results.

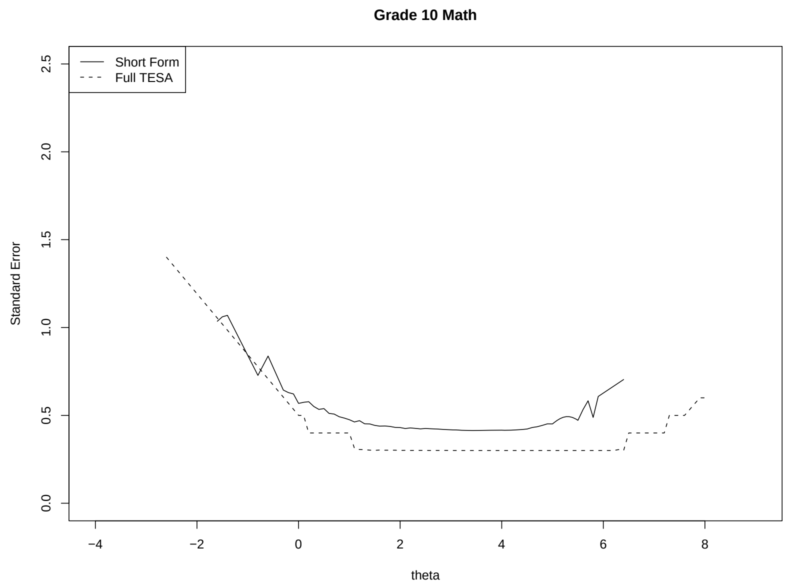
The results of this analysis are graphically presented below for each grade and subject.

Figure 6.
Standard Errors for TESA Long and Short Forms, Math (Grades 5 and 10) and Reading (Grades 3 and 8)




Grade 8 Reading
Short Form
Full TESA
Standard Error
0.0
0.5
1.0
1.5
−2
0
2
4
6
theta


Grade 5 Math
Short Form
Full TESA
Standard Error
0.0
0.5
1.0
1.5
−4
−2
0
2
4
6
8
theta



The charts illustrate the conditional standard errors for the long and short forms of the test. The results are consistent across all forms; the standard errors indicate that the short version of the test always provided less information across the ability range than the full TESA form. This is, of course, expected because a long test form will generally provide more information than a short form. Table 12 shows the average improvement in the standard error of measurement by grade and form.

Table 12.
Average Standard Error of Measurement by Form

Grade/Form	Short Form	Full TESA
Grade 3 Reading	4.6	3.2
Grade 8 Reading	4.0	3.0
Grade 5 Math	4.2	3.0
Grade 10 Math	4.3	3.0

The data in Table 12 indicate that the standard errors were approximately 1 RIT unit smaller on the long form when compared to the short form. In this case, scores are measured approximately one point more precisely, but may display up to prepare of bios due to extreme fatigue.

Similarity with the standard error curves from the paper-and-pencil forms provided in Appendix A provides additional evidence of comparability of TESA and paper-and-pencil tests. Historically, standard errors for the tests hover around 3 RIT points, and similar trends are observed in the measurement errors.

8.5 The Stability of the Within-Grade Linkages Over Time for TESA Item Pools, Short and Long (2004–05 and 2005–06)

Source: Doran, H., and Cohen, J. (2006). Analytic and Psychometric Services: Technical Report II. Technical Report for the Oregon Department of Education by the American Institutes for Research.

Each examinee should obtain equivalent scores irrespective of which item pool is used to construct his or her test form. In the equating literature, this is referred to as the principle of equity (Lord, 1980), which states that it should be a matter of indifference as to which form the student takes. One method for evaluating the consistency in scores obtained across multiple pools over time is to evaluate the error in the equating relationship, which can be evaluated with the expectation that 68% of the examinees with scores of y_1 from pool 1 should have a score of $y_2 \pm \delta$ if their form had been constructed from pool 2, where δ is the standard error of equating.

Equating error arises from two distinct sources, the sampling of common items and the sampling of students (Doran & Jiang, 2004; Michaelides & Haertel, 2003). The common items found across the pools are only a sample of the items that could have been used, and the item parameters obtained for those items were obtained from a sample of students. Consequently, there is variability in the cross-form statistical linkages that adds to the total variance in a student's score estimate.

Given that the pools share a few common items, it was possible to establish a statistical linkage across the pools using a common item nonequivalent groups design (Kolen & Brennan, 2004). However, the fundamental challenge was to estimate the variance of this linkage to establish a sense of score consistency across pools. For this analysis, the procedures of Cohen, Seburn, Antal, and Gushta (2005) were implemented.

Cohen et al. (2005) develop an analytic expression for estimating the variance of a linkage using the Rasch model in clustered samples and show that their variance estimator is consistent under joint calibration. Using AM software, the item parameters for the common items were re-estimated applying a joint calibration procedure. The size of the linking variance with respect to the total variance in student scaled scores is computed as the proportion of total variance accounted for by linking error as

$$\eta = \frac{\delta}{N^{-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}$$

where δ is the equating variance and θ is the scaled score for student *i*.

The common items from TESA-administered tests in grades 3 and 8 Reading and grades 5 and 10 Math across the 2004–05 and 2005–06 school years were used for this analysis. Data consisted of the common items in a single item pool from each grade, subject, and year (Table 14 describes the number of items in each pool) .

Table 13 provides the results of the equating error evaluation in RIT scores, with the square root of the variance (i.e., the standard error of equating) in parenthesis.

Table 13.
Equating Variance by Grade and Subject

Grade 3 Reading	Grade 8 Reading	Grade 5 Math	Grade 10 Math
.31 (.56)	1.5 (1.22)	.32 (.56)	.48 (.69)

Because the RIT scale has a standard deviation of 10 and an average standard error of measurement of approximately 3 RIT units,[1] the equating error accounted for a very small fraction of the measurement error with respect to the standard error of measurement.

The overall impact of the equating error on scores obtained from disparate pools is relatively small. From Table 10, a student with a Grade 3 Reading score of y_1 from a test form constructed using the items in pool 1 would be expected to have a score of $y_2 \pm .56$ if that same student had a test form constructed from the items in pool 2. This is a nominal difference with respect to the standard deviation on the RIT scale.

Table 14 describes the size of the linking variance with respect to the total variance in student scaled scores.

Table 14.
Total Variance and Proportion Accounted for by Linking Error

	Grade 3 Reading	Grade 8 Reading	Grade 5 Math	Grade 10 Math
Total Variance	168.38	109.39	99.13	90.85
Proportion of total variance accounted for by equating variance (η)	.18%	1.37%	.32 %	.53%

Results indicate that linking error accounts for a very small proportion of the total variance in student scores.

8.6 Consistency of Operational TESA Item Parameters Over Time (2004–05 and 2005–06)

Source: Doran, H., and Cohen, J. (2006). Analytic and Psychometric Services: Technical Report II. Technical Report for the Oregon Department of Education by the American Institutes for Research.

An important aspect in evaluating the consistency in test forms is to consider whether estimates of the operational item parameters remain consistent over time. Evaluating the degree to which item parameters remain constant over time is one source of evidence that can be used to substantiate consistency in the forms across year.

The current study evaluated the stability of common items over time. The grades evaluated included grades 3 and 8 Reading and grades 5 and 10 Math across the 2004–05 and 2005–06 school years.

[1] The TESA is adaptive, and the standard errors vary by score and instance of the test. However, the average standard error is approximately 3 units across the ability range.

Because Oregon's TESA is a progressive test format, which constructs unique test forms from a test pool within each grade, only those items that were common across the 2004–05 and 2005–06 pools are included in this analysis. Table 15 shows the number of common items across the pools for each grade as well as the number of operational items used for each administration.

Table 15.
Common Items Across 2004–05 and 2005–06 Pools

Grade/Subject	Number of Common Items	Total Number of Operational Items in 2004–05	Total Number of Operational Items in 2005–06
Grade 3 Reading	71	178	130
Grade 8 Reading	116	170	158
Grade 5 Math	47	148	133
Grade 10 Math	19	149	146

In order to evaluate the extent to which the item parameters remain stable over time, the items were recalibrated using the student responses from each year with a separate Rasch-model calibration such that a unique estimate of the item parameters was obtained for each year. Once the parameter estimates were obtained, it was necessary to recenter the item parameters because the calibrations were performed separately and consequently items were not on the same scale of measurement. This was accomplished via mean equating with the linking constant identified as follows:

$$\hat{\boldsymbol{B}} = \mu(\boldsymbol{b}_1) - \mu(\boldsymbol{b}_2)$$
$$\boldsymbol{b}_{i2}^{*} = \boldsymbol{b}_{i2} + \hat{\boldsymbol{B}}$$

where $\mu(b_m)$ is the mean of the common items in pool m and b_{i2} is the ith item in pool 2. By adding in the constant, $\hat{\boldsymbol{B}}$, items in pool 2 were placed on the same scale as the items in pool 1.

With the items now on the same scale, it was possible to proceed with a comparison of the item parameters over time. The first step was to estimate the variance of the common items across pools using the variance estimator in Equation 1 below:

$$\text{var}(\boldsymbol{L}) = \boldsymbol{N}^{-1}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}(\boldsymbol{b}_{\boldsymbol{i}1} - \boldsymbol{b}_{\boldsymbol{i}2}^{*})^2 \quad \textbf{(1)}$$

where, as before, b_{im} represents the ith common item in pool m and N is the length of the vector $b = (b_{i1}, \ldots, b_{iN})$. Note that the variance in Equation 1 is influenced by two factors: the degree to which the items behave similarly over time and the number of common items.

While the variance estimator in Equation 1 is an indicator of spread, it lacks intuitive context for evaluating whether items used in pools over time behave in a similar manner. A reliability-like statistic was used as a summary estimator for judging item consistency over time as

Error! Objects cannot be created from editing field codes. (2)

where N_m is the total number of common items in pool *m*. The variance components from Equations 1 and 2 were used to judge whether the common items accounted for an appreciable portion of the total variance as

$$\boldsymbol{r} = \frac{\mathrm{var}(\boldsymbol{T}) - \mathrm{var}(\boldsymbol{L})}{\mathrm{var}(\boldsymbol{T})}$$

This statistic is similar to the measurement of reliability in classical test theory and behaves similarly in that $0 < r < 1$. If *var(L)* accounts for a large proportion of the total variance among common items; then *r* tends to 0. If *var(L)* over time becomes smaller, then *r* tends to 1. The more consistently the items behave over time, the more likely *r* will approach 1.

Figure 7 describes the scatter plots of the item parameters for each grade/subject combination, while Appendix B provides the item parameters for all grades and subjects. The plots provide evidence that there is a strong linear relationship among the items over time, with the exception of a single grade 10 Math item (item M0206300) that behaved somewhat inconsistently across the two years.

Figure 7:
Comparing TESA Item Parameters Over Time

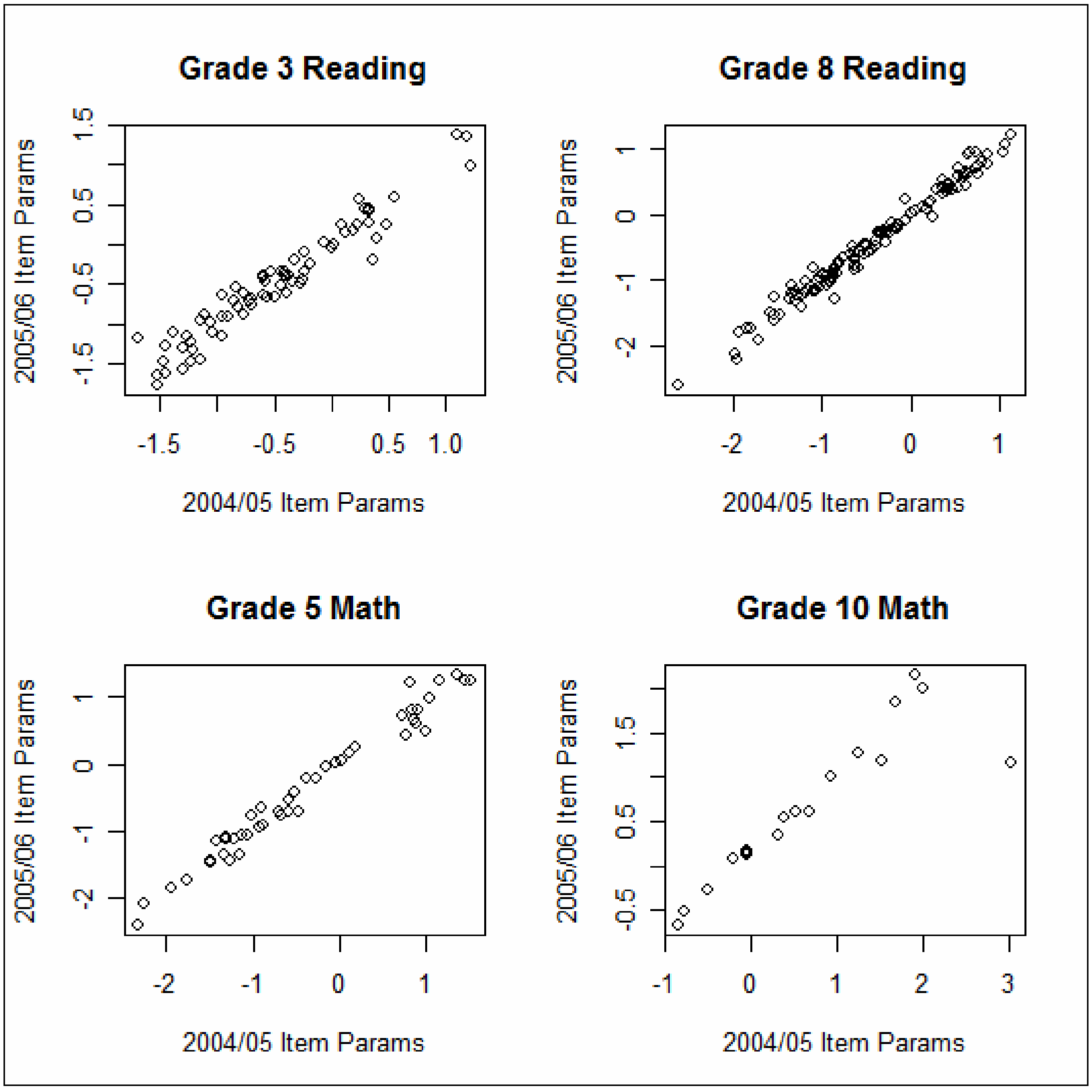


Table 16 presents the variance components for each grade/subject estimated from Equations 1 and 2.

Table 16.
Variance Components

	Grade 3 Reading	Grade 8 Reading	Grade 5 Math	Grade 10 Math
var(L)	0.03424587	0.01422786	0.03024606	0.2267902
var(T)	0.48184822	0.66712631	1.05581894	0.9162831
R	0.92892809	0.97867292	0.97135298	0.7524889

The last row of the table (*r*) is the reliability statistic summarizing the degree to which the item parameters behave over time. In grades 3 and 8 Reading and grade 5 Math, the reliabilities are near 1, indicating that the items are very consistent over time. The grade 10 Math reliability, however, is much lower than for the other grades (.75). A visual inspection of the data reveals that a single item, M0206300 (the anomalous item), is increasing the term *var(L)* and causing the reliability to decrease.

8.7 Comparability of TESA-Administered Tests, 2005–06

Source: Doran, H., and Cohen, J. (2006). Analytic and Psychometric Services: Technical Report II. Technical Report for the Oregon Department of Education by the American Institutes for Research.

Because the TESA is adaptive in nature, a unique test form may be constructed for each student participating in the test. NCLB requires evidence that the TESA-administered tests are comparable to each other, overall and at the achievement levels, and are consistent with grade-level test specifications. Evidence of comparability is provided by comparing the 2004–05 and 2005–06 TESA-administered grade 3 and 8 Reading assessments and the grades 5 and 10 Mathematics assessments to the test specifications and to each other.

Currently, the TESA algorithm is designed to select items that meet specified criteria including selection of the specified number of items from each content strand and selection of items within each content strand that increasingly approximate each student's proficiency level. Ideally, a sufficient number of items exist within the item banks for each strand for all levels of proficiency, and each test form would align to test specifications exactly.

Data were cross-tabulated and compared to the test specifications by content strand. Appendix B provides the data for this analysis, and Figures 8–11 summarize the results. In the appendix, the data are presented in two forms: (1) as the proportion of tests that meet test specifications exactly for each content strand and performance level or are off by one, two, or three or more items, and (2) as the proportion of tests that are off by zero or one item or are off by two or more items.

The figures below show the proportion of tests that meet test specifications exactly or are off, in absolute value, by 1, 2, … *N* items. A test that met the specifications exactly deviates from test specifications by zero items. Results are presented for each of the Oregon performance categories to evaluate whether there are systematic differences in the extent to which forms meet the test specifications conditional on the student's level of ability.

Figure 8:
Comparison of TESA-Administered Tests to Test Specifications: Grade 3 Reading

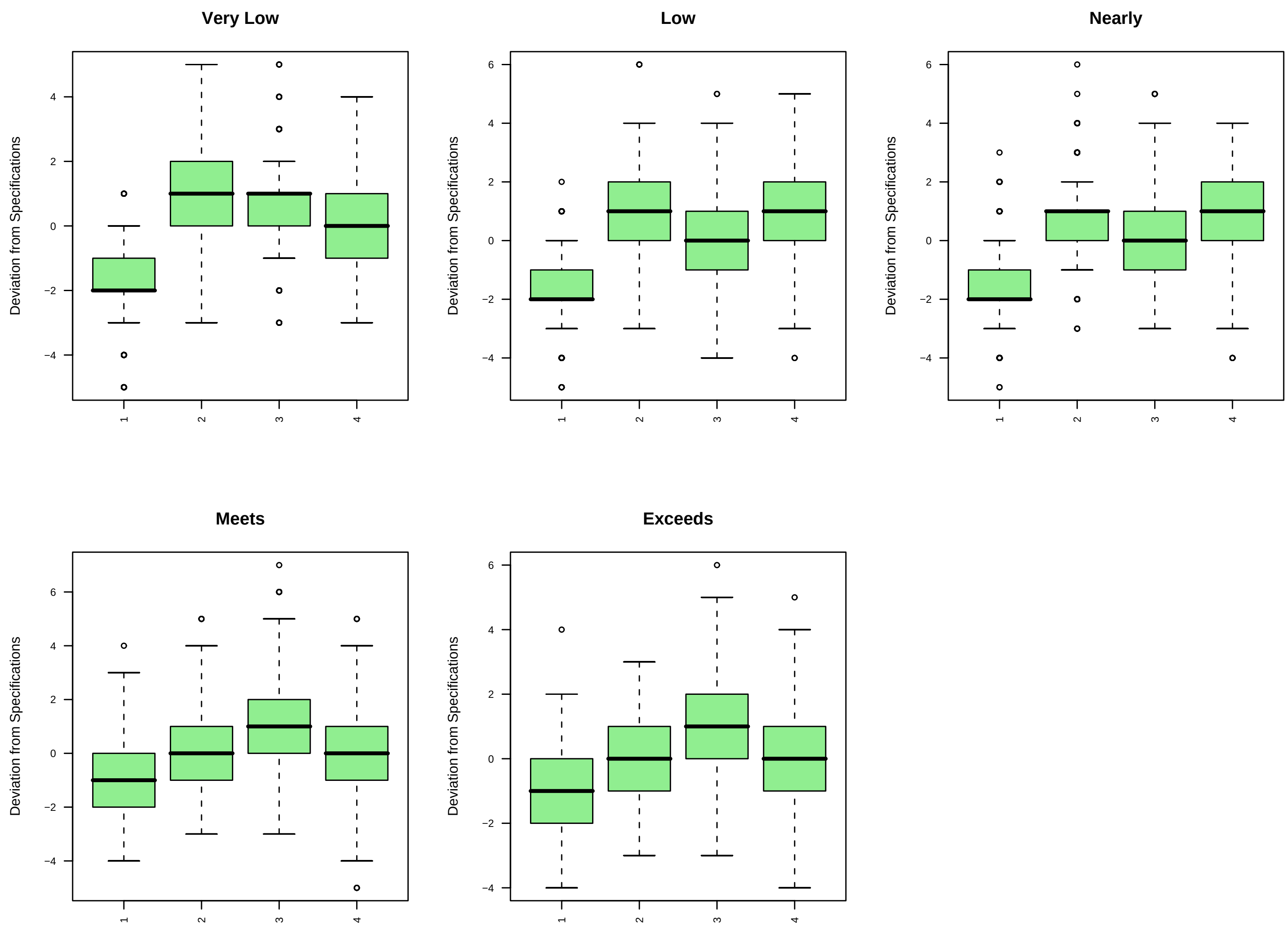


Figure 9:
Comparison of TESA-Administered Tests to Test Specifications: Grade 8 Reading

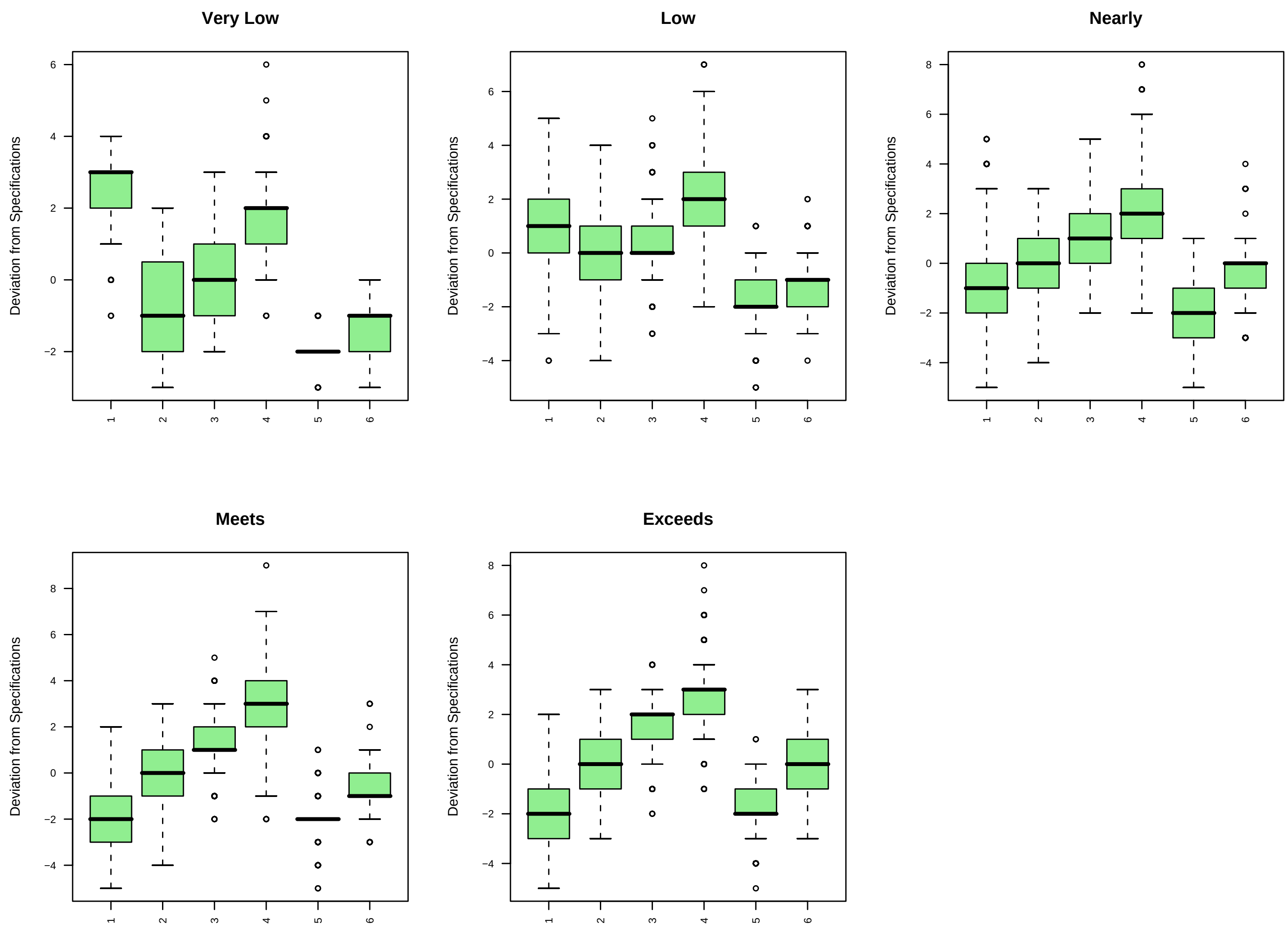


Figure 10:
Comparison of TESA-Administered Tests to Test Specifications: Grade 5 Math

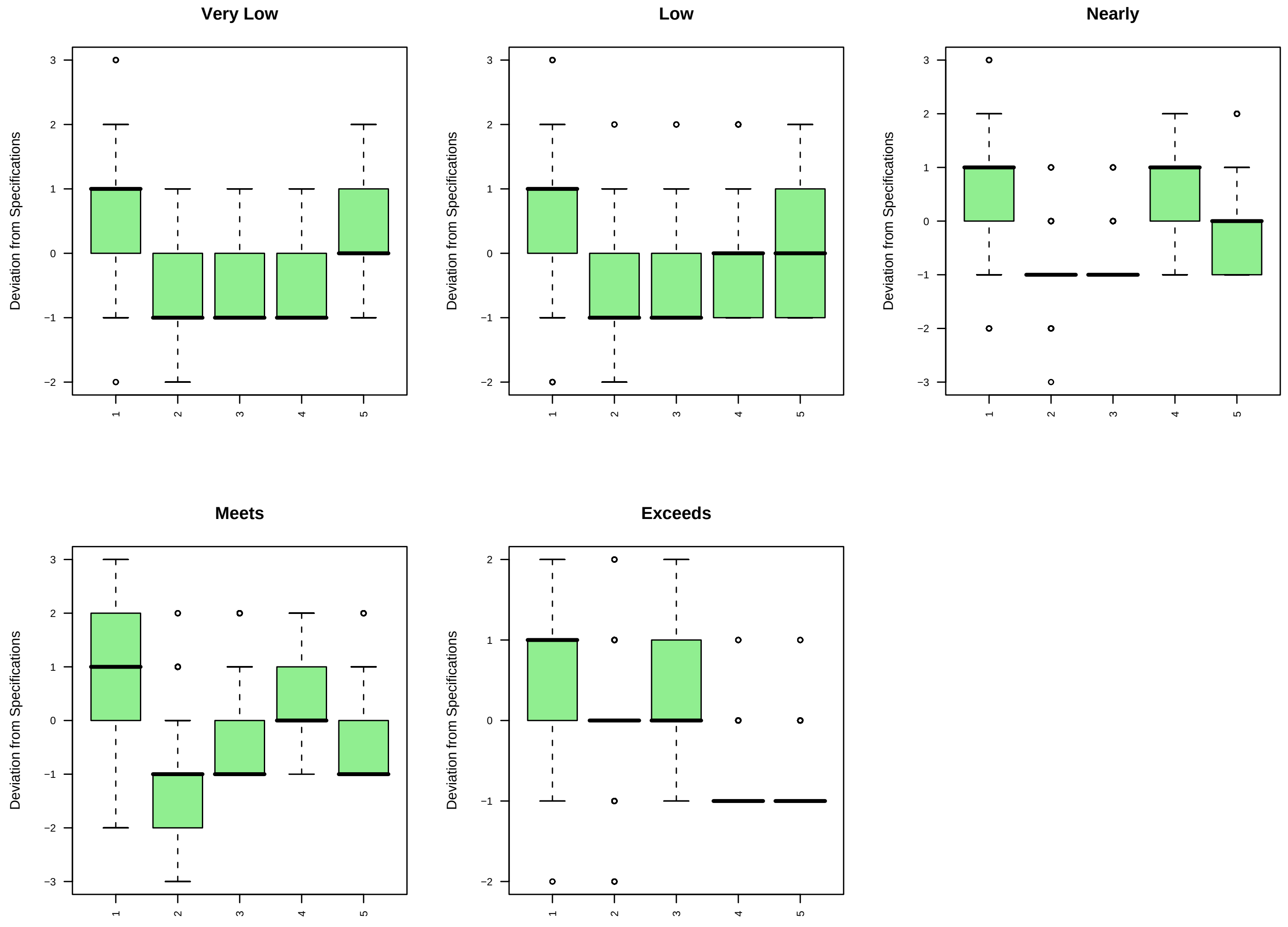


Figure 11:
Comparison of TESA-Administered Tests to Test Specifications: Grade 10 Math

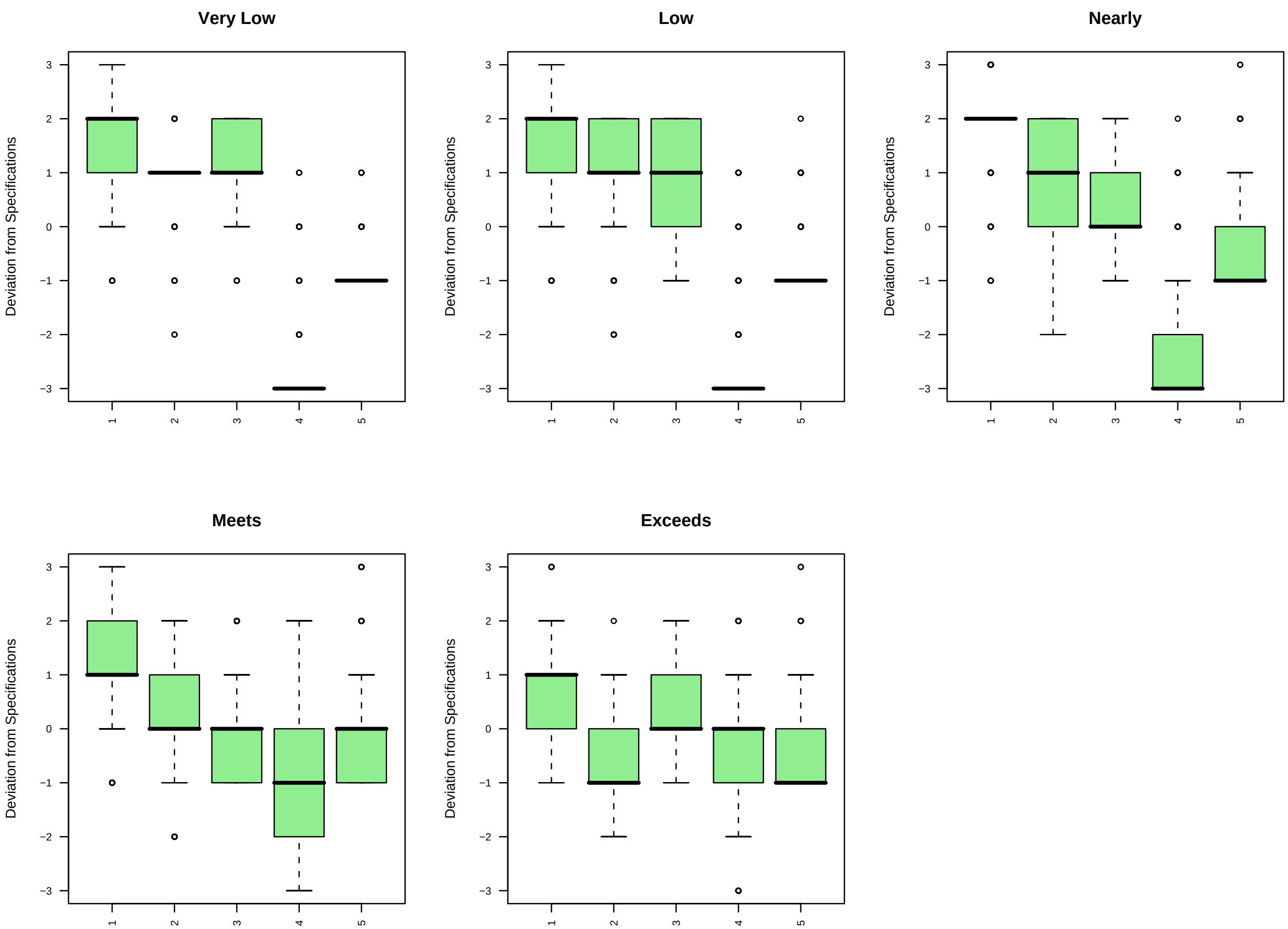


Figure 8 shows that for grade 3, most of the tests deviated from the test specifications by one or two items in each content strand at each level of proficiency. This same trend was evident in the grade 8 Reading as well as the grade 5 Math test forms. However, a number of tests did deviate from the test specifications by an amount that might be considered more than practically allowable. For example, 25% of the grade 3 test forms for students scoring in the *Low* performance category deviated from the SRC1 (vocabulary) test specifications by three or more items.

Although the proportions are very small, some tests deviate from the test specifications by eight test items, an amount that may result in significant variations in content representativeness for some students. One promising result, however, is that the patterns seem independent of performance level—that is, the patterns of deviations from test specifications for students in grades 3 and 8 Reading and grade 5 Math are similar across all five performance categories.

However, this trend did not appear to be true for grade 10 Math. Figure 11 shows that for the *Very Low*, *Low,* and *Nearly* categories, most of the test forms deviate from the test specifications in content strand 4 (Algebraic Relationships). While more than 80% of the test forms for students scoring in the *Very Low* category deviated from the test specifications by three items, the test forms for students scoring in the *Meets* and *Exceeds* categories were much more closely aligned to the specifications.

8.8 Comparability of Side-by-Side (Russian and Spanish) and TESA Tests

Source: Choi, S.W., and McCall, M. (2002). In Large Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation. Tindal, G., and Haladyna, T., Eds. Linking Bilingual Mathematics Assessments: A Monololingual IRT Approach. Lawrence Earlbaum Associates, Mahwah, NJ.

Source: Oregon Department of Education. (2007). Comparability of English and Spanish/English Mathematics Tests.

In 1998, Oregon completed a study investigating the comparability of the Side-by-Side Spanish/English Math tests. Presenting the problem in both languages gives the student the best chance to provide answers based on math proficiency without language interference. For example, a native Spanish speaker who has received math instruction in English may not know the Spanish word for diagonal or quotient.

Side-by-side tests are considered standard versions of the statewide math tests. Scores earned on the side-by-side test are included in group statistics and are considered evidence of meeting (or failing to meet) state performance standards. In the following sections, items and tests in English are compared to Spanish/English side-by-side tests and items, not to Spanish only tests and items; the terms Spanish test and Spanish item both refer to the side-by-side Spanish-English format.

The study consisted of three parts: (1) a DIF study, (2) a calibration analysis to identify equating items, and (3) an analysis of test difficulty and results providing evidence of comparability in the assessment of student's mathematics ability using the Spanish and English versions of the test. The first part of the study evaluated the entire data set for the presence of DIF. The object of this part of the study was to see how much group difference remained when size and ability density differences were taken into account. The second part compared the results of using different types of anchor items in equating. Much of the study involved the search for a defensible set of linking items. The third part of the study investigated the effect of the equating designs on student scores and on judgments about student proficiency.

In 1998, English Language Learners (ELLS) participated in the Spanish side-by-side version of the Oregon statewide Mathematics multiple-choice assessments. The tests were administered at four grade levels (third, fifth, eighth, and 10th). However, only third and fifth grades had a large enough sample of Spanish participants; hence grade 3 and grade 5 tests are the focus of the present study. Altogether, 314 third graders and 308 fifth graders took the translated form of the Mathematics multiple-choice test. The English version was taken by 8,895 third graders and 8,762 fifth graders.

Because the Spanish sample was smaller than the English sample and the patterns of raw scores were different (on average, Spanish speakers scored lower than English speakers), the first analyses were replicated under two conditions. Both used the entire Spanish sample and 10 random samples of English examinees equal in size. In the first condition (ELL population), the 10 English samples were drawn from the overall sample of students who took the English-only test. In the second condition (matched population), the proficiency distribution of the groups taking the English and Spanish side-by-side versions was matched (the Spanish side-by-side examinees were ranked according to raw score and divided into deciles, and 10 random samples from the English group were drawn to reflect the density of the Spanish side-by-side population within the score groups formed by Spanish decile groupings). The matched random samples from the English population had distributions that were similar to the Spanish sample.

The first part of the study evaluated differential item functioning (DIF) in the data set. An item exhibits DIF if the probability of a correct response is affected by group membership and the item is biased. Comparing the response functions determines whether items function differently in the two versions of the test for two or more groups matched on ability.

The IRT likelihood ratio (LR) procedure compared the likelihood that DIF does not exist in the data to the likelihood that it does (Thissen, Steinberg, & Wainer, 1988, 1993) and was used to detect DIF and evaluate the equivalence of the items across languages. Forty items in third grade and 60 items in fifth grade were calibrated according to the one-parameter logistic model (Rasch, 1960), using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), and likelihood ratio chi-square tests were conducted to provide a statistical test of translation DIF (Sireci, Foster, Olsen, & Robin, 1997):

$$G^2(d.f.) = 2\log \frac{Likelihood[Augmented]}{Likelihood[Compact]},$$

where *d.f.* is the difference between the number of item parameters in the augmented (general) number and the compact (constrained) model for two groups.

In BILOG-MG, DIF analysis is similar to nonequivalent groups equating: The group proficiency levels are assumed to be different, and the relative differences in item difficulties between groups are examined under the constraint that the mean item difficulties of the groups are equal. First, all responses were calibrated according to the one-parameter Rasch model in a single group as if they came from the same population. This model (the NO DIF condition) treats the data as if no item exhibits group differences. Next, the two groups were analyzed separately using the augmented (ALL DIF) model, which assumes that all items may exhibit DIF, and the difference between the maximum marginal likelihoods of the item parameters of the two models were compared.

Under the null hypothesis, the difference between the log likelihoods of the augmented and compact models is distributed in large samples as chi-square with [(n - l)(m - l)] degrees of freedom, where n is the number of items and m is the number of groups. A significant chi-square indicates that the model containing group item effects (ALL DIF) fits the data significantly better than the model that contains no such effects (NO DIF), providing evidence that differential item effects are present.

In addition to the LR test, the root mean squared error (RMSE) of item-difficulty estimates was computed as another measure of DIF in the data set. Item-difficulty estimates calibrated on the entire English population served as known parameters, and the item-difficulty estimates from both samples (the Spanish and 10 English) under both conditions (full and matched population) were equated to the English metric, using the test characteristic curve equating procedure (Stocking & Lord, 1983) implemented in the EQUATE program (Baker, 1995):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(b_i - \hat{b}_i)^2}{k}},$$

where b_j is the known item difficulty for item *i* estimated based on the entire English population, and b_i is the equated difficulty estimate from the Spanish or English samples.

The second part of the study explored the data for the most appropriate linking items. When English and Spanish side-by-side items are calibrated separately, item-difficulty estimates are not necessarily on a common metric. To the extent that a strong linear relationship exists between difficulty estimates for Spanish side-by-side and English-only items, a common item equating procedure can adequately remove the differences in the centering and dispersion of item-difficulty estimates. An assumption underlying the use of common item equating in cross-lingual assessments is that the translation of items into another language does not make the items uniformly more difficult or easier than the source language items. However, if difficulty estimates are offset by factors other than group ability differences, arbitrarily aligning them can, in fact, induce a bias into the comparison of scores across languages. This study sought to minimize error by finding the least biased set of items.

Multiple-group IRT analyses can be used to identify individual items showing DIF. The BILOG-MG computer program compares item-difficulty estimates for two groups under the assumption that the mean difficulty of the translated items is equal to that of original items. This study used Stocking and Lord's (1983) test characteristic curve equating method to place difficulty estimates on a common metric. Test characteristic curve equating provides linking constants that are less susceptible to outlying items with a significant level of DIF. Assuming asymptotic normality of difficulty item estimate, *bj* for item *i*, the null hypothesis, H,: *bj* = *bj* can be evaluated in a simple asymptotic significance test (Lord, 1980) as shown below by referring *dj* to the standard normal distribution table:

$$d_i = \frac{\hat{b}_{ei} - \hat{b}_{si}}{\sqrt{Var\,\hat{b}_{ei} + Var\,\hat{b}_{si}}},$$

where *ei* is the English difficulty estimate for item, *i* is the Spanish side-by-side difficulty estimate, and Var b_{ei} and Var b_{si} are the squared standard errors associated with item-difficulty estimates for English and Spanish, respectively. Using a criterion of about **2** in absolute value to judge an item to exhibit DIF, all items showing an absolute value of d_i greater than 1.96 were identified as DIF items. Additionally, the English and Spanish-bilingual data were calibrated separately using the BILOG-MG computer program and were equated to the existing bank (RIT) scale using the common-item equating procedure. The RIT (Rasch Unit) scale has a mean of approximately 200 and a standard deviation of 10 at third grade and is vertically linked to higher grades with common anchor items.

Difficulty estimates for the Spanish side-by-side items were then equated to the English metric, using the test characteristic curve equating procedure (Stocking & Lord, 1983), presuming a common-item nonequivalent group equating design. Three different sets of common items were identified and used in subsequent equating procedures: (1) all 40 items in third grade and all 60 items in fifth grade were used as anchor items to equate Spanish item-difficulty estimates to the English metric; (2) only items with low DIF ($|d_i| < 1.96$) were anchor items; and (3) a set of anchor items was identified by expert bilingual content specialists most amenable to clear translation and least dependent on language and culture across the two language groups. Items excluded from the linking blocks were treated as unique items.

The third and final part of the study compared the impact on scores of various scaling and equating methods. To determine whether different equating methods yielded very different student scores or affected the percentage of students meeting state proficiency standards, following the series of common item linking studies, tests were rescored under the three equating conditions. The resulting item-difficulty estimates were used to estimate three separate scaled scores for students who took the Spanish side-by-side version. The similarity in impact on meeting the standard was compared between the separately scaled scores.

Results for the DIF analysis indicated significant DIF under both full- and matched-population conditions but found that LR chi-square statistics were substantially smaller under the matched-population condition. The matched condition reduces differences in item-difficulty estimates obtained based on the Spanish sample and 10 random samples of English students. Table 17 summarizes the results of the translation DIF analyses. Under both conditions, the LR chi-square statistics (G2) were significant ($p < .001$) across all 10 replications, indicating that the completely unconstrained model (ALL DIF) fits the data significantly better than the completely constrained model (NO DIF). Thus, the null hypothesis that there is no significant or reliable difference between the item parameters for the English and Spanish groups was rejected, indicating that one or more items function differently due to translation error or group membership or both.

Table 17.
Summary of Translation DIF Analysis

Summary of Translation DIF—Likelihood Ratio (LR) Analyses

		Full Population			*Matched Population*		
Grade	*Sample*	G^2	*df*	*p*	G^2	*df*	*p*
3	1	298.45	40	<.001	137.10	40	<.001
	2	256.47	40	<.001	140.52	40	<.001
	3	321.88	40	<.001	101.11	40	<.001
	4	281.85	40	<.001	109.63	40	<.001
	5	268.67	40	<.001	161.62	40	<.001
	6	259.20	40	<.001	120.13	40	<.001
	7	329.08	40	<.001	110.22	40	<.001
	8	295.39	40	<.001	146.08	40	<.001
	9	280.43	40	<.001	128.43	40	<.001
	10	315.92	40	<.001	146.49	40	<.001
	Mean	264.30			118.30		
5	1	525.95	60	<.001	263.01	60	<.001
	2	538.15	60	<.001	231.97	60	<.001
	3	531.11	60	<.001	250.44	60	<.001
	4	519.75	60	<.001	247.28	60	<.001
	5	541.35	60	<.001	315.30	60	<.001
	6	554.02	60	<.001	246.36	60	<.001
	7	510.58	60	<.001	272.54	60	<.001
	8	487.41	60	<.001	268.99	60	<.001
	9	468.38	60	<.001	284.44	60	<.001
	10	535.81	60	<.001	248.45	60	<.001
	Mean	521.25			262.88		

When differences in population ability were controlled (matched population) in LR analyses, DIF effects became smaller but were still significant. The average LR chi-square statistic ($G^2_{(40)}$) was 264.3 for third grade under the full-population condition, whereas average ($G^2_{(40)}$) dropped to 118.3 for the matched-population condition. A similar result was found for fifth grade; the average ($G^2_{(60)}$)was 521.25 for the full-population condition compared with 262.88 for the matched-population condition.

Table 18 shows the RMSE of Spanish side-by-side item calibrations and the RMSE of the 10 English-only samples.

Table 18.
RMSE of Item-Difficulty Estimates

	Grade 3		*Grade 5*	
Language	*Full Population*	*Matched Population*	*Full Population*	*Matched Population*
Spanish	0.283		0.387	
English 1	0.112	0.204	0.128	0.269
English 2	0.122	0.213	0.115	0.277
English 3	0.121	0.227	0.143	0.242
English 4	0.132	0.227	0.154	0.245
English 5	0.099	0.218	0.148	0.267
English 6	0.099	0.225	0.135	0.251
English 7	0.138	0.195	0.138	0.274
English 8	0.132	0.230	0.148	0.231
English 9	0.119	0.203	0.149	0.246
English 10	0.124	0.227	0.141	0.269
English Mean	0.120	0.217	0.140	0.257
English *SD*	0.013	0.013	0.011	0.016

The RMSE for Spanish calibrations for third grade was 0.283 in logit unit, compared with the average RMSE of 0.120 for English samples under full-population condition. However, when English samples were drawn from the population matched to the Spanish distribution, the average RMSE increased to 0.217, which is more closely comparable to that of the Spanish group. A similar trend was found in fifth grade; the average RMSE for the matched-English samples (0.257) was closer to the RMSE for the Spanish group (0.387) and larger than the average RMSE for the full-English samples (0.140). It is worth noting that the difference in RMSEs for the Spanish group and matched-English samples was larger for fifth grade (0.130) than for third grade (0.066). The matched population condition is a more accurate representation of the effect of translation error. It shows how much error the Spanish sample exhibits beyond that of an English sample of similar range and density.

Figure 12 graphically displays the RMSEs presented above in Table 18. The RMSEs for the item-difficulty estimates for both full- and matched-population conditions across 10 replications are plotted, as well as the RMSE for the Spanish item-difficulty estimates.

Figure 12.
RMSE for Spanish and 10 English Replications for Both Full- and Matched-Population Conditions

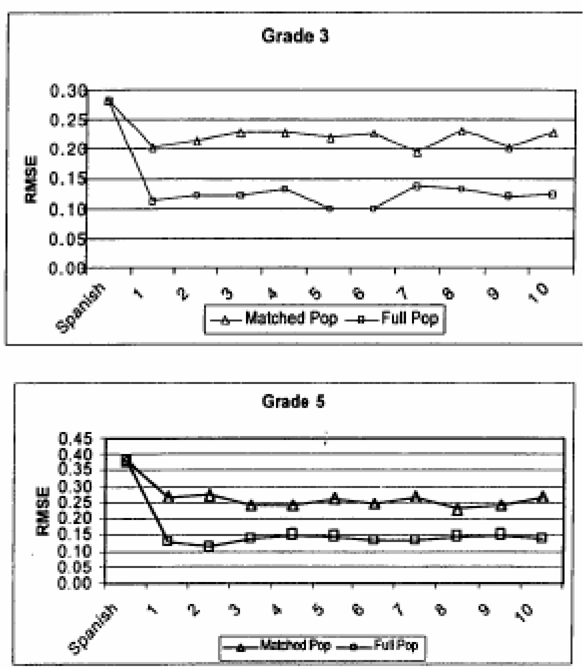


Table 19 provides the results of the common item linking analyses and shows the number of items included in each common anchor item block and descriptive statistics, including correlation coefficients between the English and Spanish item-difficulty estimates after the initial linking study. The correlation coefficients were acceptably high and ranged from 0.86 to 0.98. The 17 items selected by expert judges in grade 3 had a correlation of 0.95, which is slightly higher than that for all 40 items. However, the 16 judge-selected items in grade 5 had a correlation of 0.86. This somewhat lower correlation was caused by an outlying item, which showed a significant amount of DIF. Dropping that item from the common anchor block raised the correlation to 0.94.

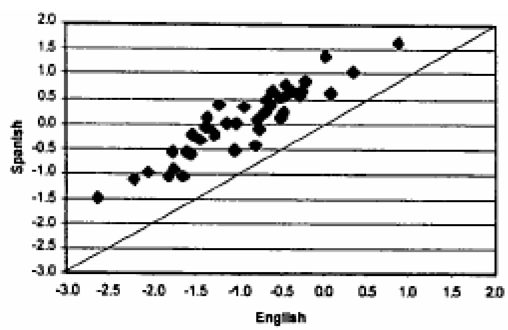
Table 19.
Descriptive Statistics of Linking Items

	Grade 3			Grade 5		
Variable	*All Items*	*NO-DIF*	*Expert*	*All Items*	*NO-DIF*	*Expert*
Number of items	40	28	17	60	39	16
Correlation	0.93	0.98	0.95	0.90	0.97	0.86
Mean	196.05	195.79	193.13	212.24	212.95	212.03
SD	7.71	8.38	7.29	8.54	8.50	7.57

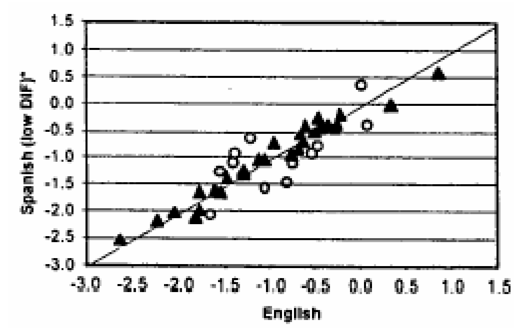
Figure 13 displays scatter plots of grade 3 English and Spanish item-difficulty estimates before equating and for all three equating conditions, with the identity line projected through the scatter plots. Figure 14 displays the same information for grade 5 items.

Figure 13.
Scatter Plots of English and Spanish Item Difficulties: Grade 3

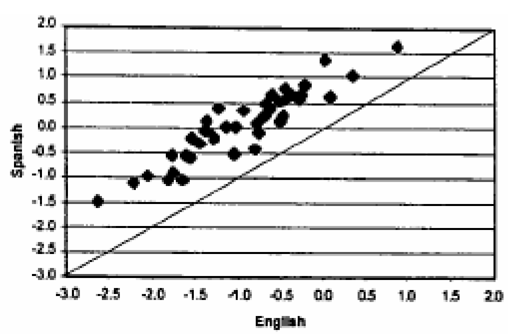
Before equating



Low DIF linking items (condition 2)



Spanish items equated to English scale (condition 1)



Expert chosen linking items (condition 3)

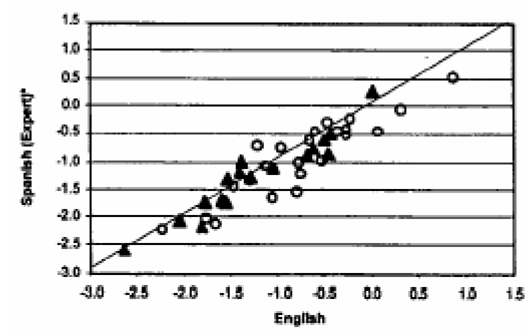
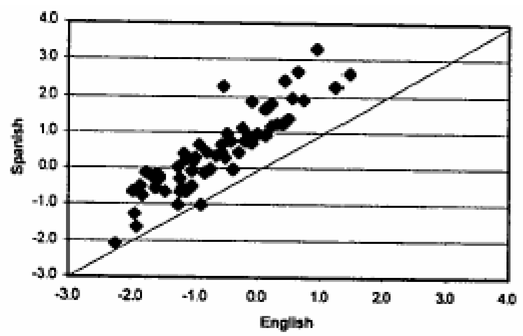
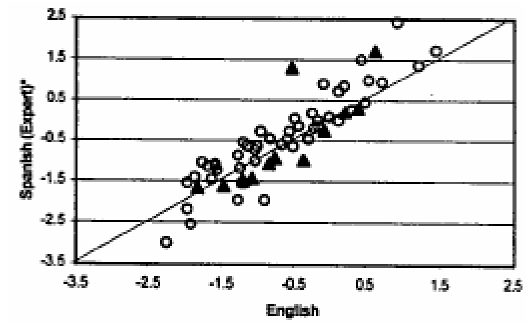


Figure 14.
Scatter Plots of English and Spanish Item Difficulties: Grade 5

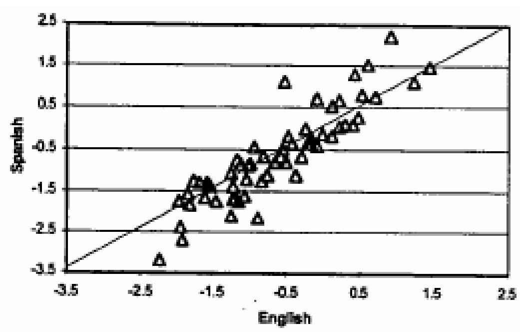
Before equating



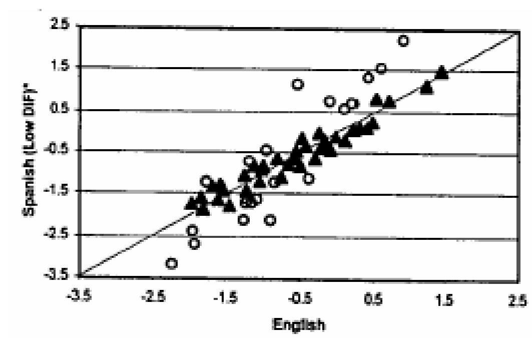
Low DIF linking items (condition 2)



Spanish items equated to English scale (condition 1)



Expert chosen linking items (condition 3)



Consistent with the high correlations, the scatter plot of item-difficulty estimates was quite tight for grade 3. No distinct outliers were present. In conditions 2 and 3, the items included in the common anchor block were denoted as black triangles.

Items identified with Lord's chi-square showed the most compact pattern, falling closely to the identity line. The scatter plot of item-difficulty estimates for grade 5 was not as tight as that for grade 3, consistent with the slightly lower correlation. There appeared to be a few items in grade 5 that were significantly deviant from the order of relative difficulties between languages.

Finally, Tables 20 and 21 summarize the results of the comparison of scaled scores. Raw number-correct scores were transformed into the RIT (Rasch unit) scores. Four raw to RIT score conversions were run for Spanish raw scores, based on different sets of linked item-difficulty estimates.

Tables 20 and 21 also include the mean proficiency estimates obtained from three different sets of linked item-difficulty estimates. These mean proficiency measures ranged from 194.38 to 195.34 for grade 3. The difference between the lowest and highest mean proficiency estimates in grade 3 was approximately 1 RIT or 0.1 logit, indicating practically no variation in mean proficiency measures across various linking procedures. In grade 5, the mean proficiency measures for Spanish students ranged from 208.54 for the condition where all 60 items were used as common anchor items to 210.34 for the expert-chosen linking items.

To meet the Mathematics proficiency standard in third grade, students must have a RIT score greater than or equal to 202; fifth graders need a score of 215 or above to meet the standard. Condition 1 resulted in the same percents meeting criteria when linked scores were compared to scores based on English-only calibration in both third and fifth grades. Seven third-grade students met the standard when the English calibrations were used but failed to meet the standard under Conditions 2 and 3. In fifth grade, the equating based on expert-chosen anchor items yielded somewhat different results compared with the English calibration and other equating procedures. About 16.56% fifth graders met the standard under English calibration, equatings based on all items and on a DIF statistic (d_i), whereas 21.75% would meet the standard on the basis of the equating with expert-chosen anchor items. Sixteen fifth-grade students who did not meet the standard based on the English calibration did meet it when the items identified by experts were used in equating. However, two grade 5 items chosen by experts were outliers. This makes the set of items exhibit less dependability than the grade 3 set.

Table 20.
Comparison of Scaled Scores, Grade 3

		Spanish Group	*Equated Spanish Group*		
Variable	*English Group*	*English Calibration*	*All Items*	*NO DIF*	*Expert*
N	8,895	314	314	314	314
Mean raw score	27.02	19.34	19.34	19.34	19.34
Mean scaled score	206.61	195.34	195.35	195.04	194.38
SD of scaled scores	12.32	10.21	10.16	10.16	10.16
% Meeting standard	65.7	24.52	24.52	22.29	22.29
% Not meeting standard	34.3	75.48	75.48	77.71	77.71

Table 21.
Comparison of Scaled Scores, Grade 5

		Spanish Group	*Equated Spanish Group*		
Variable	*English Group*	*English Calibration*	*All Items*	*NO DIF*	*Expert*
N	8,762	308	308	308	308
Mean raw score	36.74	25.80	25.80	25.80	25.80
Mean scaled score	218.43	208.82	208.54	208.60	210.34
SD of scaled scores	10.44	7.25	7.75	7.75	7.75
% Meeting standard	60.8	16.56	16.56	16.56	21.75
% Not meeting standard	39.2	83.44	83.44	83.44	78.25

The results of this study indicate that the English and Spanish-English side-by-side versions of the third- and fifth-grade Mathematics multiple-choice tests are similar. When the performance of students was reviewed, the order of difficulty for both the English and Spanish-English was predominantly the same. Although some items functioned differently between the two groups of examinees, when tests were equated under a number of common item linking designs, there was little difference in the impact on overall score or judgment of mastery for the Spanish group.

Although a series of LR chi-square tests comparing ALL DIF and NO DIF models indicated that one or more translated items performed differently, DIF was substantially reduced when item difficulties were estimated from samples with matched ability levels. However, the extent of differential item functioning that remained when populations were matched was still statistically significant.

To determine whether equating the Spanish and English tests has an impact on the use and interpretation of scores in the school setting, IRT common item linking procedures were used under the assumption that translation or adaptation does not affect the overall level of item difficulties. Translated tests were equated to the common metric and criterion of the source language tests, using three types of anchor items. No substantive differences were found between the English calibration and the Spanish calibrations linked through the various sets of common anchor items.

8.7.1 Translation Accuracy

Information on translation accuracy is included in a separate report included as part of this package.

9. Fairness and Accessibility

Fairness concerns occur throughout testing. Standardization itself is intended to ensure that no examinees are given advantages or impediments through administration practices. Nevertheless fairness issues arise simply because uniform conditions trigger different levels of comfort in examinees. Although absolute fairness cannot be guaranteed, sources of bias should be investigated and controlled to the extent practicable. The following section includes several components of fairness and accessibility:

- 9.1 Administration
- 9.2 Test Content
- 9.3 Statistical Bias
- 9.4 Score Reporting and Use

9.1 Fairness in Administration

In general, consistency in administration enhances fairness. There may, however, be variations if test procedures create more valid conditions for certain groups or individuals. These include changes in setting or process for disabled students as well as the use of small-group settings or native language instructions if these serve to promote more accurate performance on the test.

In accordance with the standards and NCLB critical elements, test manuals stipulate that all students should be treated equitably during testing. Accommodations allow more students to participate in assessment as required by federal law. In 2001, the Oregon Department of Education entered into a legal settlement agreement with Advocates for Special Kids (ASK) to change the way ODE identifies accommodations for the Oregon State Assessment.

The effect of this settlement was to allow accommodations to assessments unless research proved that they affected test validity. This was a change from previous policy, which required changes in testing to undergo validity checks before being accepted as accommodations. In response to the agreement, ODE set up a panel to review modifications and accommodations as requests arose from the field.

Oregon devotes a lot of effort to developing and providing a comprehensive, well-researched set of accommodations (changes in testing that do not affect score validity) and modifications (changes that affect the meaning and use of scores). The process of developing, documenting, and disseminating information concerning appropriate accommodations and promoting their use is described in detail in the Annual Technical Report on the Department's website at http://www.ode.state.or.us/search/page/?id=787 and in the accommodations and modifications tables available at http://www.ode.state.or.us/search/page/?id=487.

Also available from the same sources are current accommodations, modifications, and eligibility requirements. Administration manuals are updated annually to reflect changes and are maintained on the ODE website at http://www.ode.state.or.us/search/page/?id=486.

The Accommodations Panel maintains and executes a research agenda; records of meetings can be obtained from the Oregon Department of Education.

9.2 Fairness in Test Content

Test items and prompts are checked to ensure that they are culturally and linguistically appropriate and that they are not biased or offensive to examinee ethnic, religious, or cultural groups. Documentation of this process is part of the overall validity argument.

During training for item writing, participants are trained in sensitivity to fairness considerations. Like most large-scale testing programs, Oregon conducts fairness review of items and prompts prior to field testing. A description of this process, the composition of the bias review committees, and the general criteria for approving items are publicly available on the Department's Web site and in technical reports.

9.3 Statistical Bias Issues

Statistical bias is present when the measurement model does not work the same way for different groups. Tests can be scrutinized for bias by conducting differential item function (DIF) studies and dimensionality analysis. DIF studies look at how item statistics differ among groups. Test items are *unbiased* when the probability for success on the items is the same for equally able examinees of the same population regardless of their subgroup membership. In IRT contexts, bias can be measured by departures from model fit indicated by different item response functions for students belonging to one group than for those belonging to another.

There are often no common characteristics that explain DIF, so each item exhibiting DIF is looked at separately. At the point when item DIF is detected, items undergo a number of reviews, but only infrequently is an apparent bias factor discovered. DIF may arise because of differences in opportunity to learn or in how groups of students respond to instruction. In these cases, DIF may be a tool for informing instruction.

When the discrepancy between calibrations for groups reaches a threshold, content specialists are notified. Because items exhibit DIF for a variety of reasons, content experts usually make the final decision about DIF items, deciding whether to delete or rewrite and retest these items. In any case, ODE conducts annual DIF analyses on sets of items within all tests to examine item DIF for each of the required subgroups plus gender.

REFERENCES

Baker, 1995

Choi, S. W., and Tinkler, T. (2002). Evaluating comparability of paper and pencil and computer based assessment in a K-12 setting. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Choi, S.W., and McCall, M. (2002). In Large Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation. Tindal, G., and Haladyna, T., Eds. Linking Bilingual Mathematics Assessments: A Monololingual IRT Approach. Lawrence Earlbaum Associates, Mahwah, NJ.

Cohen, J. and the American Institutes for Research (2002). AM Statistical Software (Beta Version 0.06.00). Washington DC: American Institutes for Research. Software and manual available free of charge at http://am.air.org.

Cohen, Seburn, Antal, and Gushta. (2005).

Doran & Jiang, 2004

Doran, H., and Cohen, J. (2006). Analytic and Psychometric Services: Technical Report II. Technical Report for the Oregon Department of Education by the American Institutes for Research.

Doran, H., and Cohen, J. (2006). Oregon Technical Report I. Technical Report for the Oregon Department of Education.

Kingsbury, G.C., and Hauser, C. (2004). Computerized adaptive testing and No Child Left Behind, Northwest Evaluation Association. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolen, M. J. and Brennan, R. L. (2004). Test equating, scaling and linking: Methods and practices. Springer-Verlag

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

McCall, M.S. (n.d.). Technical review of the Oregon Department of Education assessment system. Northwest Evaluation Association.

Measuring Up: A Report on Education Standards and Assessments for Oregon. (2000). Technical Report Prepared for the Oregon Department of Education by Achieve, Inc.

Michaelides & Haertel, 2003

North Central Regional Educational Laboratory, (n.d.). *Surveys of enacted curriculum*. Retrieved from http://www.secsupport.org/ April 25, 2007.

Oregon Department of Education. (2001) *Technical Digest—Validity—Oregon Statewide Assessment Program*.

Oregon Department of Education. (2003). The first year: Student performance on the 10th grade benchmark standards and subsequent performance in the first year of college. Oregon Office of Community Colleges and Workforce Development, The Oregon University System.

Oregon State Curriculum Standards and Areas of Alignment with the PSAT/NMSQT, SAT I and SAT II. (2002). Report for the Chancellor's Office of the Oregon State University System by Educational Testing System and The College Board.

Oregon Statewide Assessment System. *(2005)*. Standard errors of measurement, reading and mathematics knowledge and skills tests.

Relationships between student's performance on Oregon's 10th grade Certificate of Initial Mastery (CIM) Assessments and Selected Oregon Business Pre-Employment Tests. (1999). The Oregon Business Council, The Oregon Department of Education, and The Oregon University System.

Roeber, E. D. (1996). Review of the Oregon Content and Performance Standards. Report for the Oregon Department of Education by the National Standards Review Team.

Samejima, F. (1972)

Sireci, Foster, Olsen, & Robin, 1997

Stocking & Lord, 1983

Thissen, Steinberg, & Wainer, 1988

Thissen, Steinberg, & Wainer, 1993

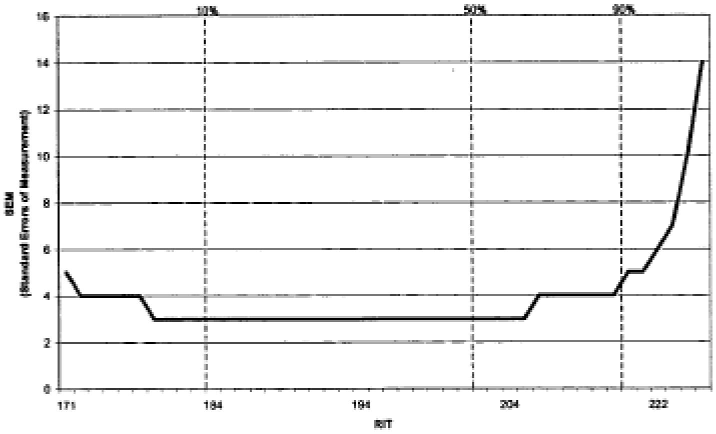
Zimowski M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (1999) BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items. Scientific Software International, Inc. Chicago, IL

APPENDIX A
STANDARD ERRORS OF MEASUREMENT, PAPER-AND-PENCIL AND TESA, OVERALL AND BY SUBGROUP, 2003–04

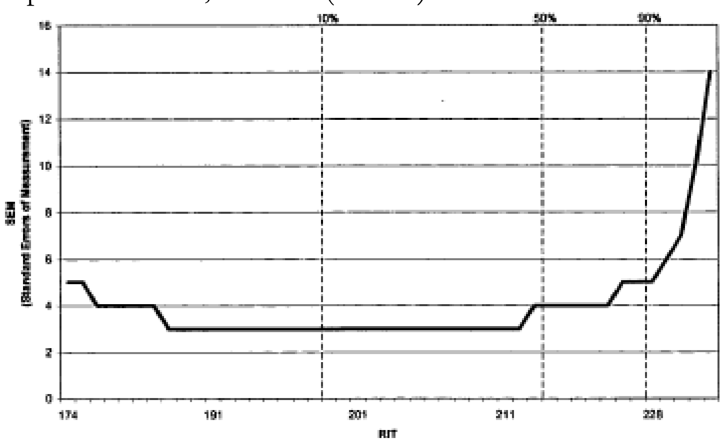
A.1 Grade 3 Reading

Test Form

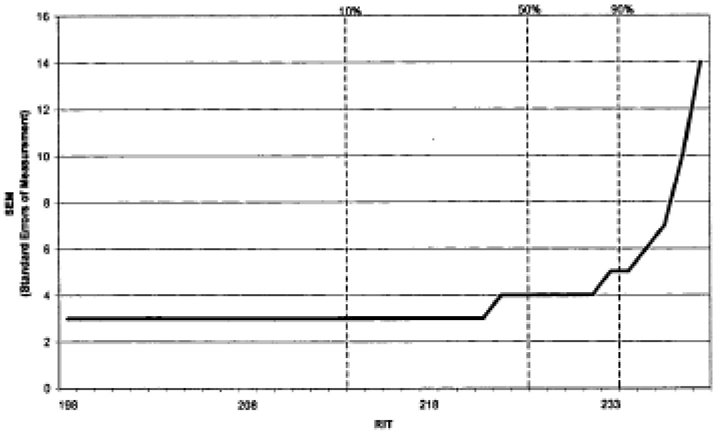
Paper and Pencil, Form A (lower)



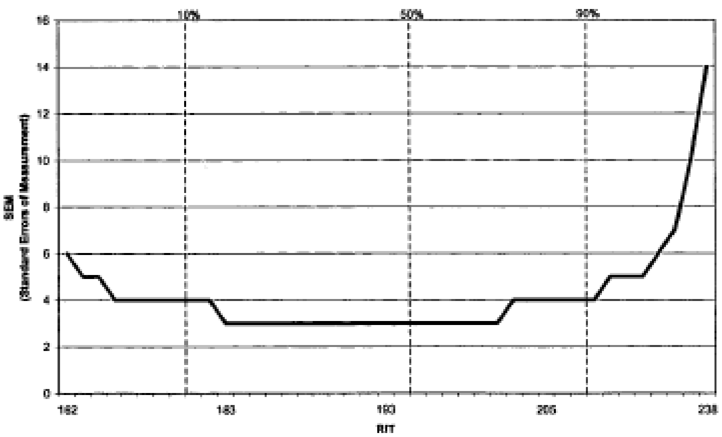
Paper and Pencil, Form B (middle)



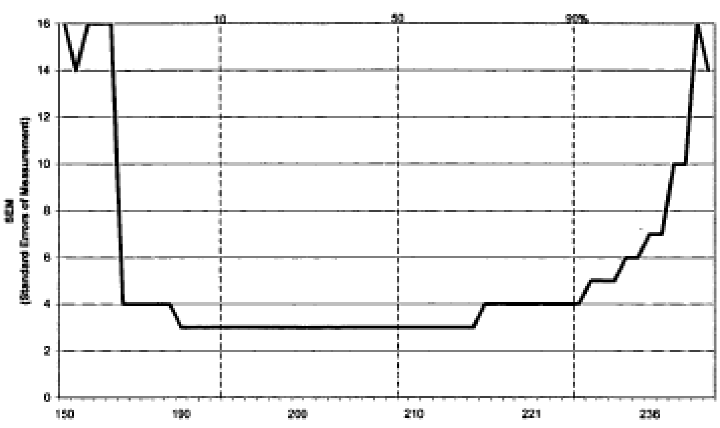
Paper and Pencil, Form C (upper)



Paper and Pencil, Plain Language (lower)

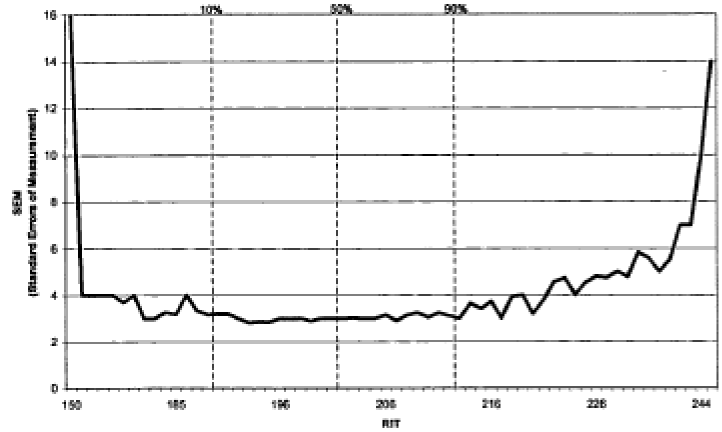


TESA (Wide Range)

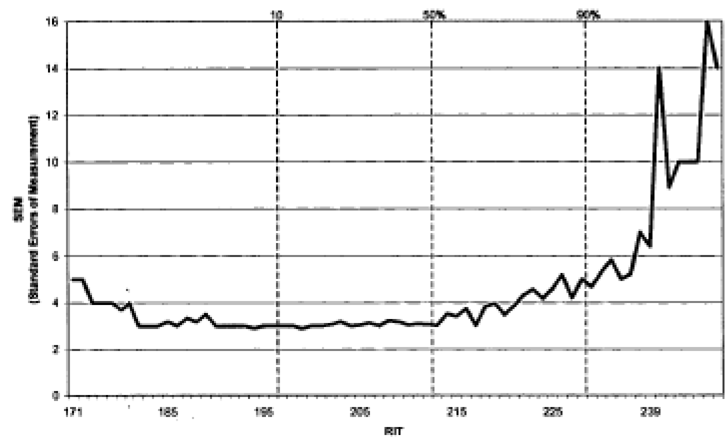


Ethnicity

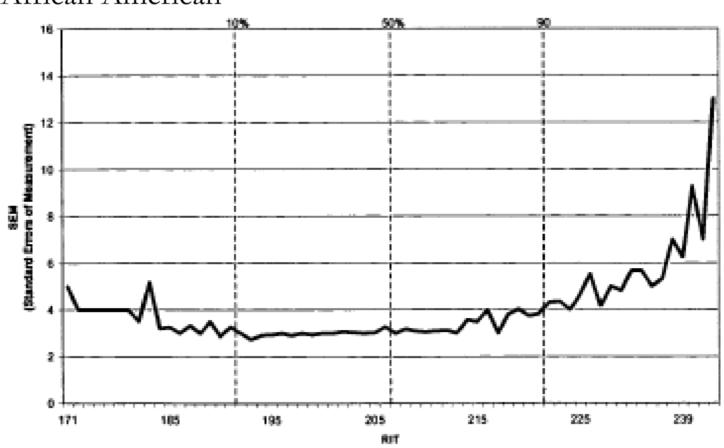
Native American Students



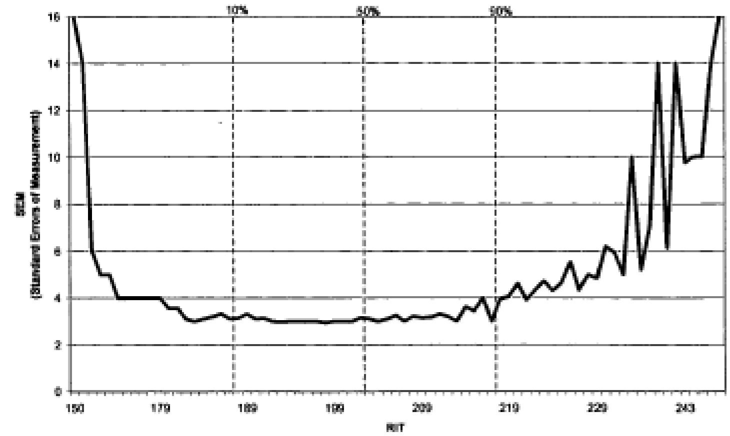
Asian



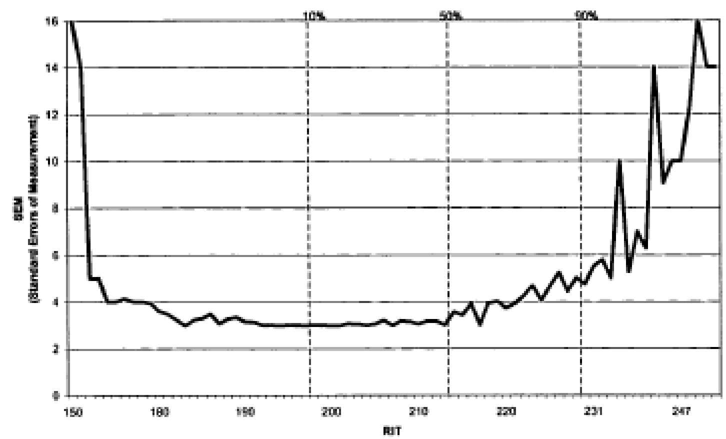
African American



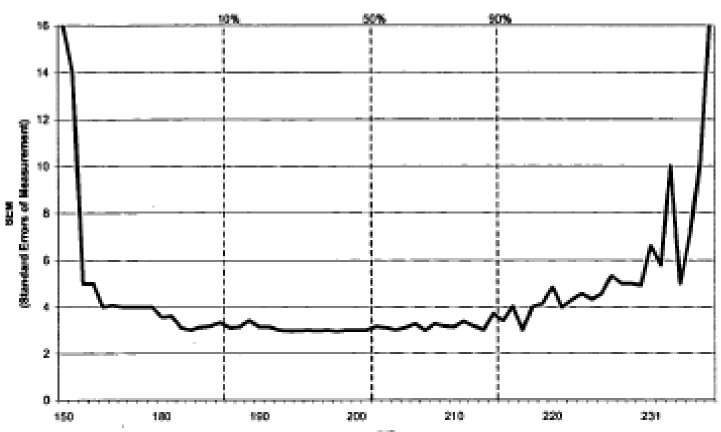
Hispanic



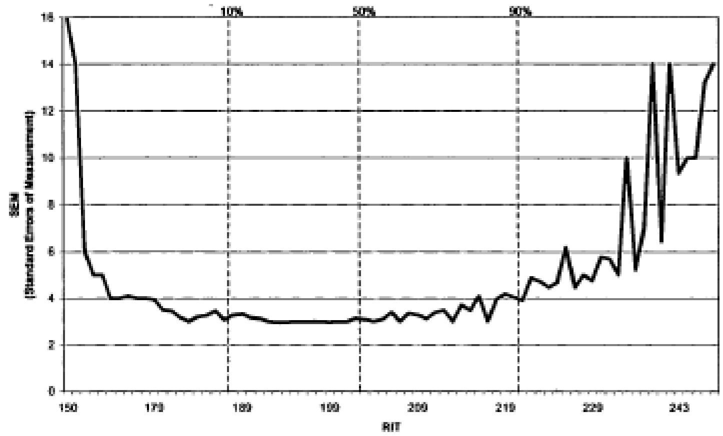
White



LEP Students



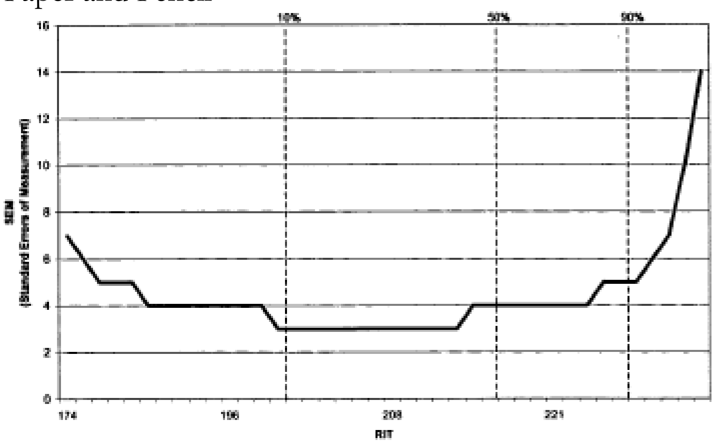
Special Education Students



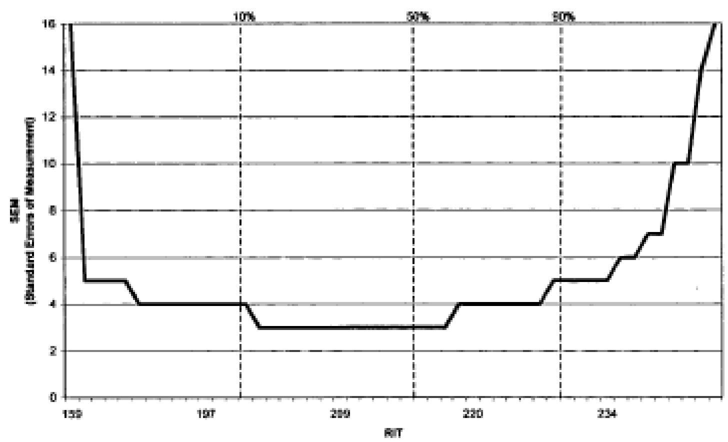
A.2 Grade 4 Reading

Test Forms

Paper and Pencil

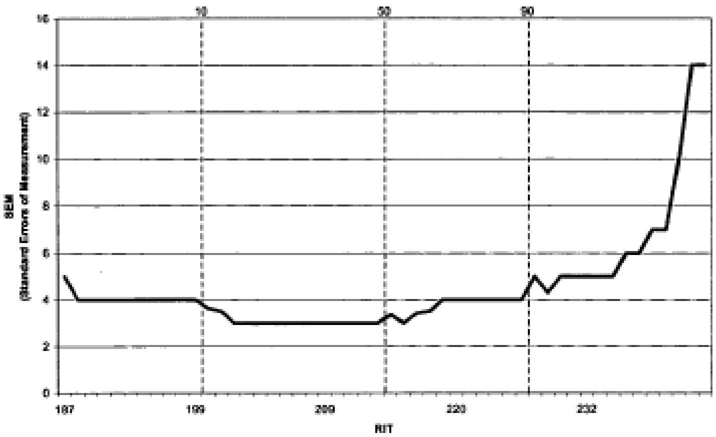


TESA

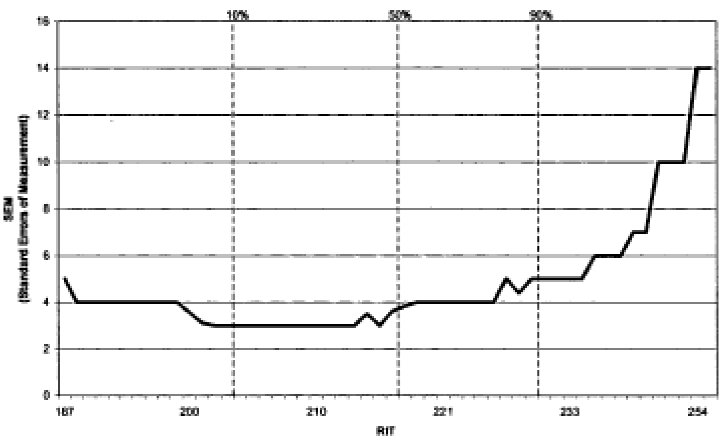


Student Ethnicity

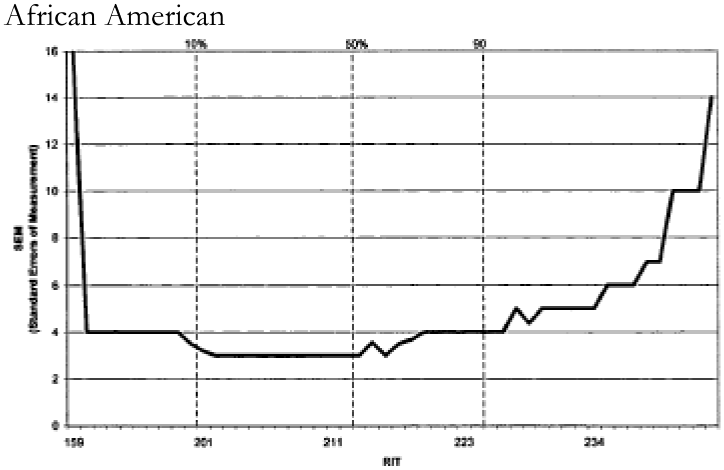
Native American



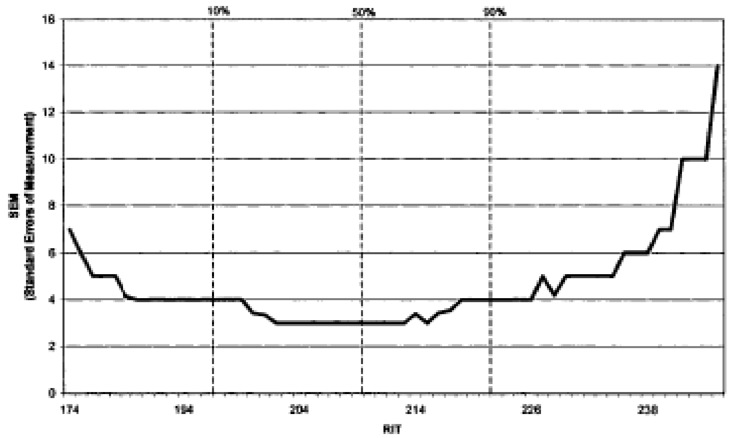
Asian



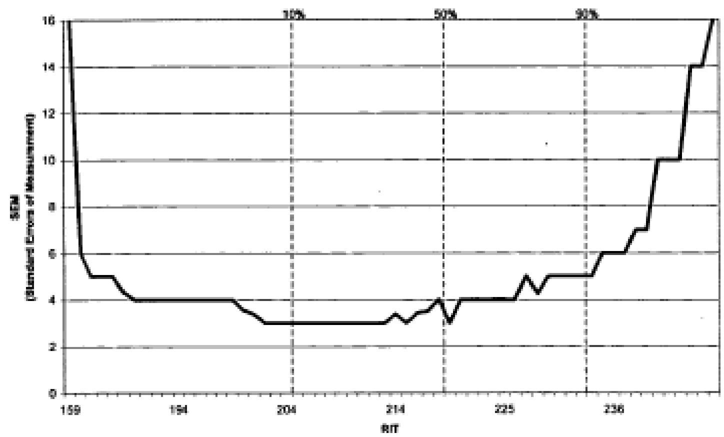
African American



Hispanic

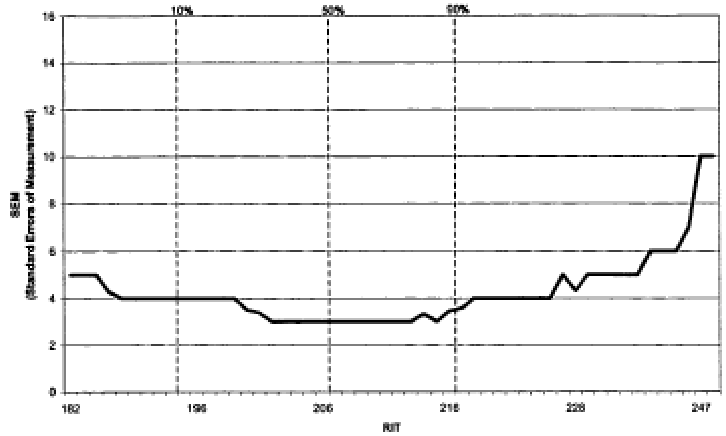


White

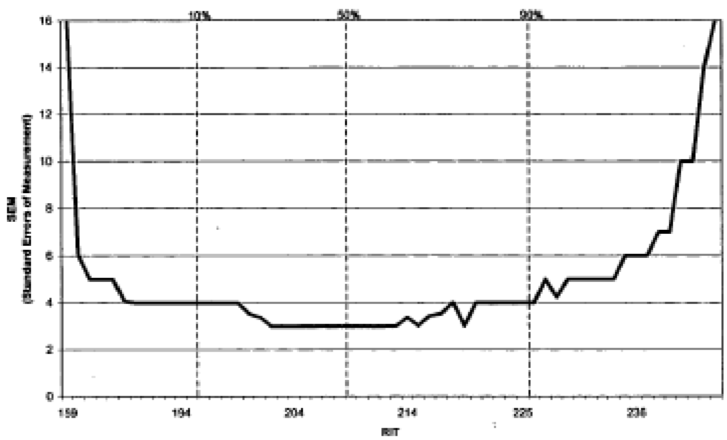


Special Populations

LEP Students



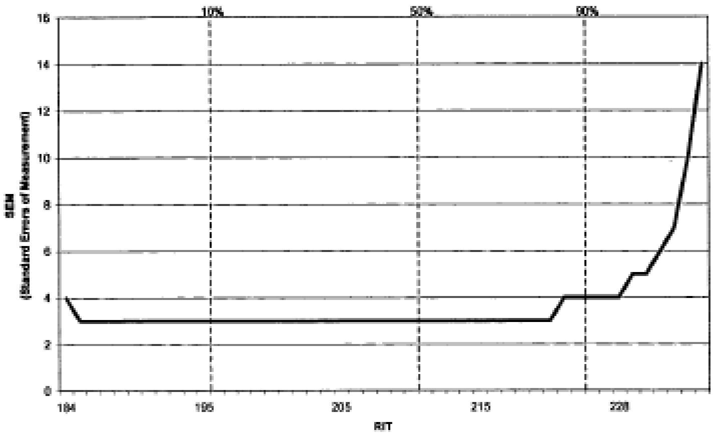
Special Education Students



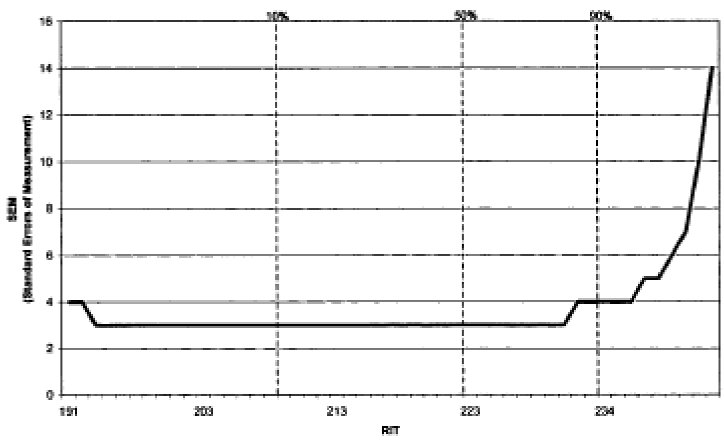
A.3 Grade 5 Reading

Test Forms

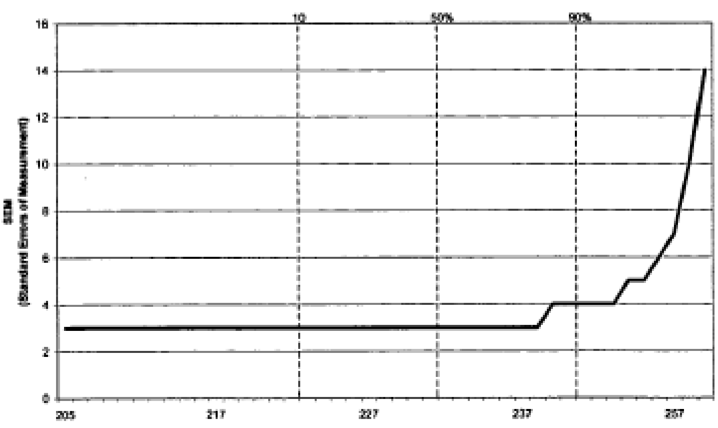
Paper and Pencil, Form A (lower)



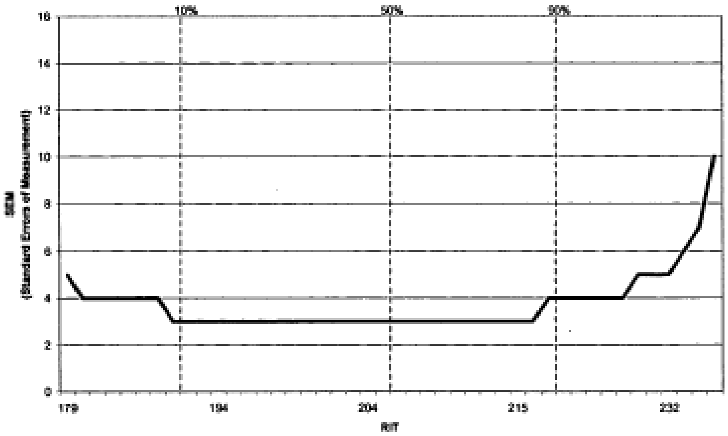
Paper and Pencil, Form B (middle)



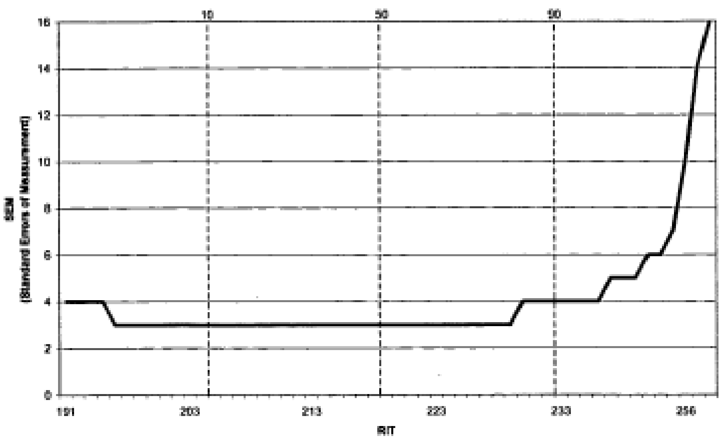
Paper and Pencil, Form C (upper)



Plain Language

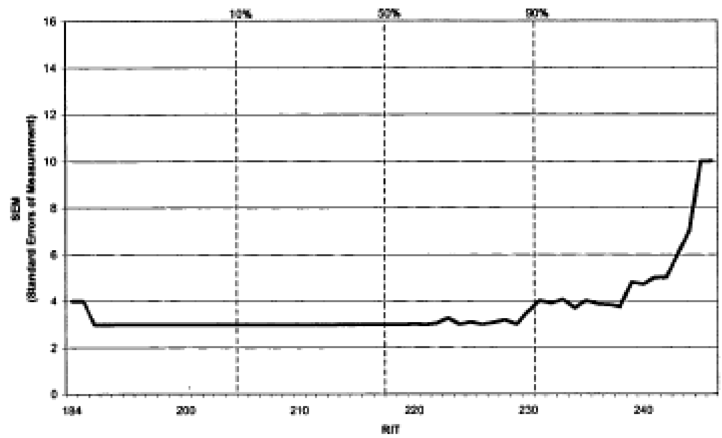


TESA

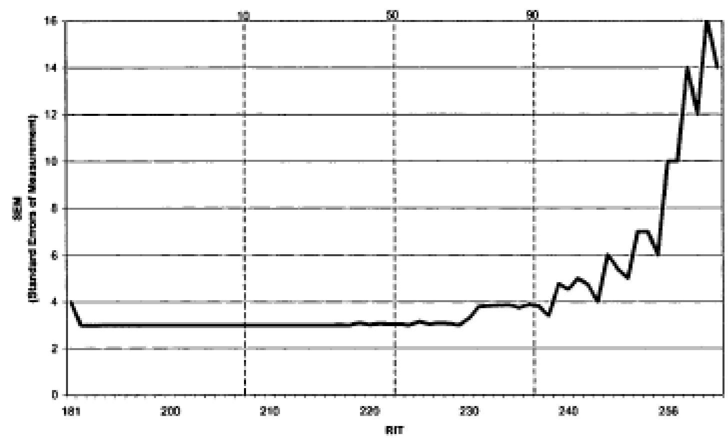


Student Ethnicity

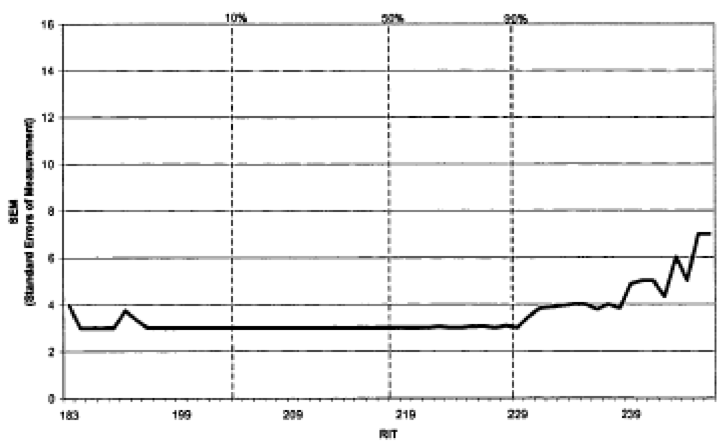
Native American



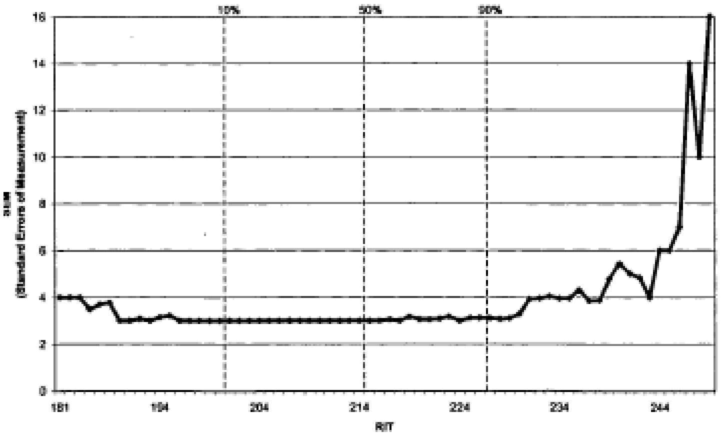
Asian



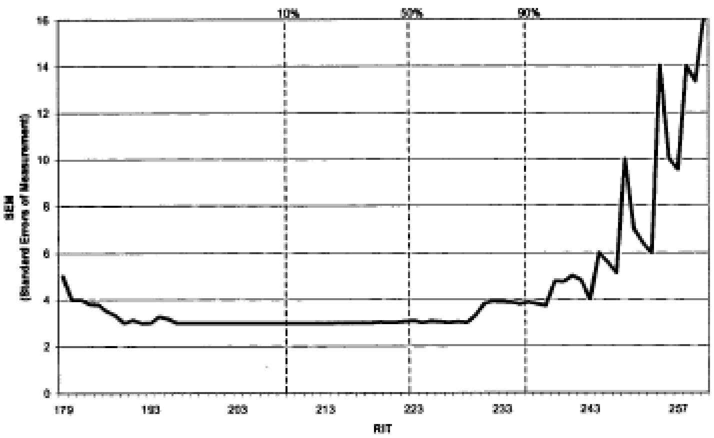
African American



Hispanic

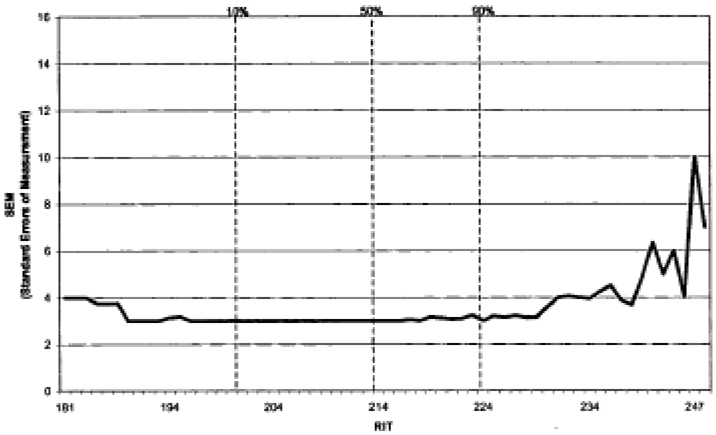


White

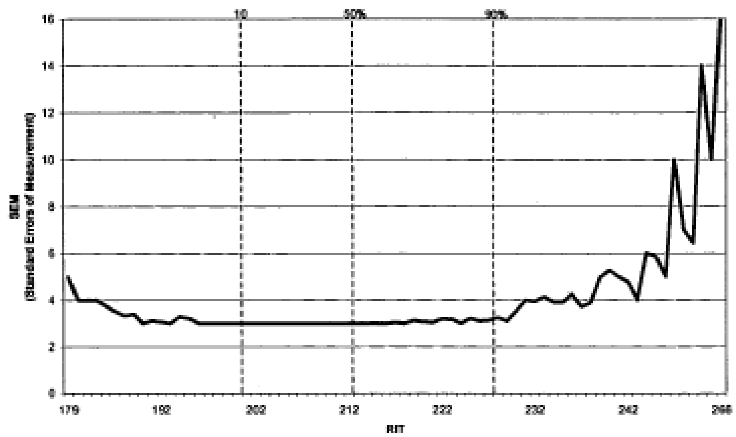


Special Populations

LEP Students



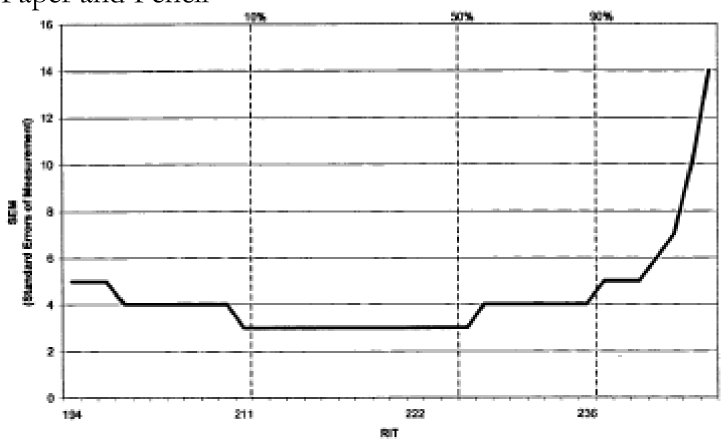
Special Education Students



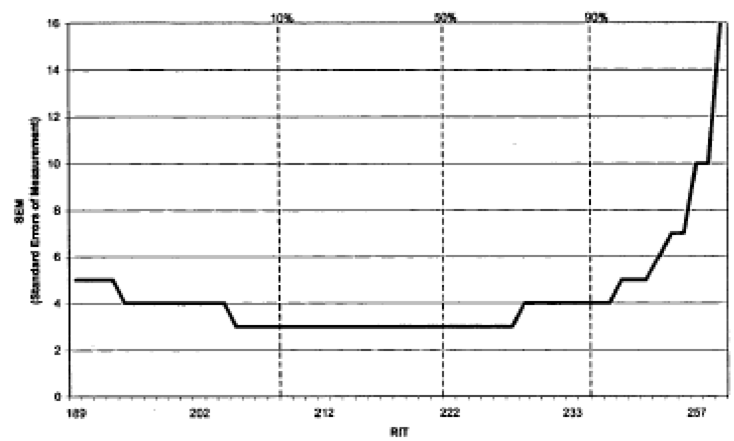
A.4 Grade 6 Reading

Test Forms

Paper and Pencil

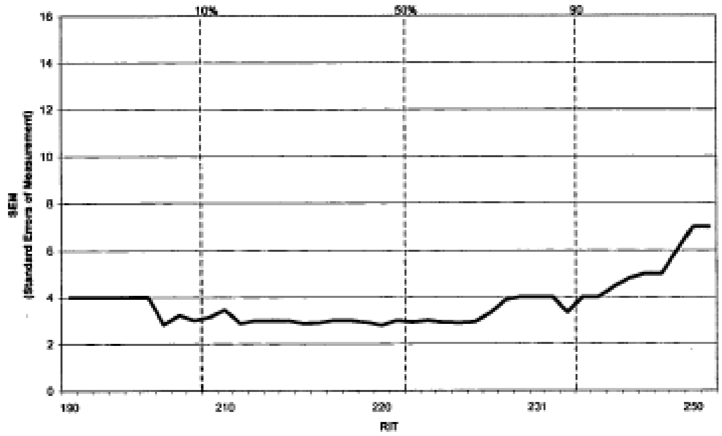


TESA

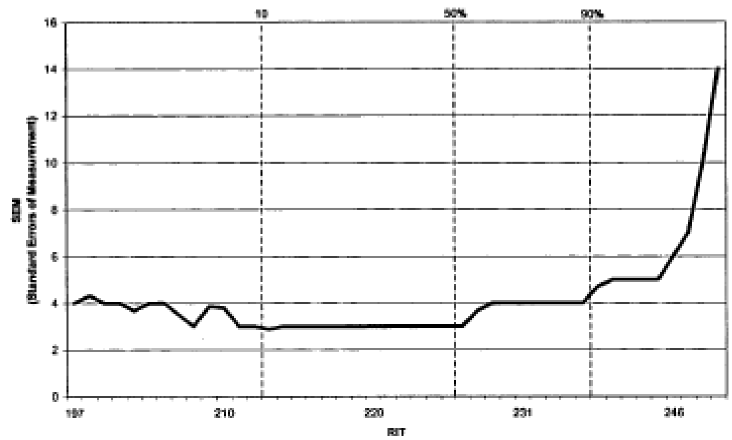


Student Ethnicity

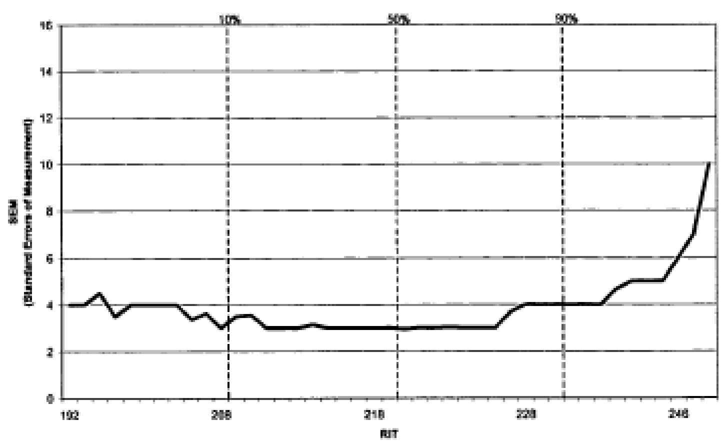
Native American



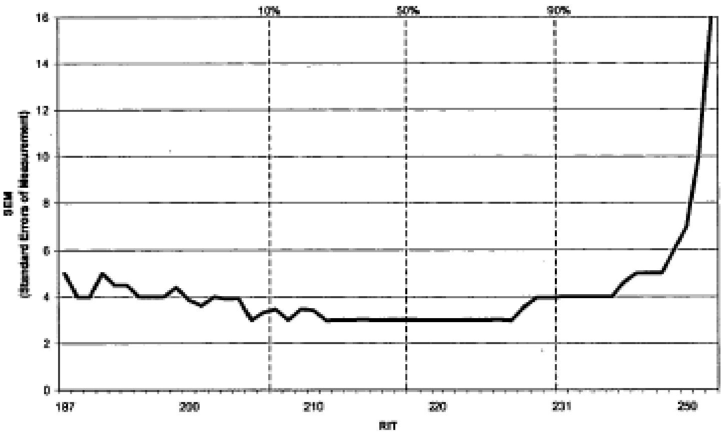
Asian



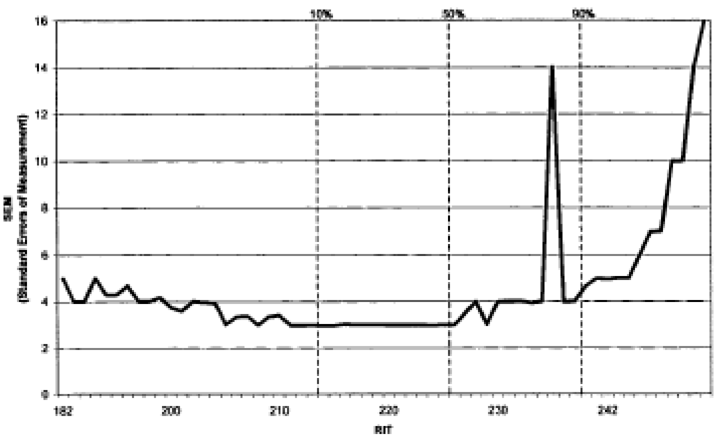
African American



Hispanic

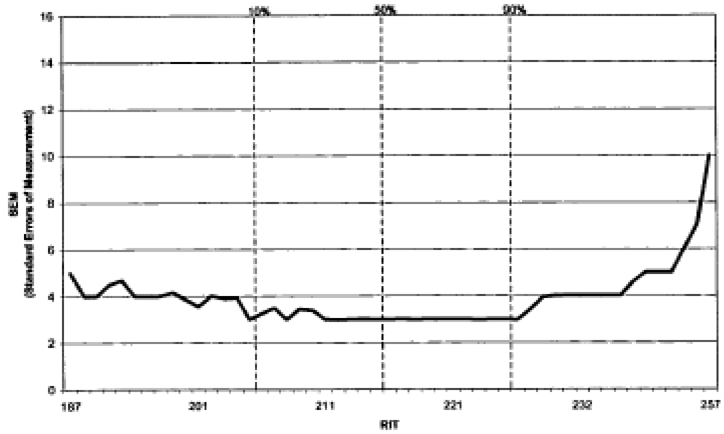


White

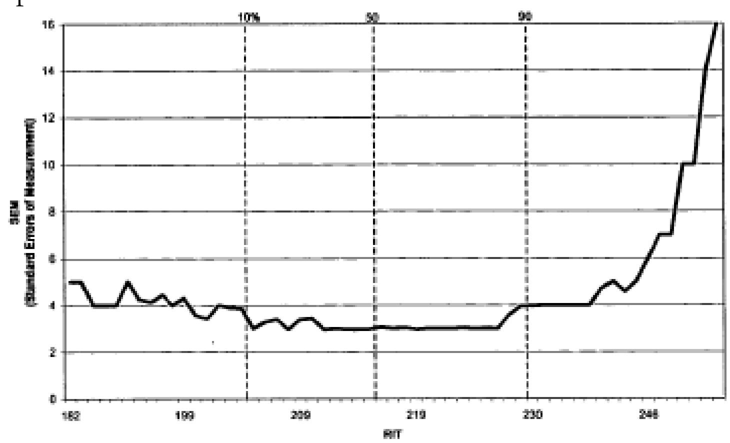


Special Populations

LEP Students



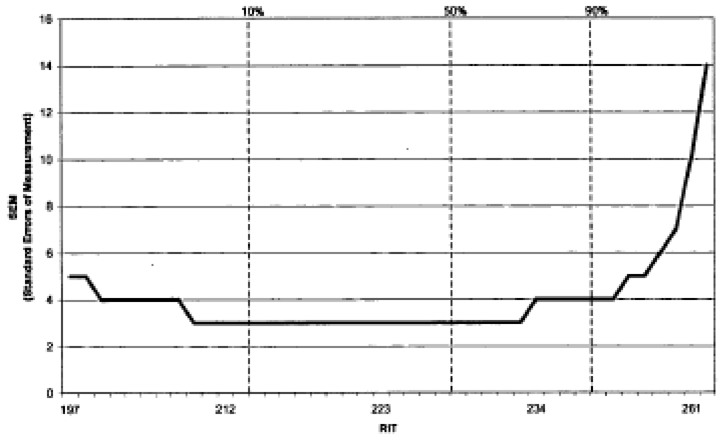
Special Education Students



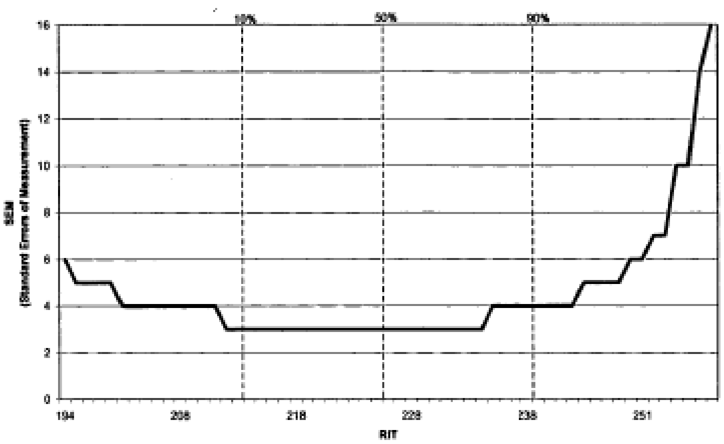
A.5 Grade 7 Reading

Test Forms

Paper and Pencil

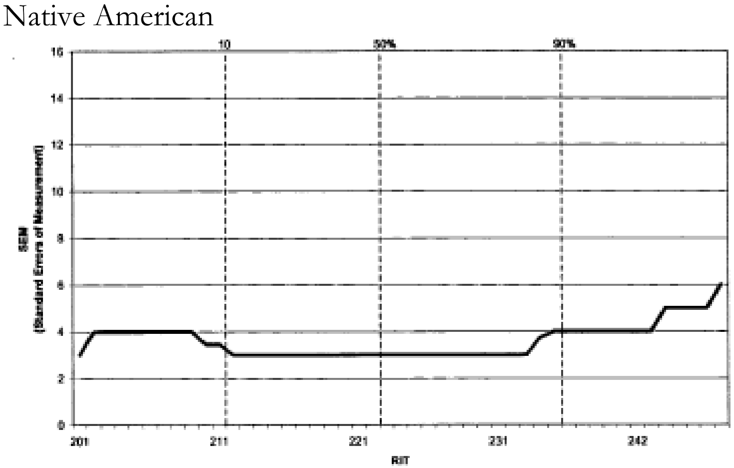


TESA

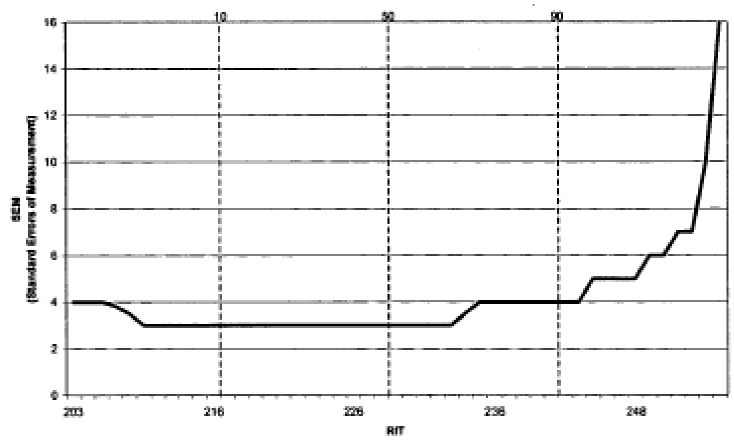


Student Ethnicity

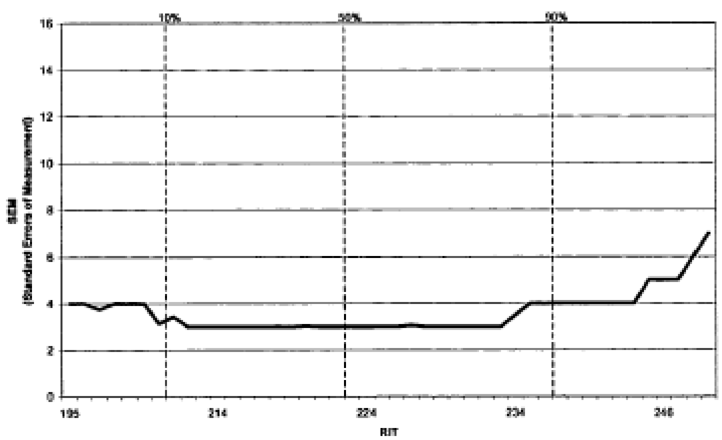
Native American



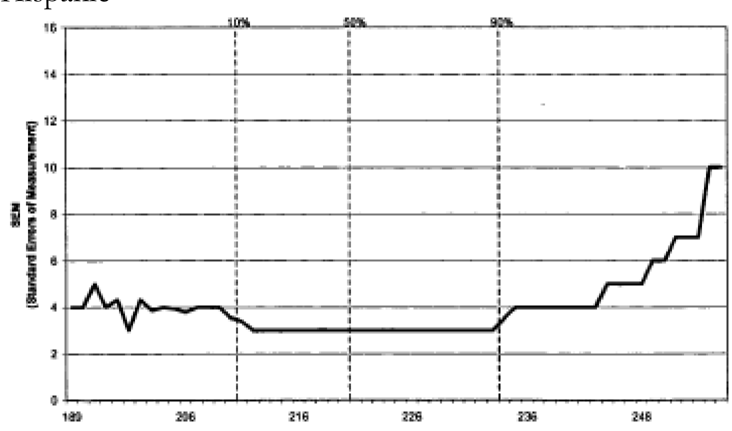
Asian



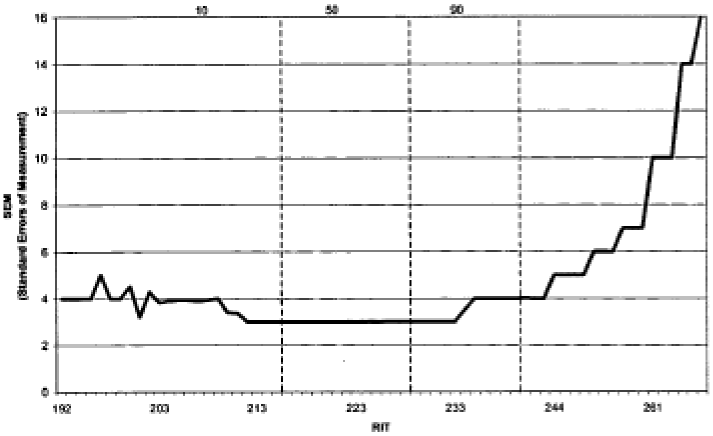
African American



Hispanic

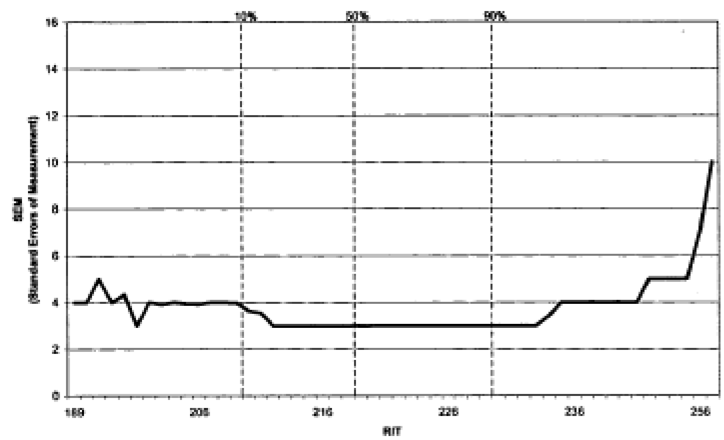


White

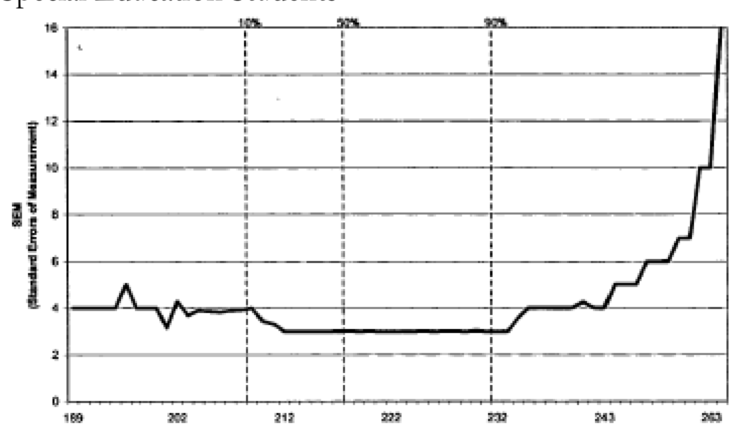


Special Populations

LEP Students



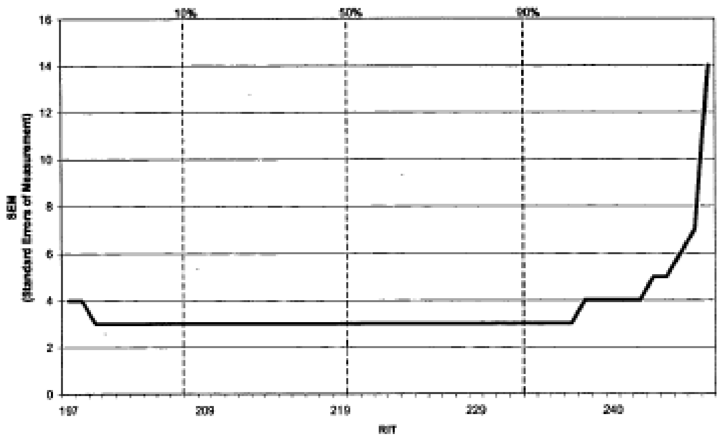
Special Education Students



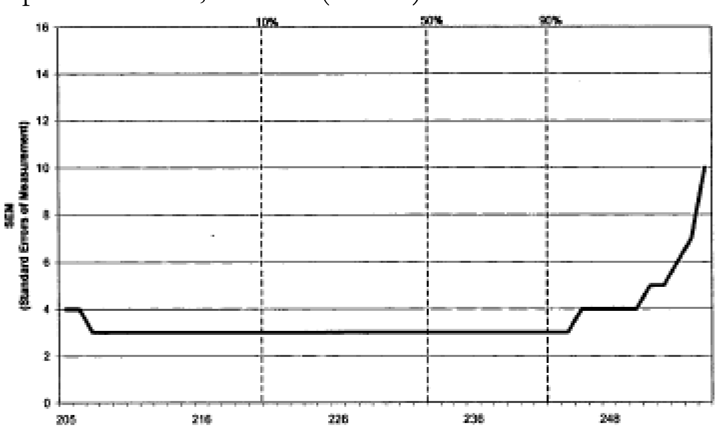
A.6 Grade 8 Reading

Test Forms

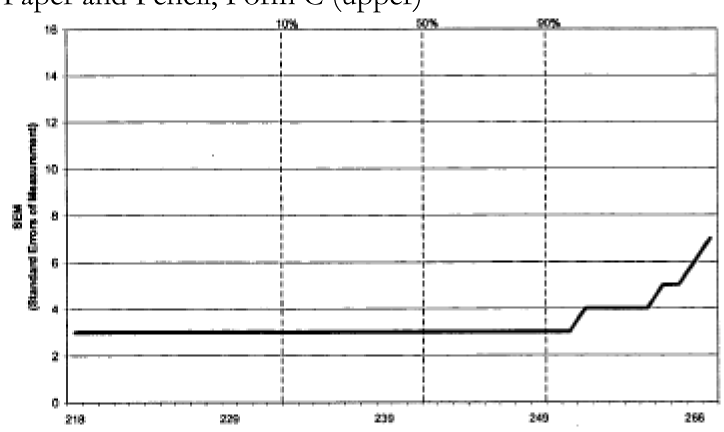
Paper and Pencil, Form A (lower)



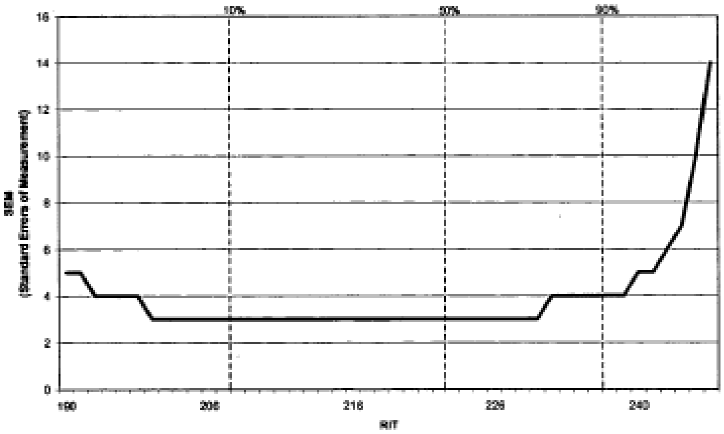
Paper and Pencil, Form B (middle)



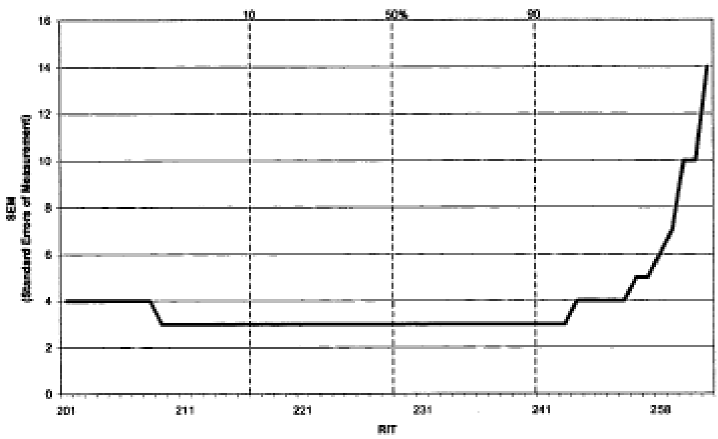
Paper and Pencil, Form C (upper)



Plain Language

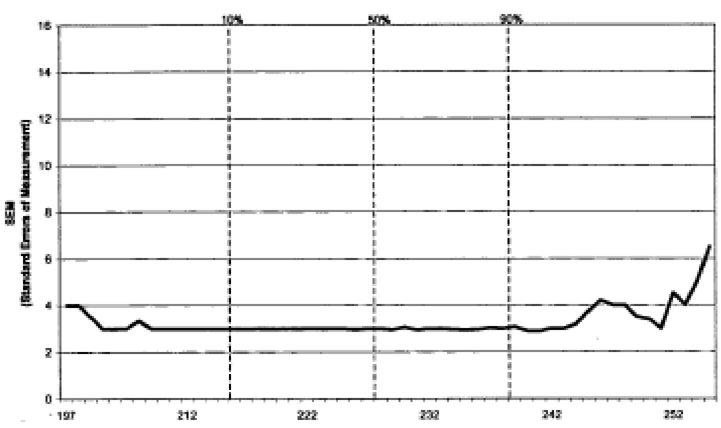


TESA

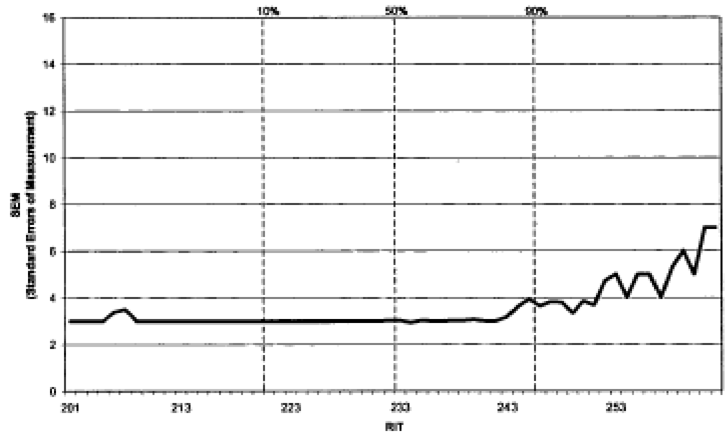


Student Ethnicity

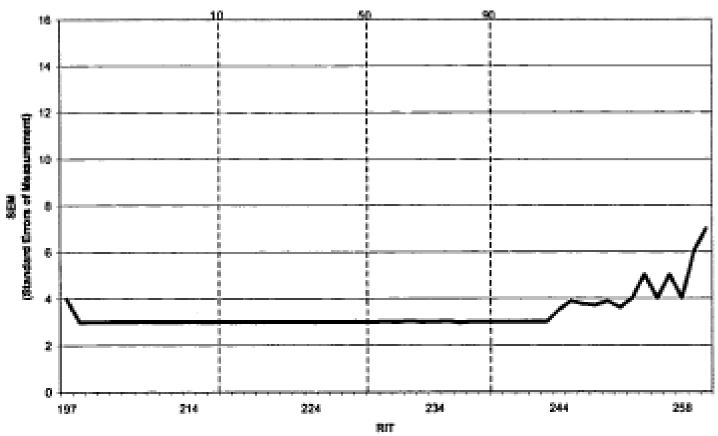
Native American



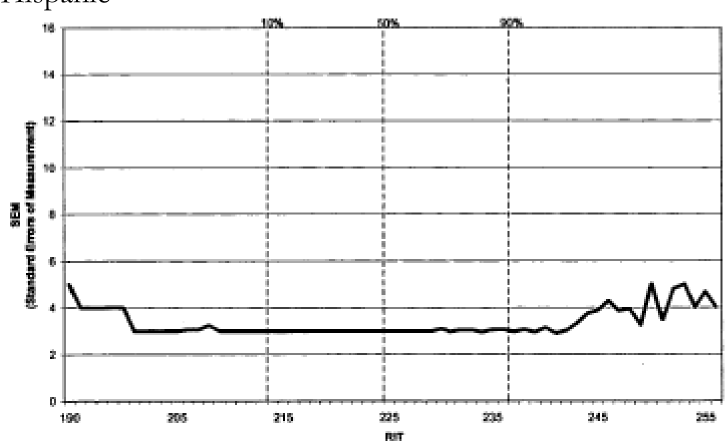
Asian



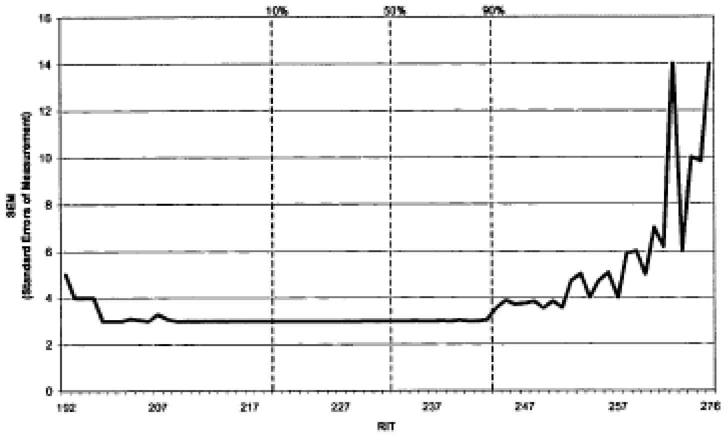
African American



Hispanic

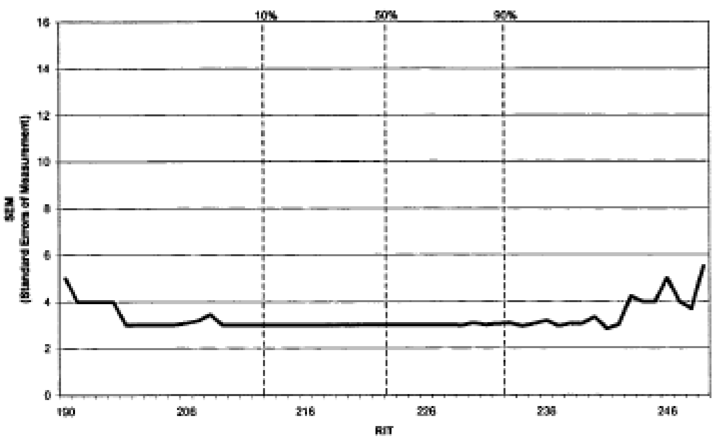


White

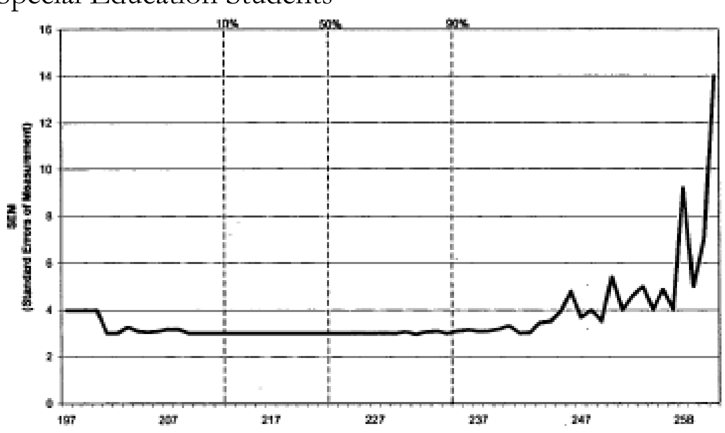


Special Populations

LEP Students



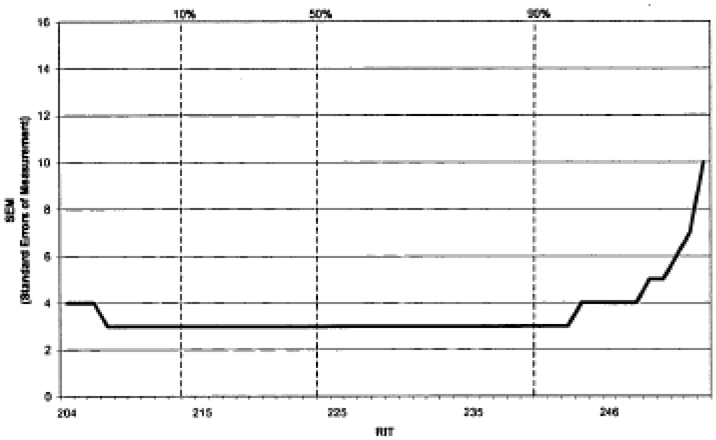
Special Education Students



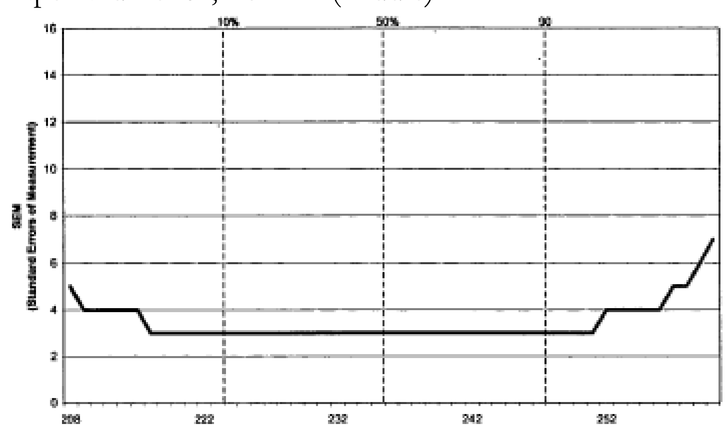
A.7 Grade 10 Reading

Test Forms

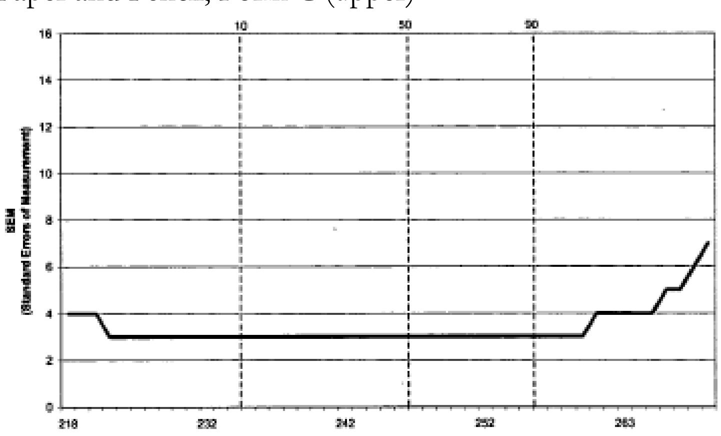
Paper and Pencil, Form A (lower)



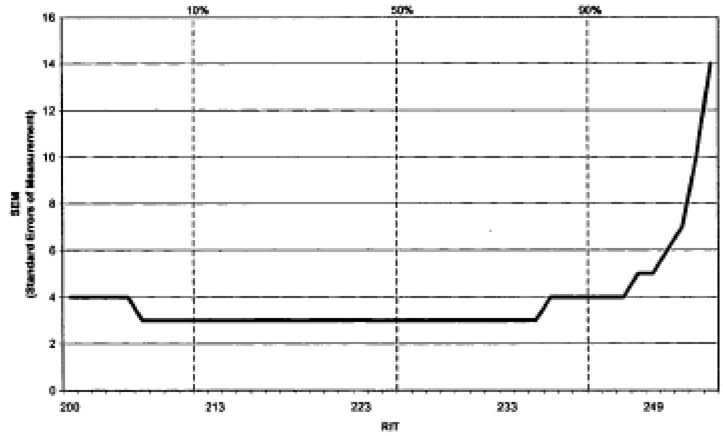
Paper and Pencil, Form B (middle)



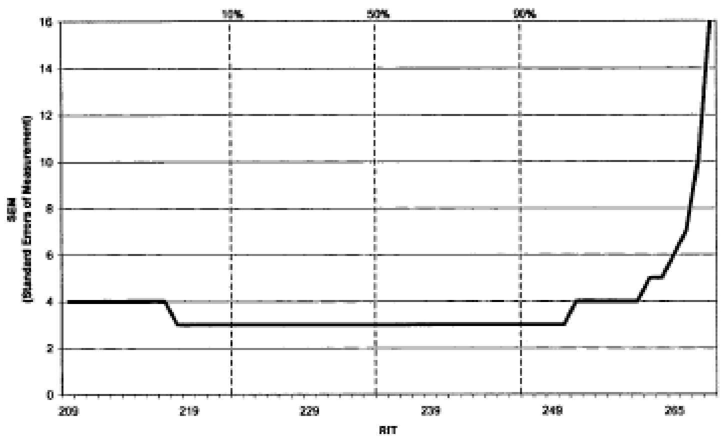
Paper and Pencil, Form C (upper)



Plain Language

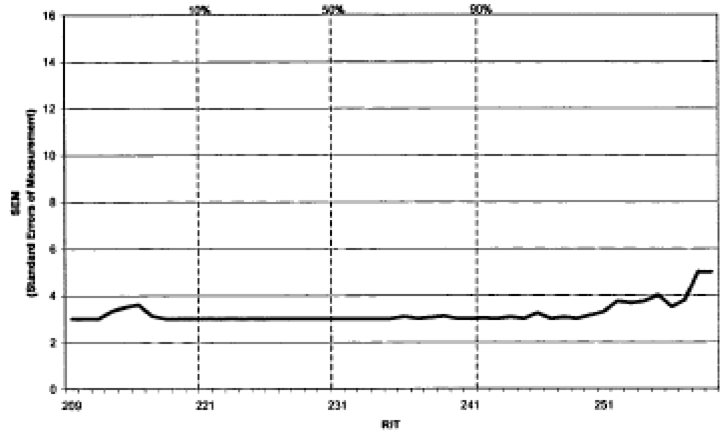


TESA

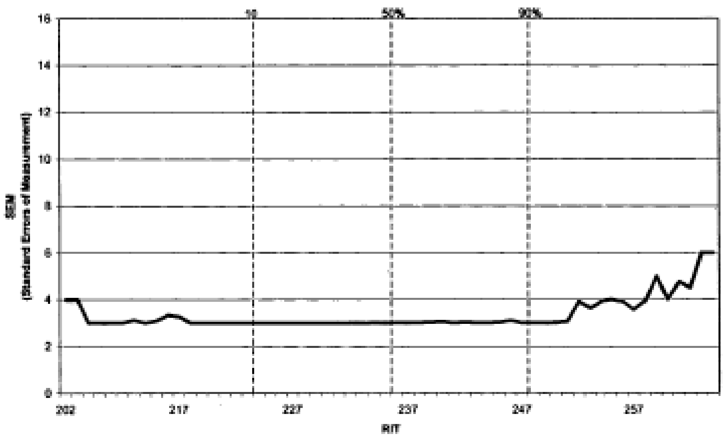


Student Ethnicity

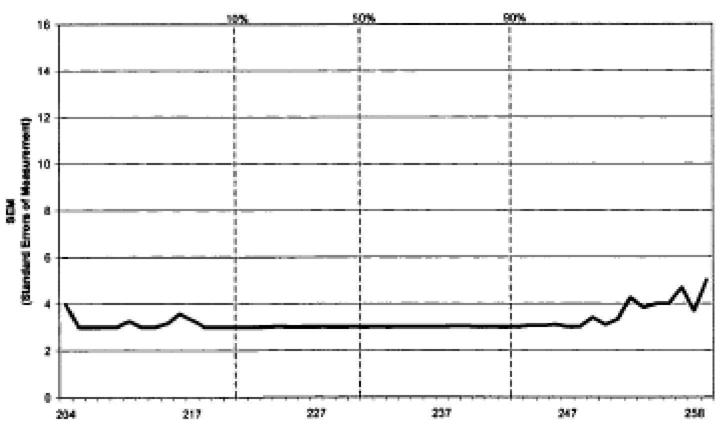
Native American



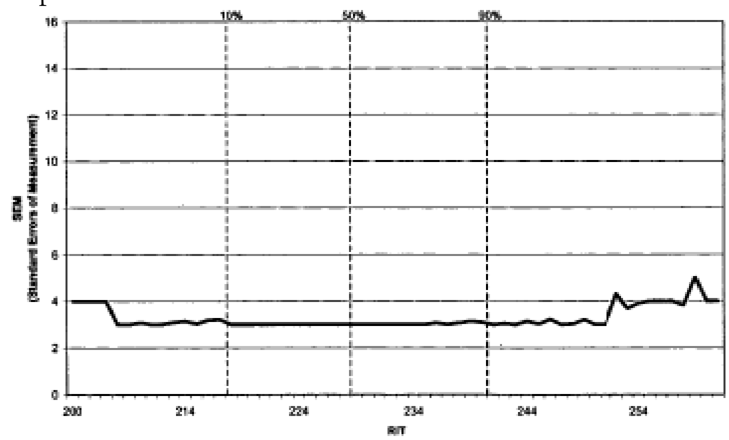
Asian



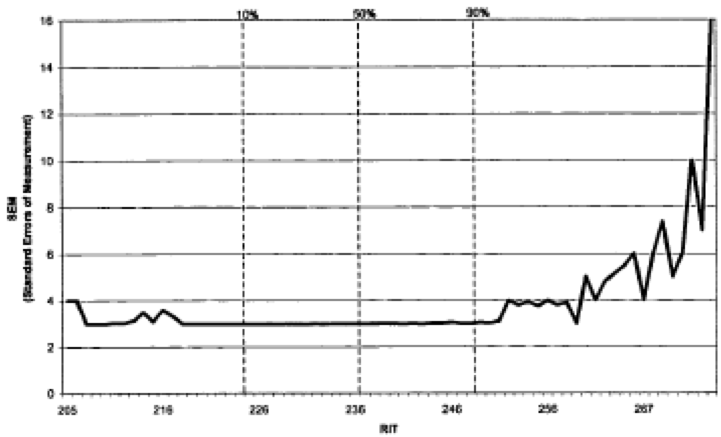
African American



Hispanic

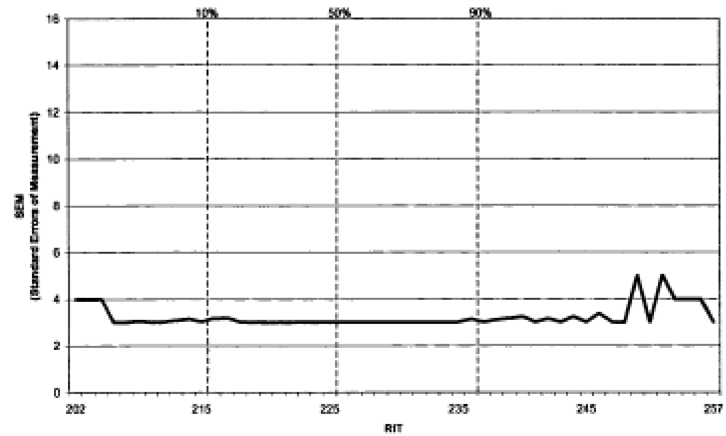


White

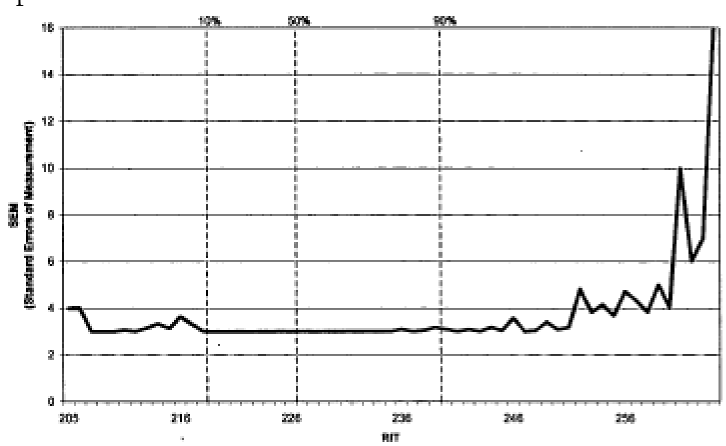


Special Populations

LEP Students



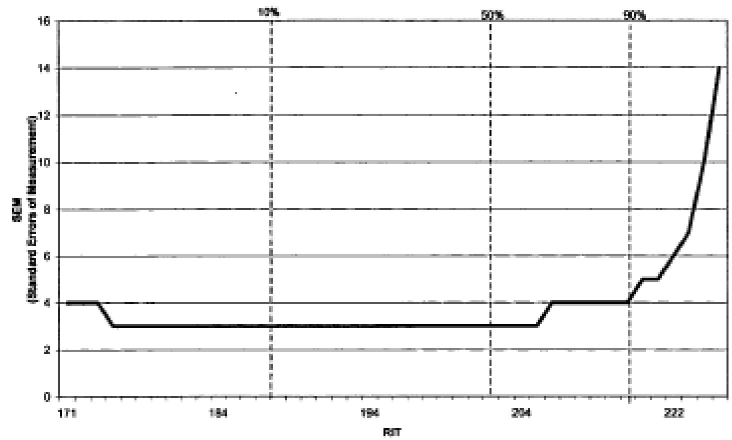
Special Education Students



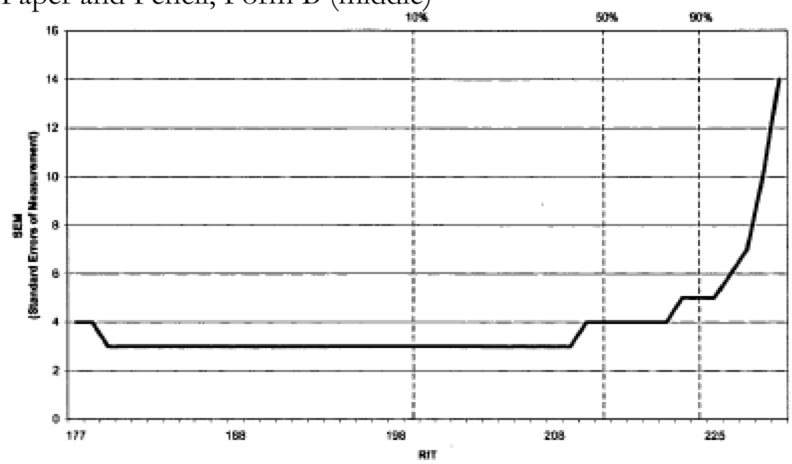
A.8 Grade 3 Mathematics

Test Forms

Paper and Pencil, Form A (lower)

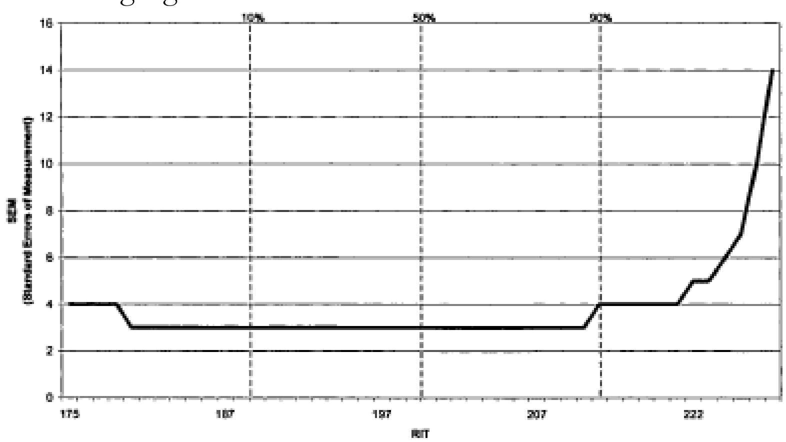


Paper and Pencil, Form B (middle)

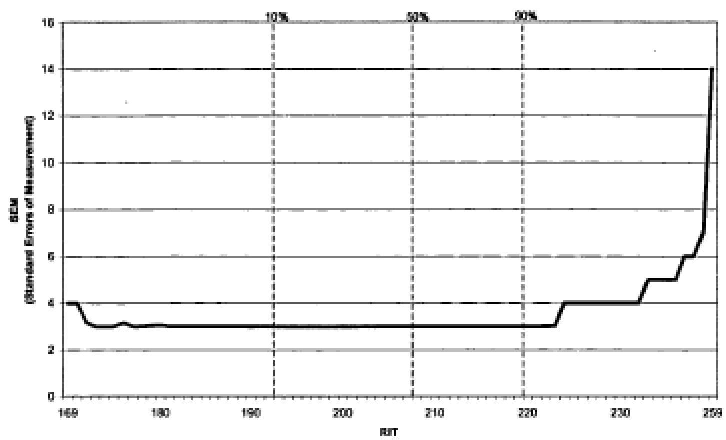


Paper and Pencil, Form C (upper)

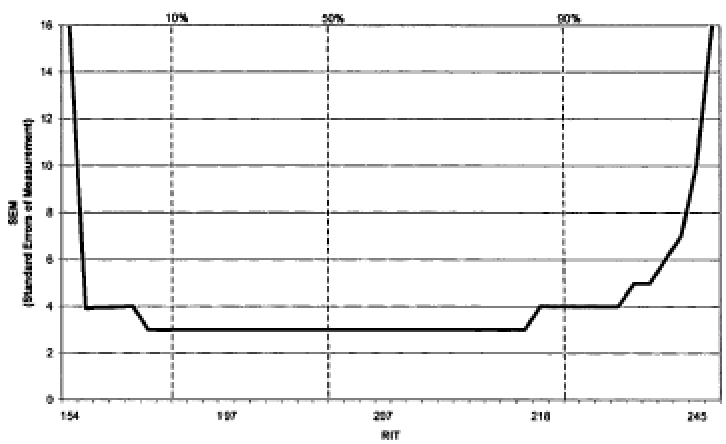
Plain Language



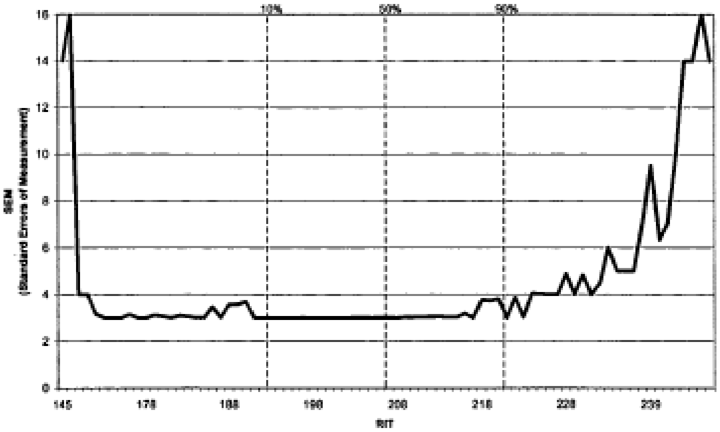
TESA Adaptive



TESA Constrained

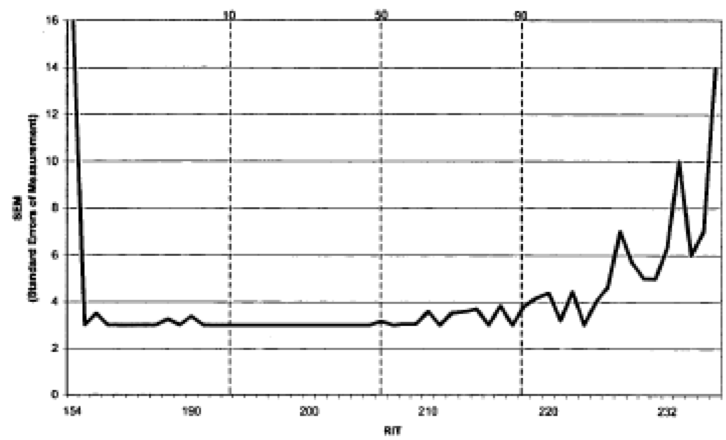


TESA Wide Range

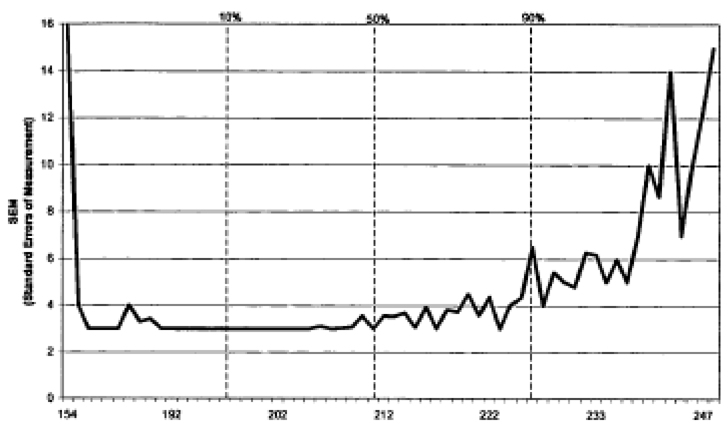


Student Ethnicity

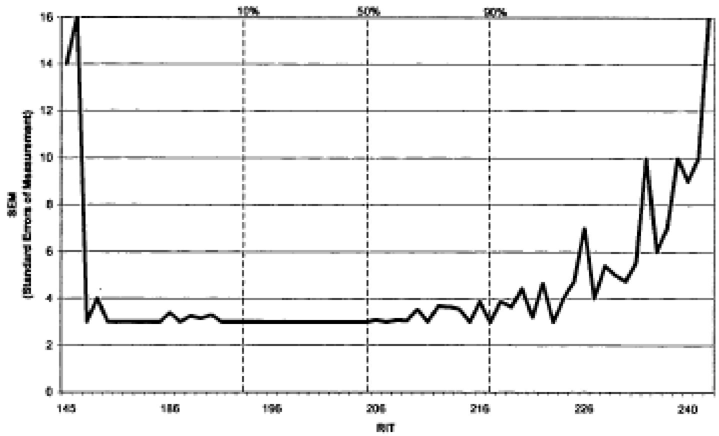
Native American



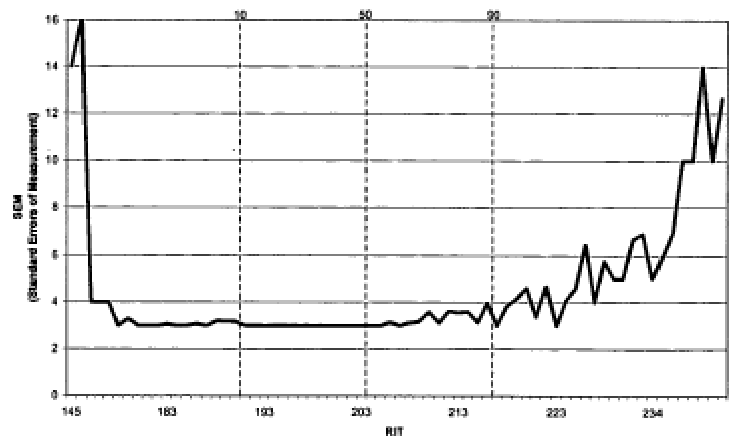
Asian



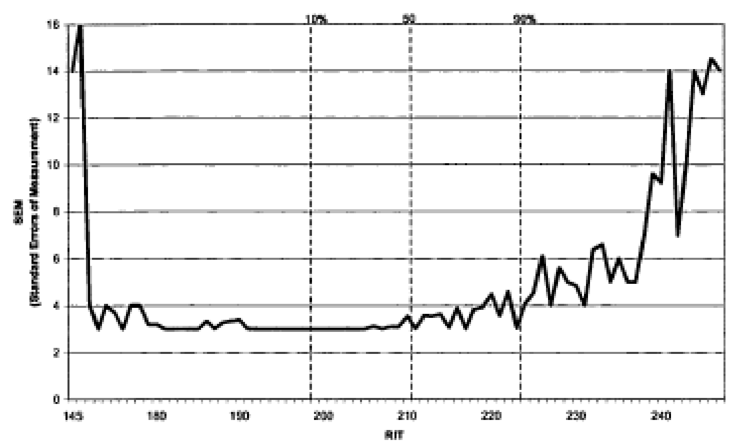
African American



Hispanic

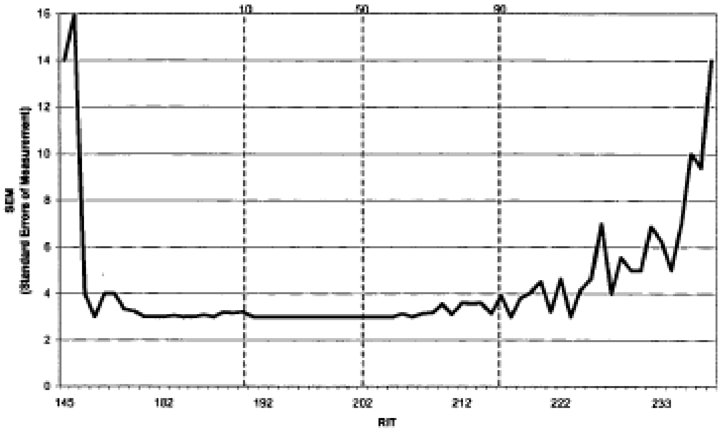


White

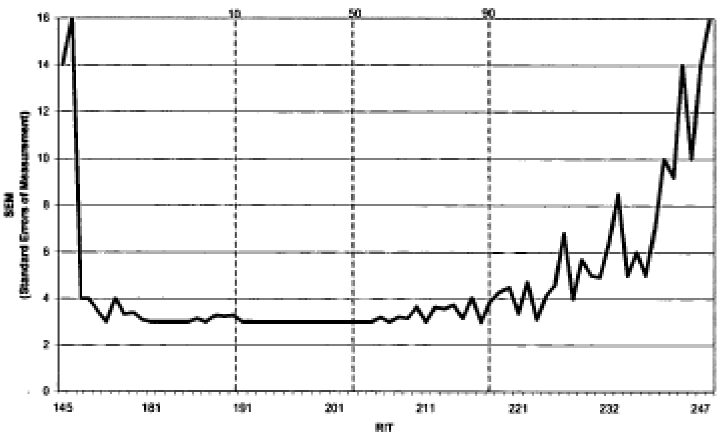


Special Populations

LEP Students



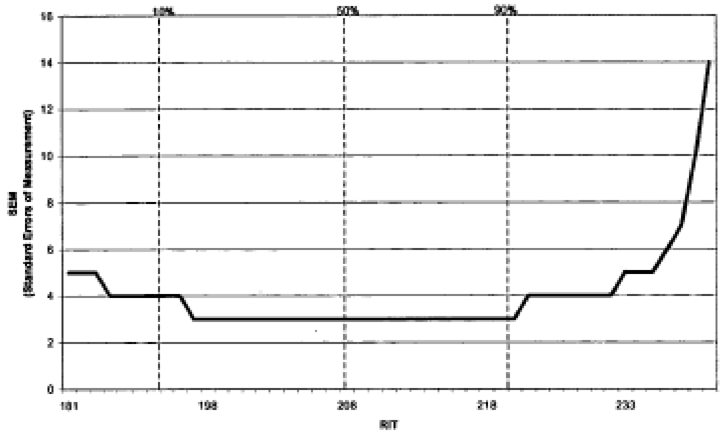
Special Education Students



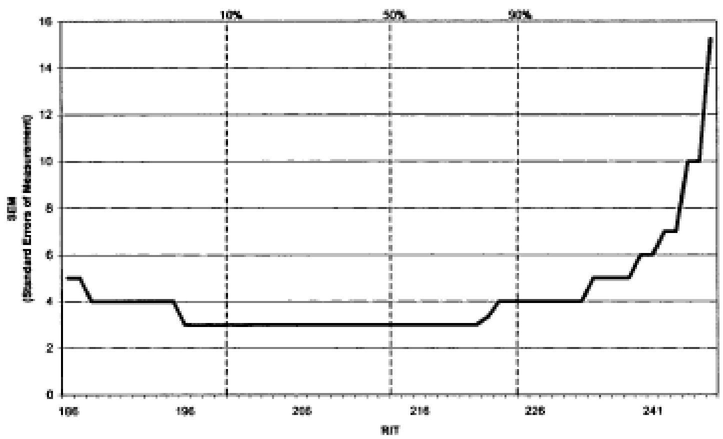
A.9 Grade 4 Mathematics

Test Forms

Paper and Pencil

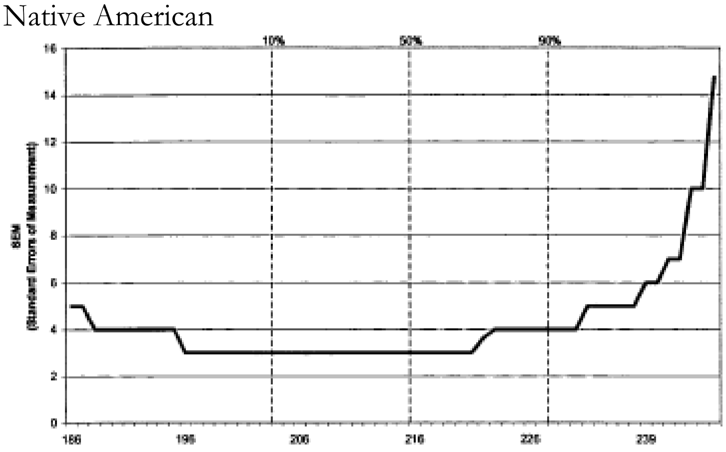


TESA

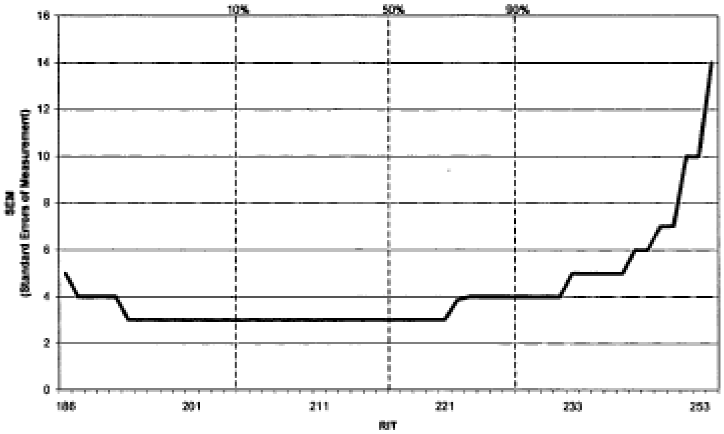


Student Ethnicity

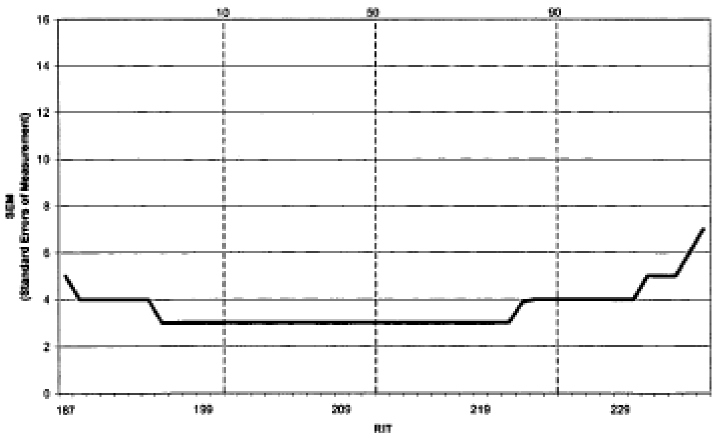
Native American



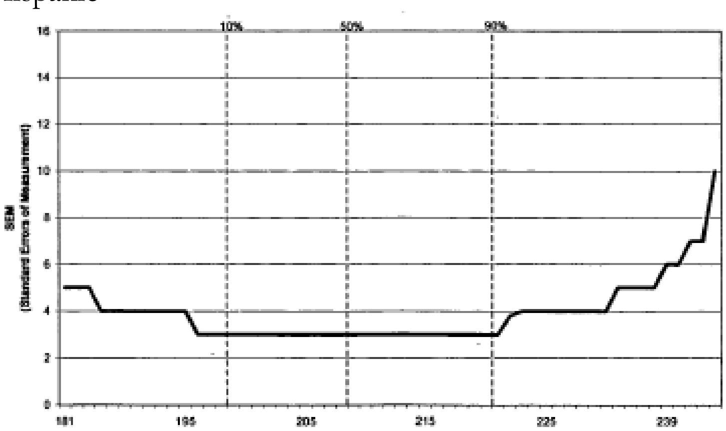
Asian



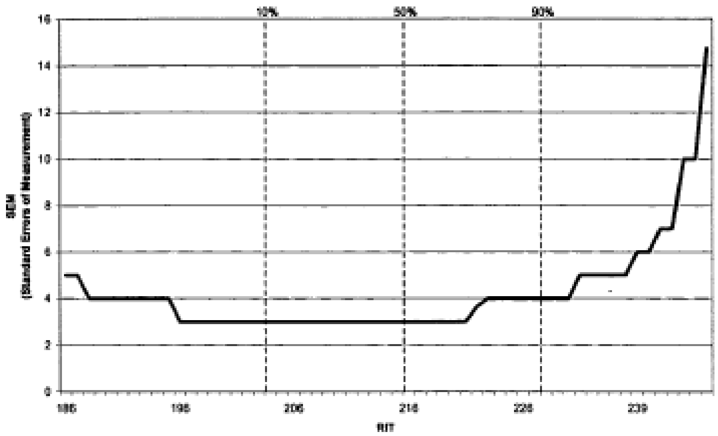
African American



Hispanic

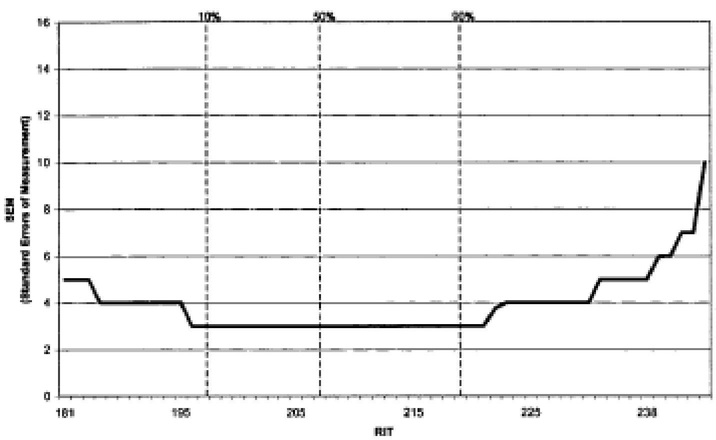


White

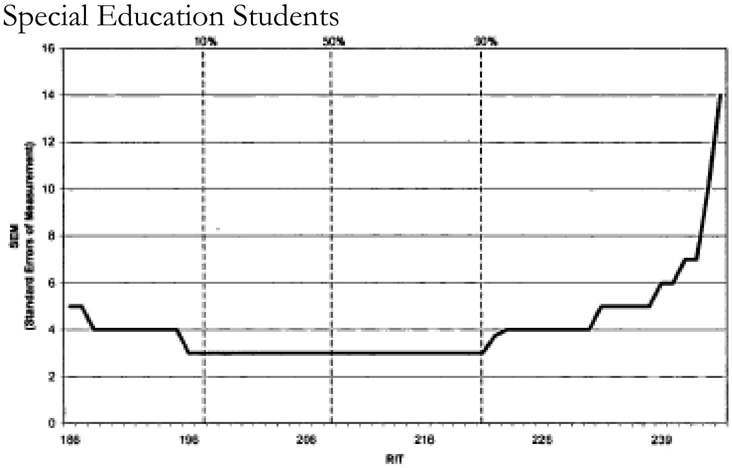


Special Populations

LEP Students



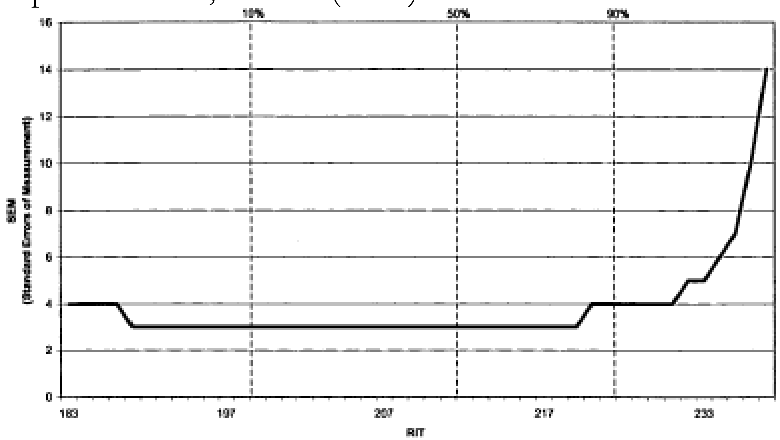
Special Education Students



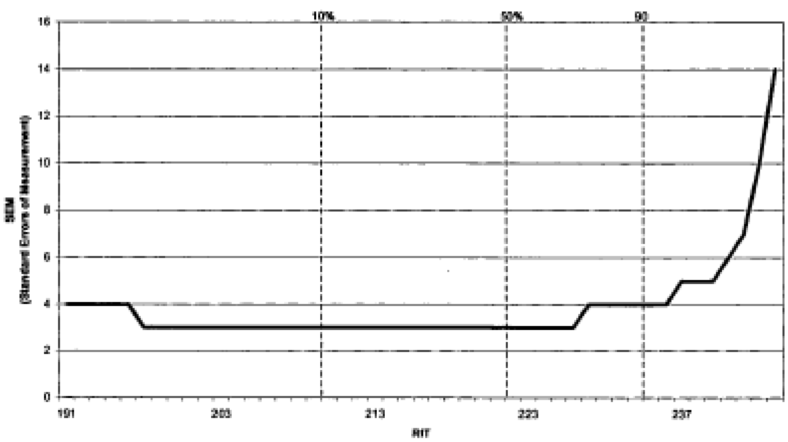
A.10 Grade 5 Mathematics

Test Forms

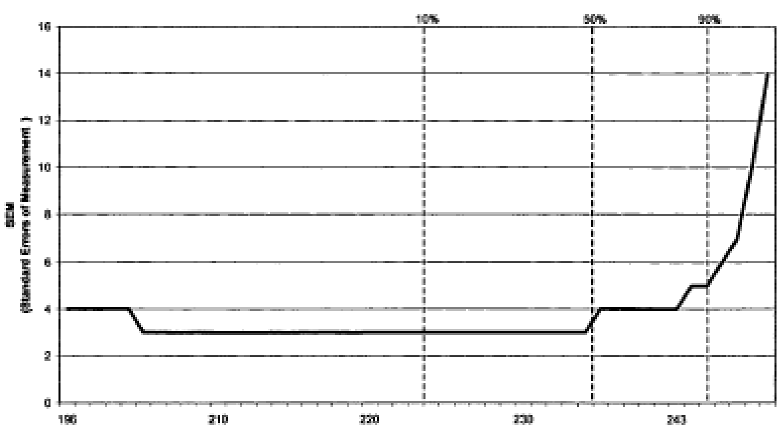
Paper and Pencil, Form A (lower)



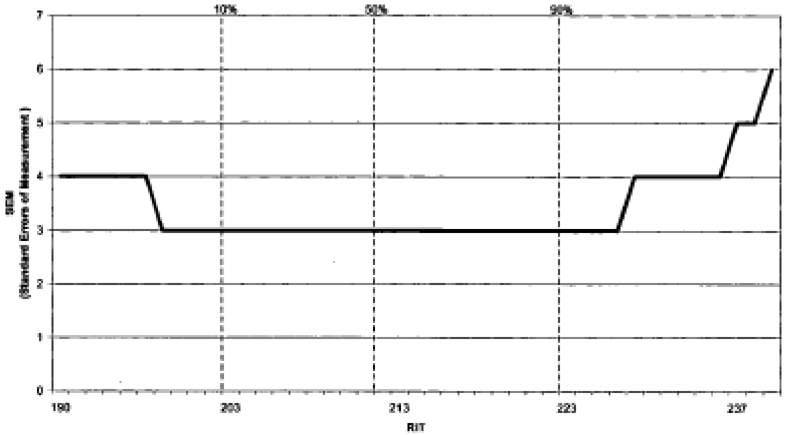
Paper and Pencil, Form B (middle)



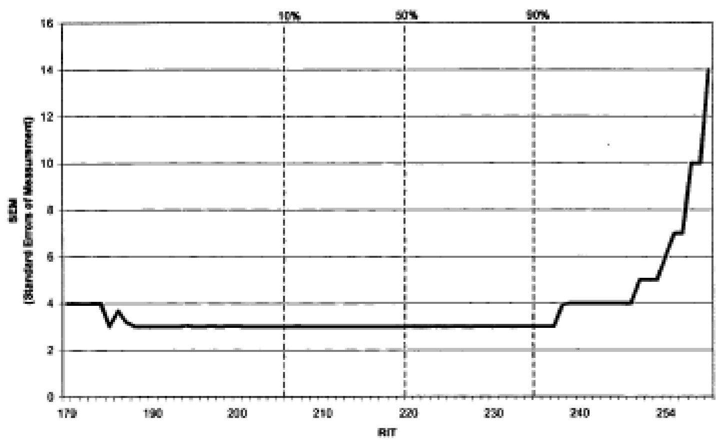
Paper and Pencil, Form C (upper)



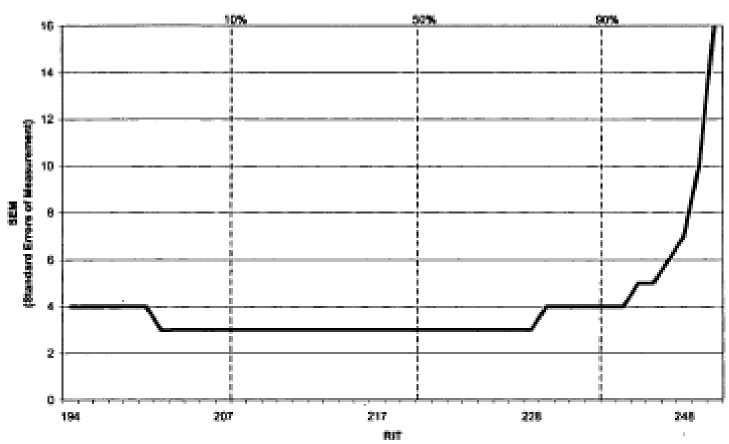
Plain Language



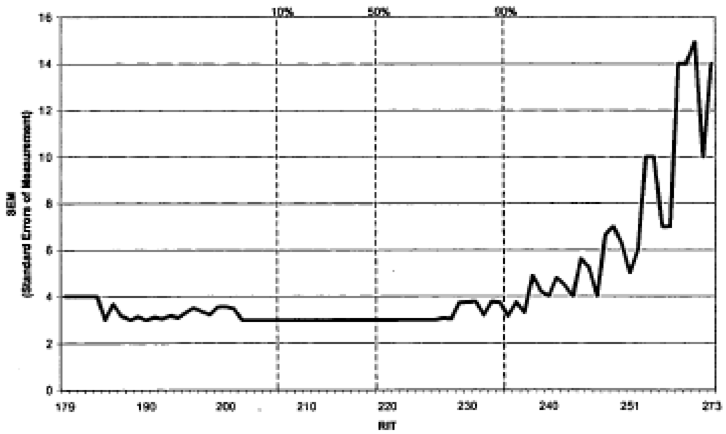
TESA Adaptive



TESA Constrained

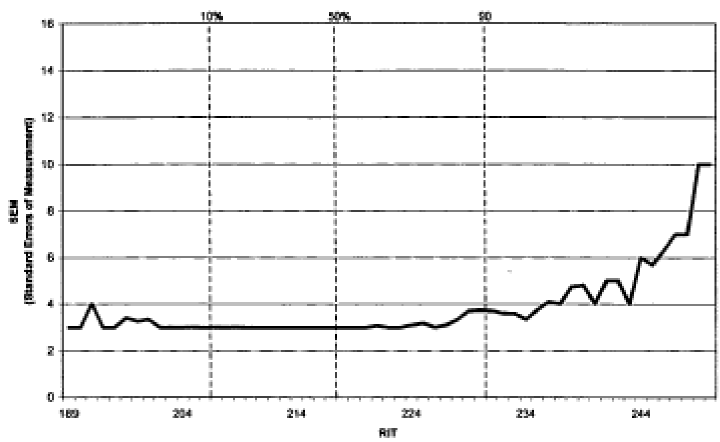


TESA Wide Range

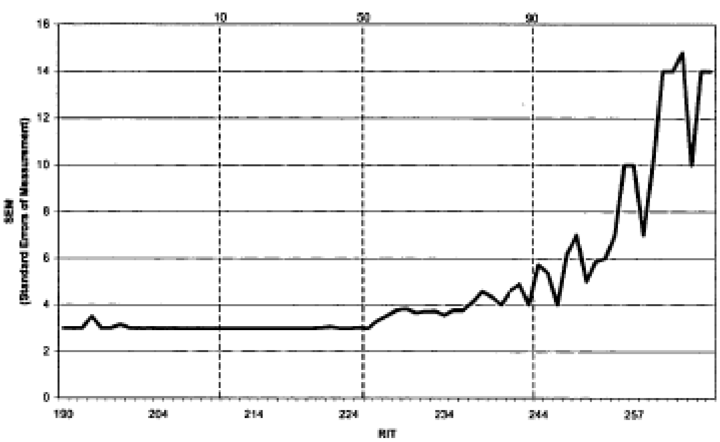


Student Ethnicity

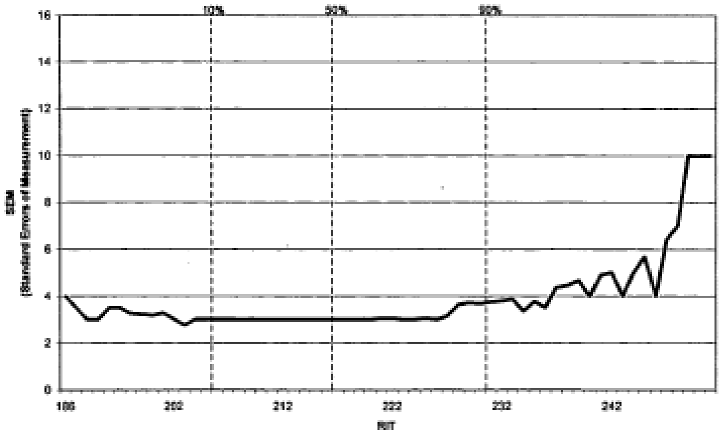
Native American



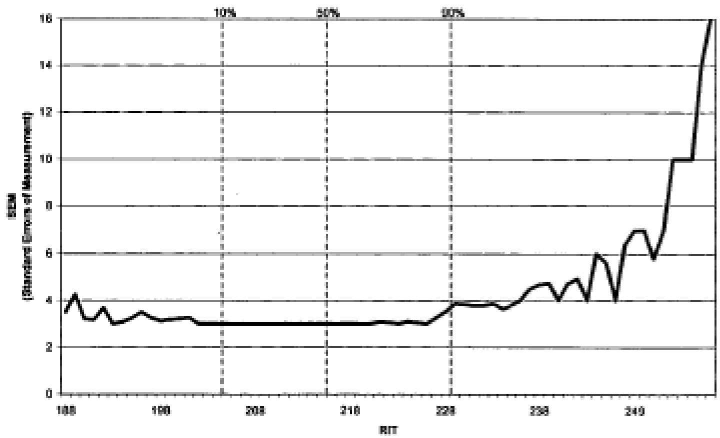
Asian



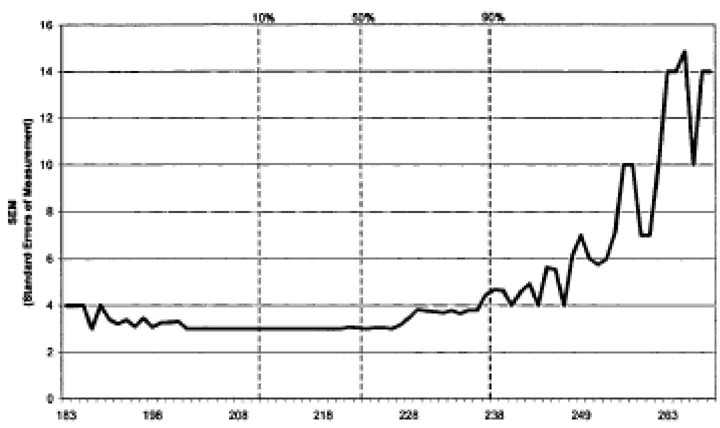
African American



Hispanic

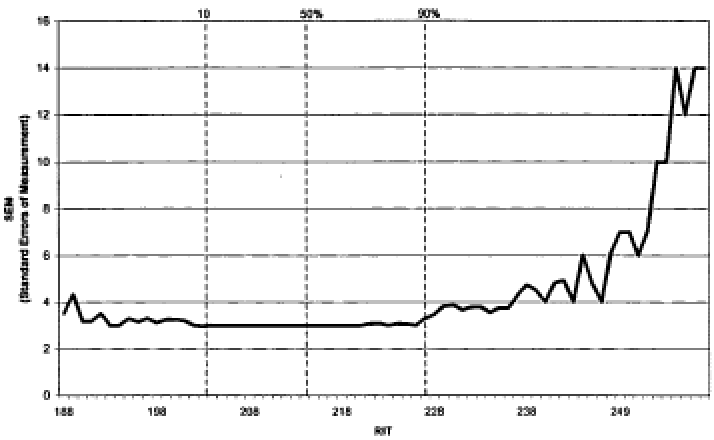


White

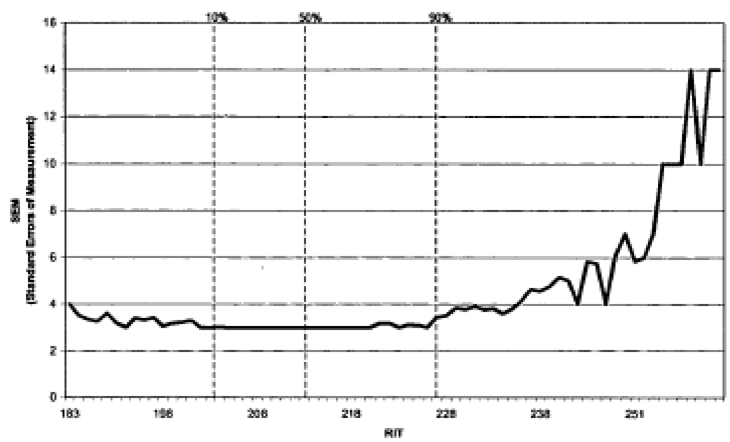


Special Populations

LEP Students



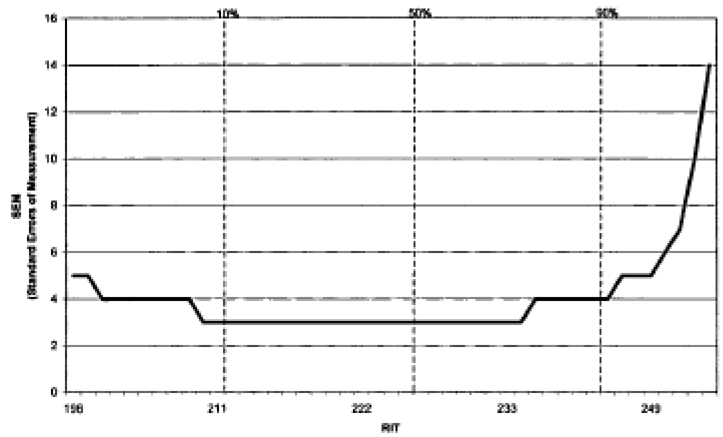
Special Education Students



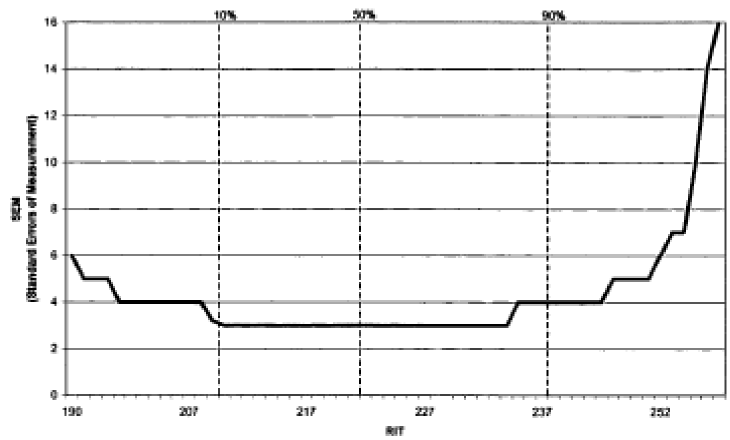
A.11 Grade 6 Mathematics

Test Forms

Paper and Pencil

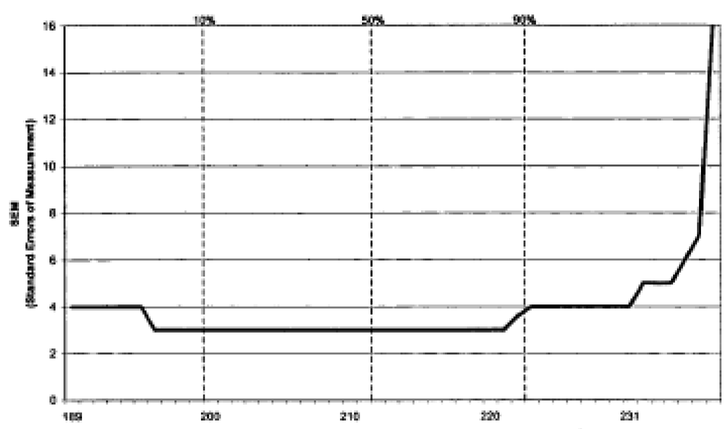


TESA

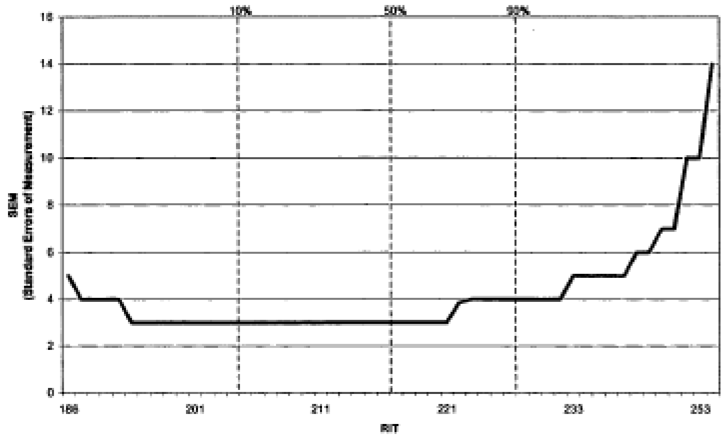


Student Ethnicity

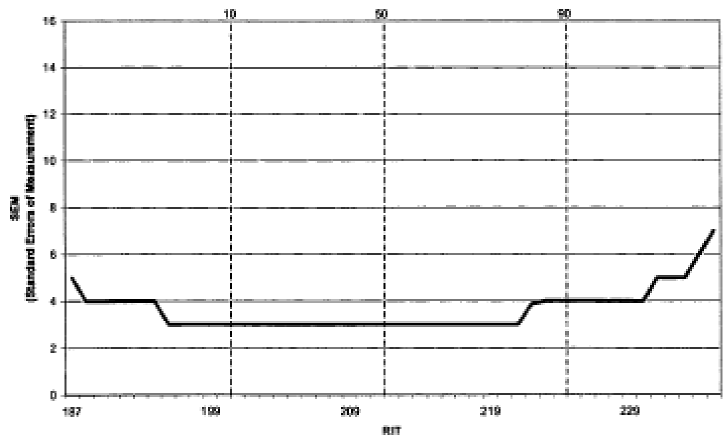
Native American



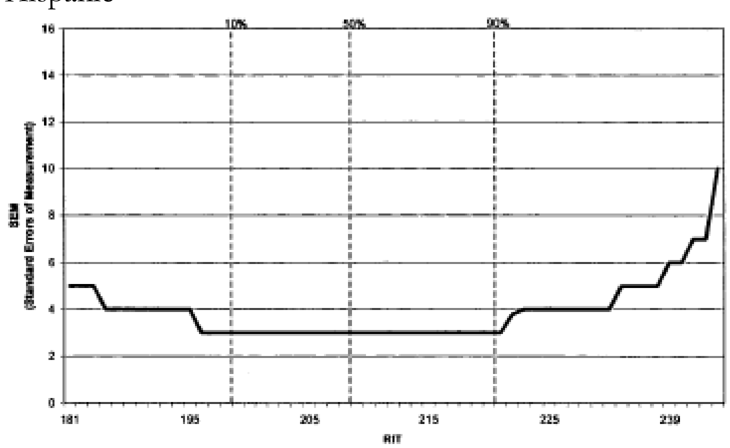
Asian



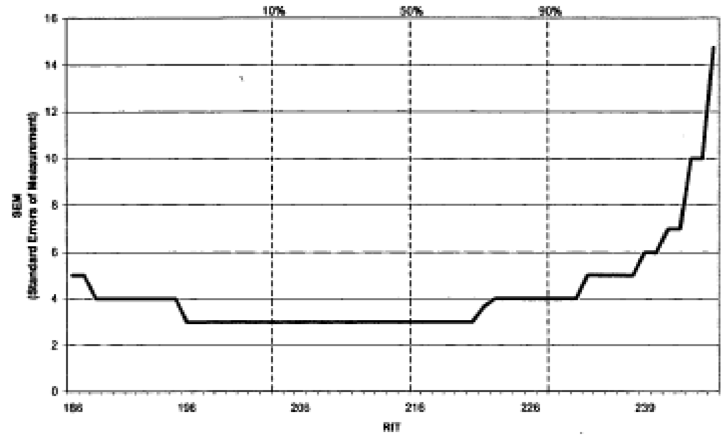
African American



Hispanic

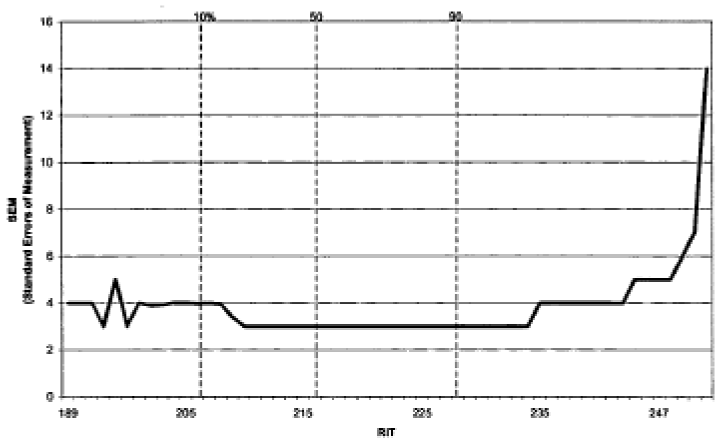


White

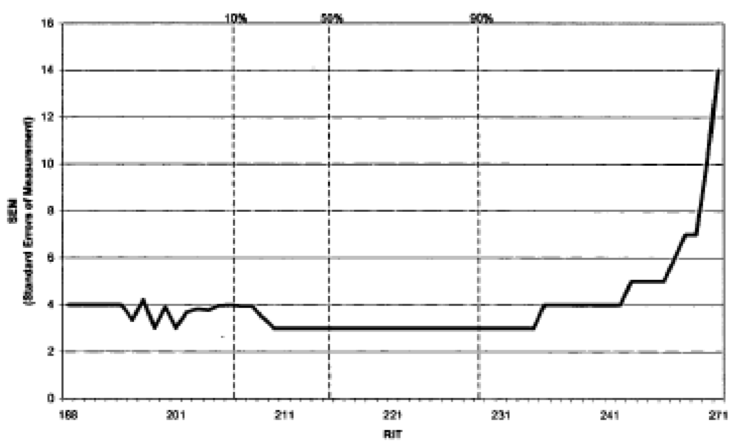


Special Populations

LEP Students



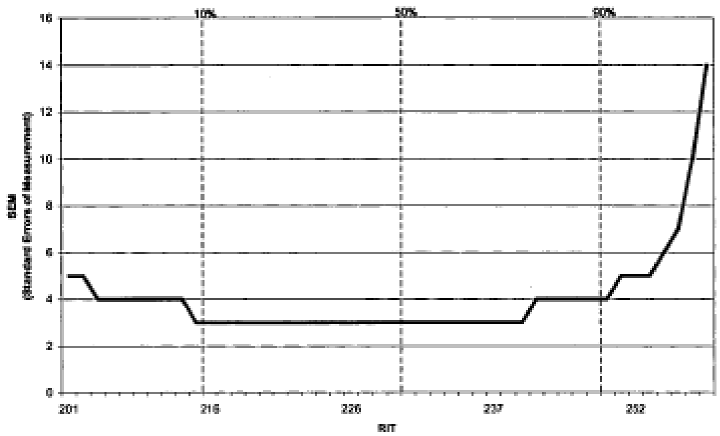
Special Education Students



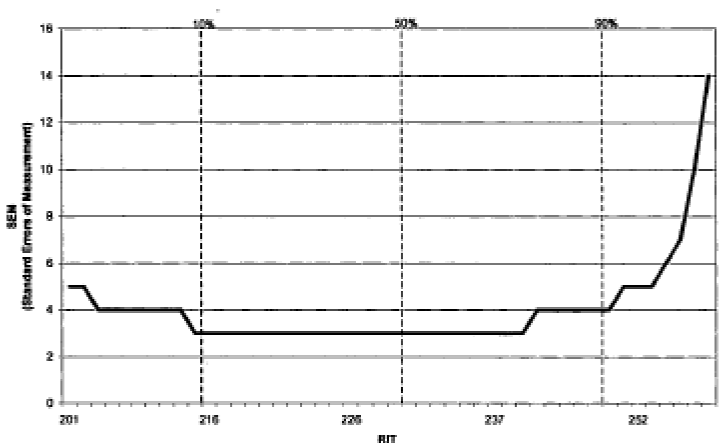
A.12 Grade 7 Mathematics

Test Forms

Paper and Pencil

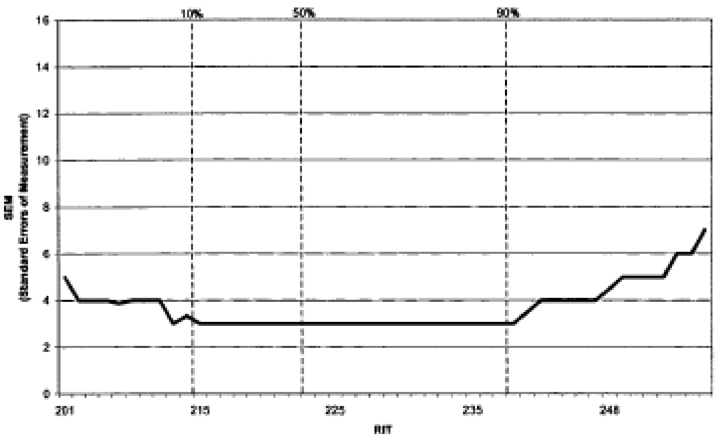


TESA

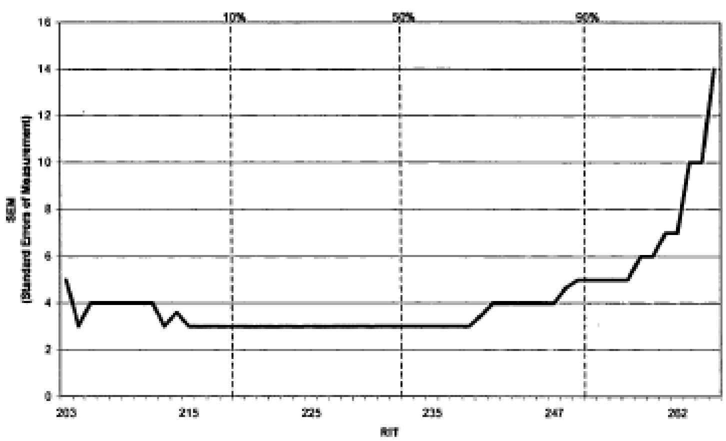


Student Ethnicity

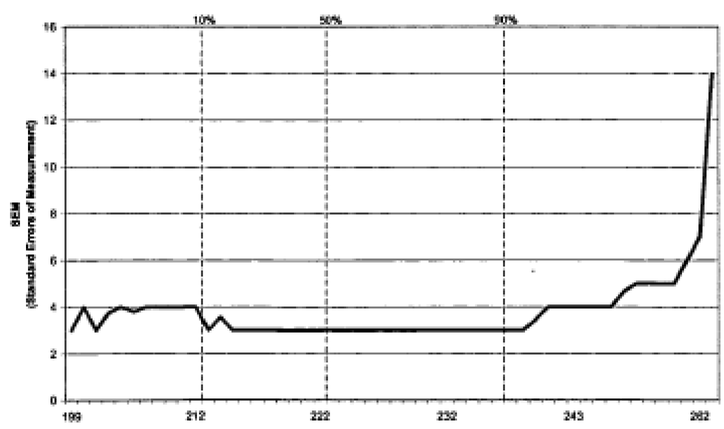
Native American



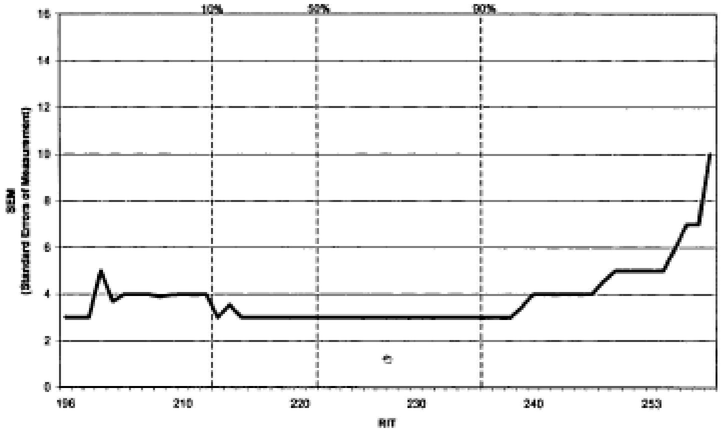
Asian



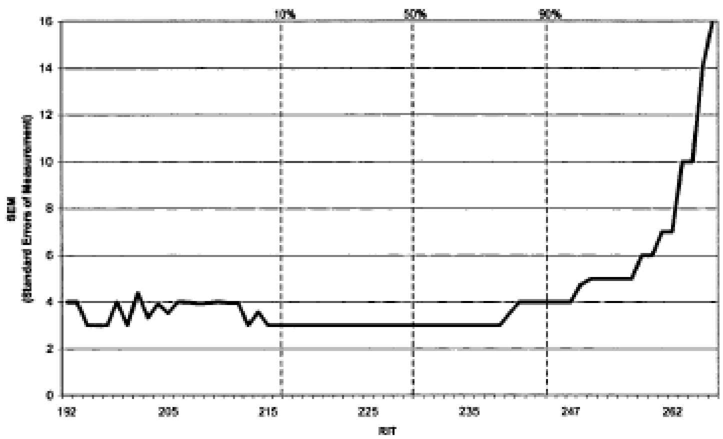
African American



Hispanic

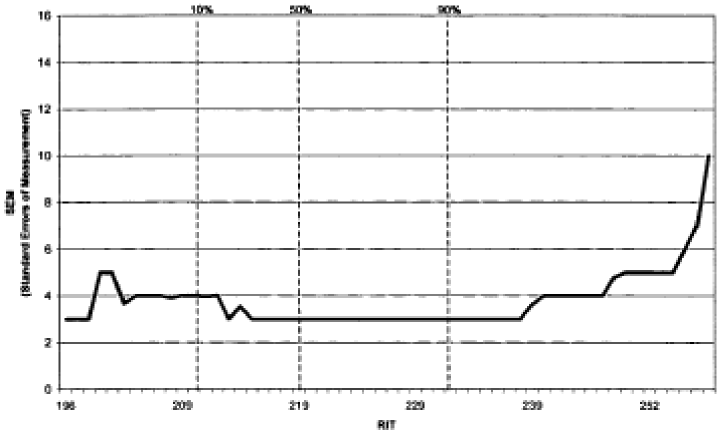


White

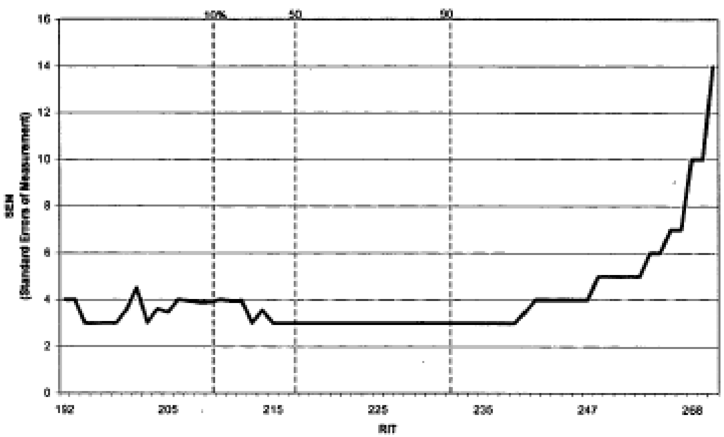


Special Populations

LEP Students



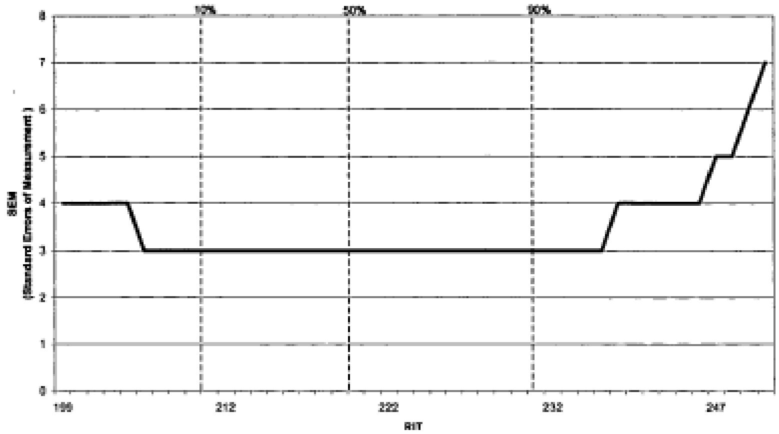
Special Education Students



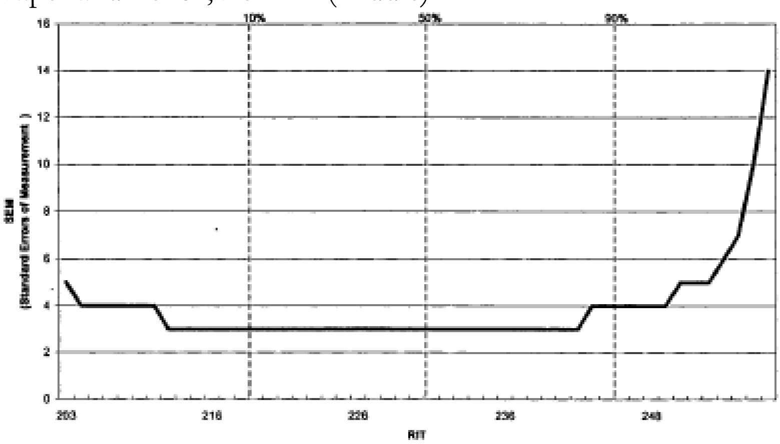
A.13 Grade 8 Mathematics

Test Forms

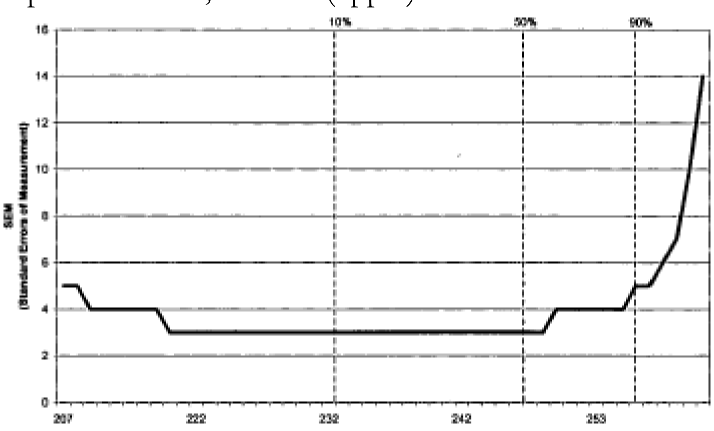
Paper and Pencil, Form A (lower)



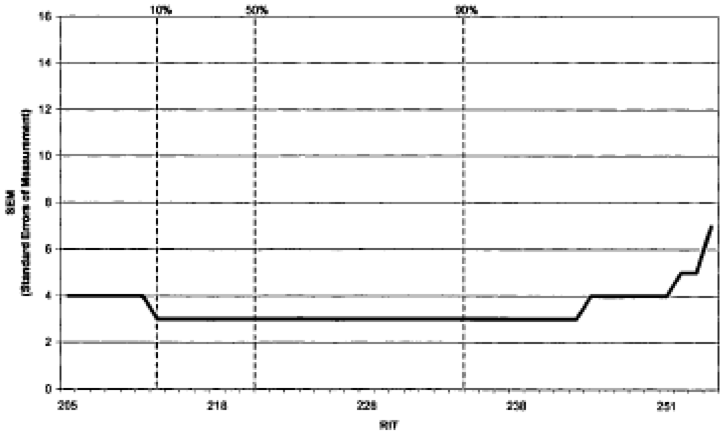
Paper and Pencil, Form B (middle)



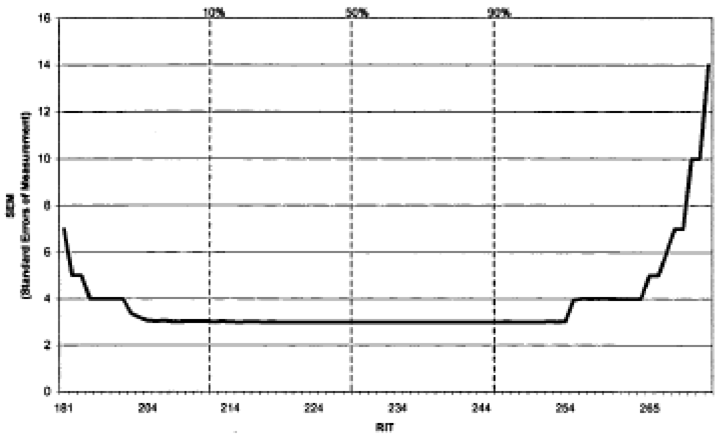
Paper and Pencil, Form C (upper)



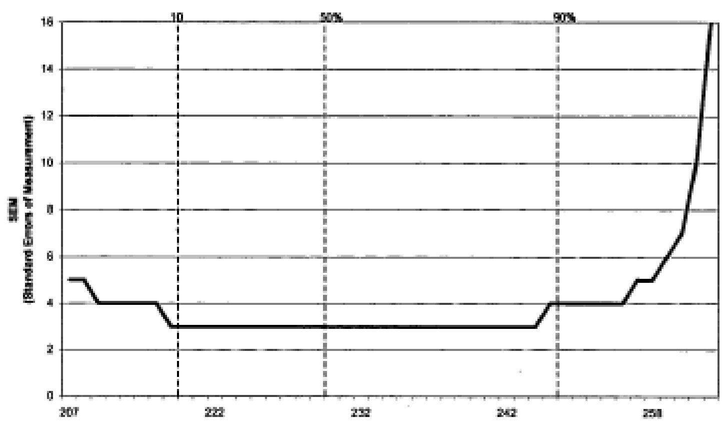
Plain Language



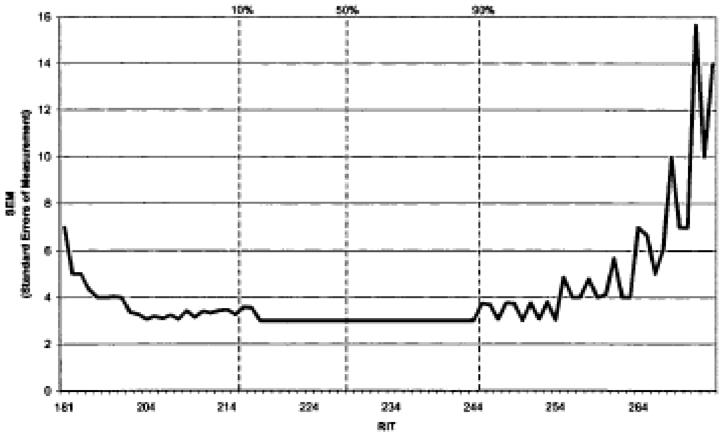
TESA Adaptive



TESA Constrained

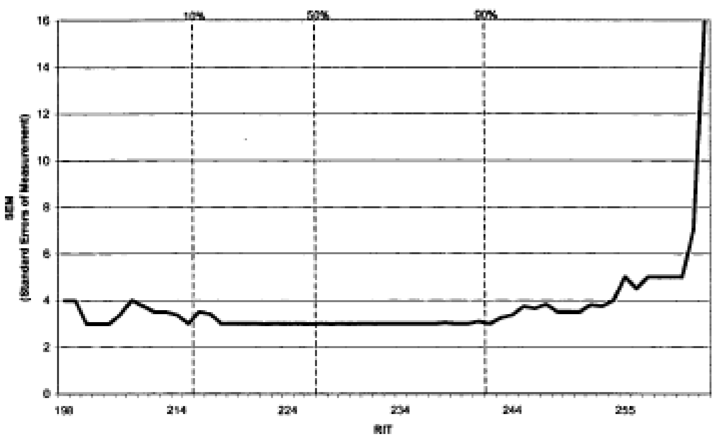


TESA Wide Range

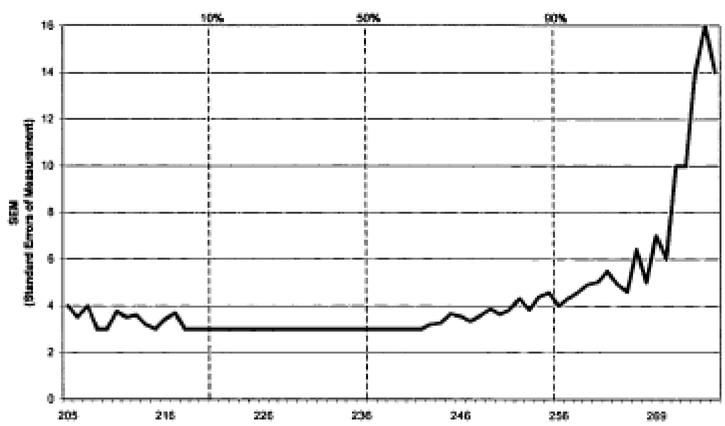


Student Ethnicity

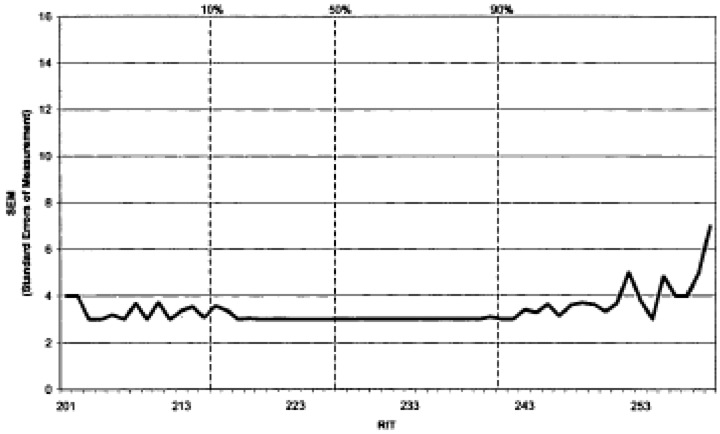
Native American



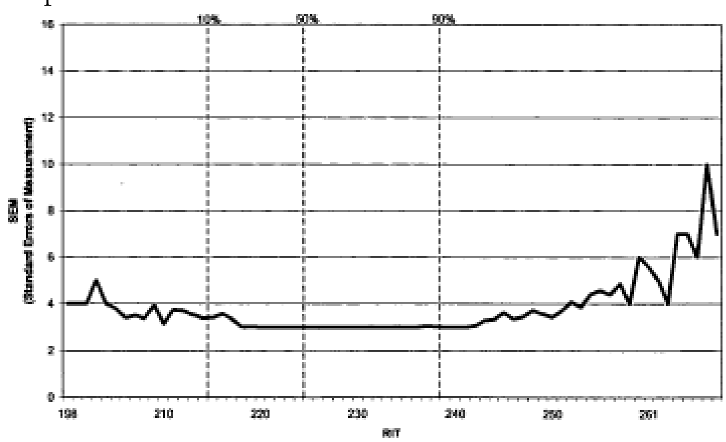
Asian



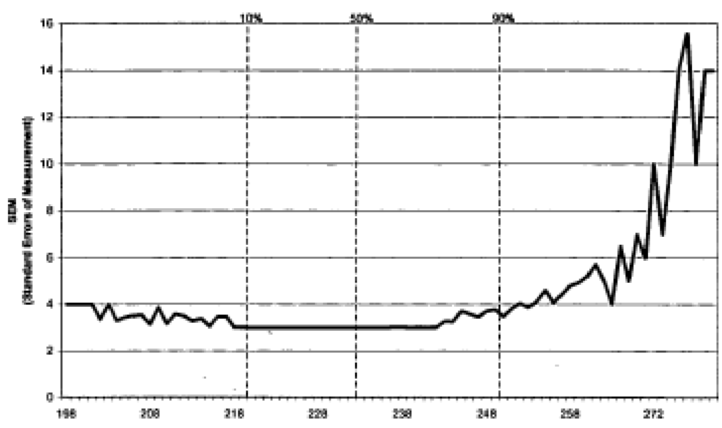
African American



Hispanic

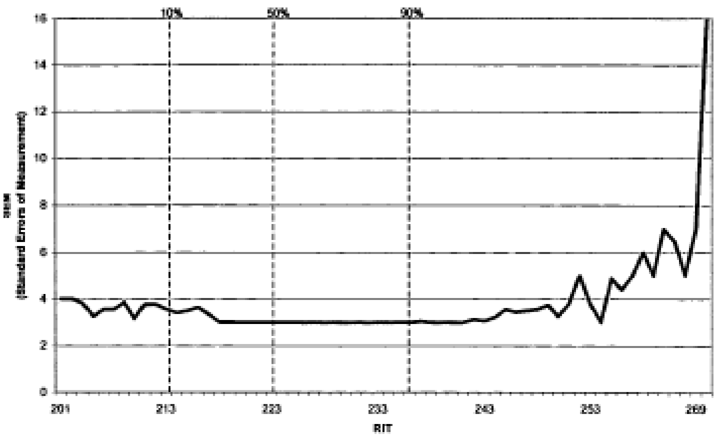


White

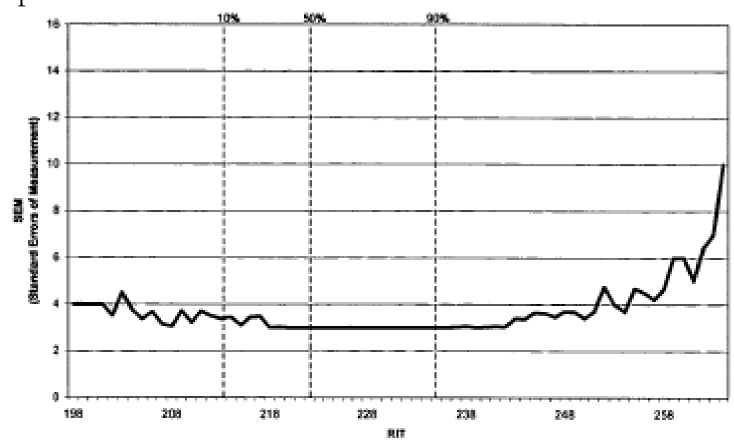


Special Populations

LEP Students



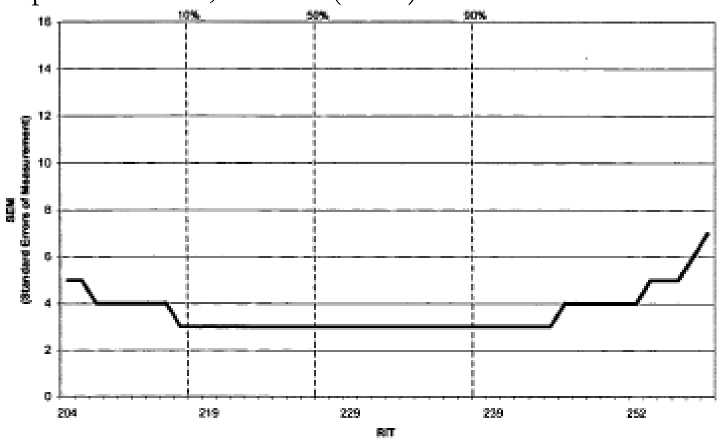
Special Education Students



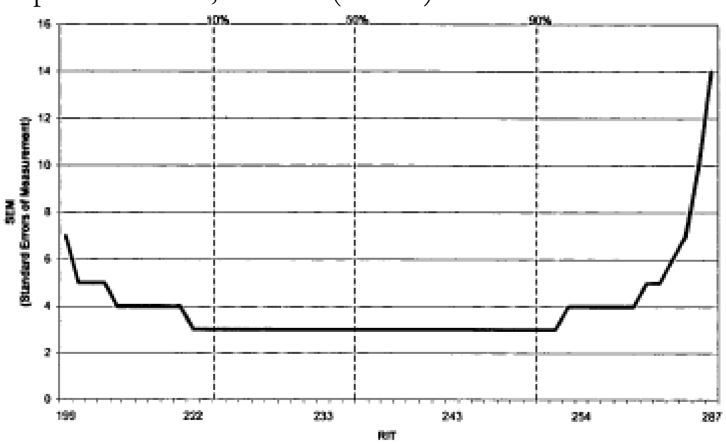
A.14 Grade 10 Mathematics

Test Forms

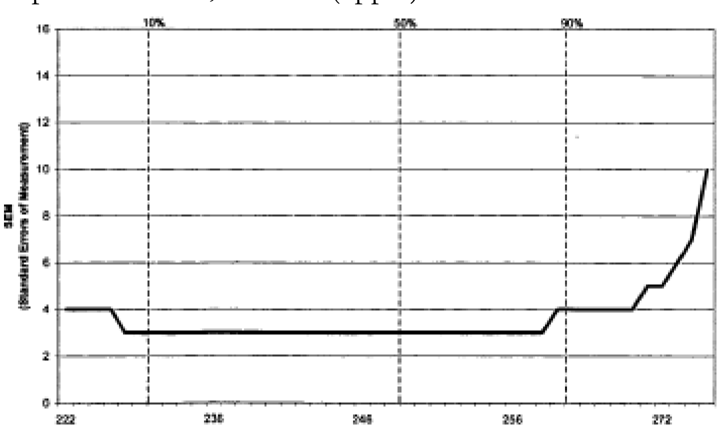
Paper and Pencil, Form A (lower)



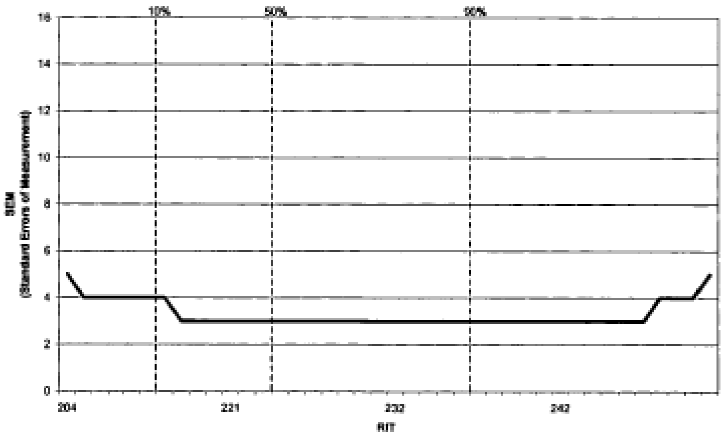
Paper and Pencil, Form B (middle)



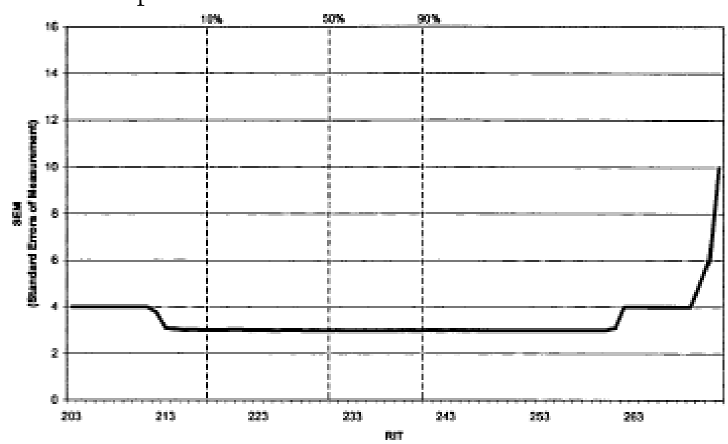
Paper and Pencil, Form C (upper)



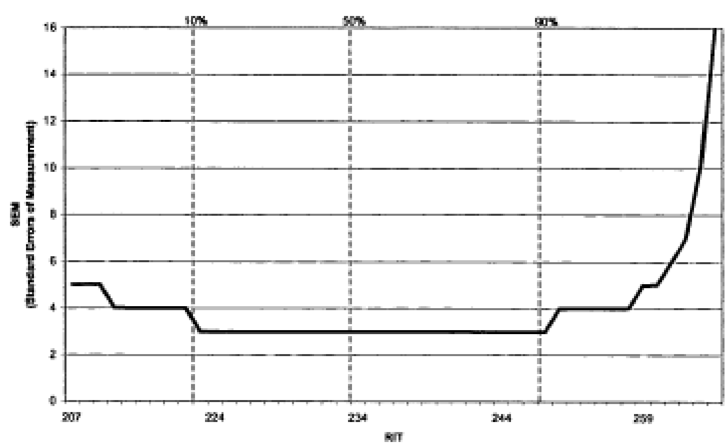
Plain Language



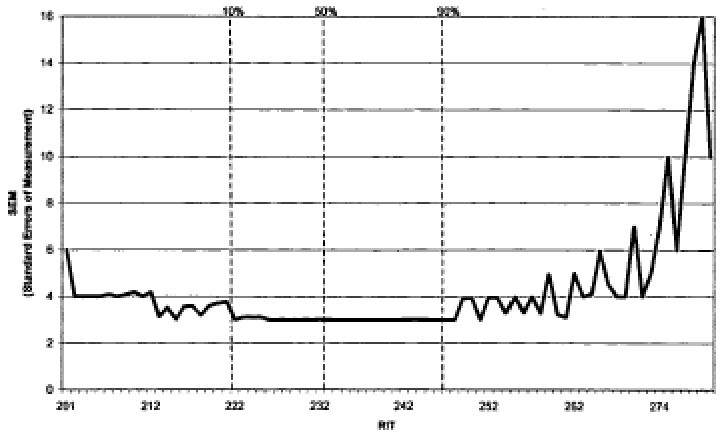
TESA Adaptive



TESA Constrained

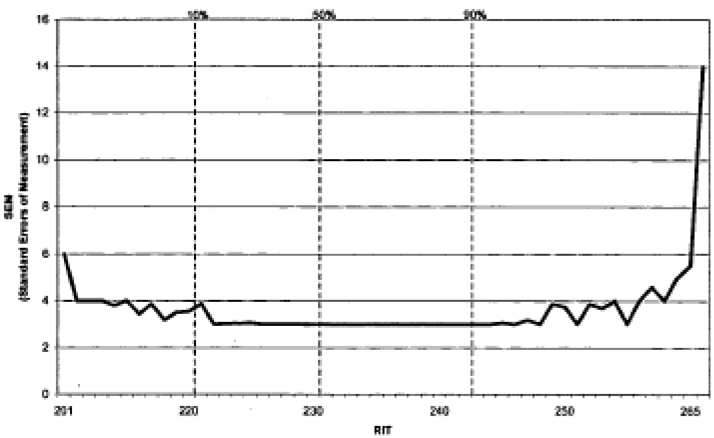


TESA Wide Range

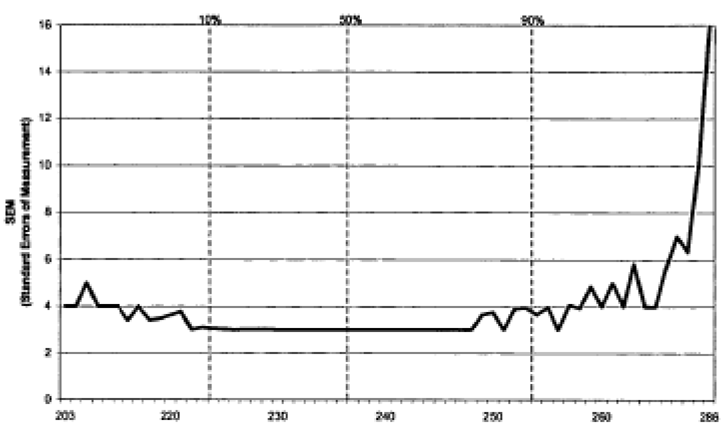


Student Ethnicity

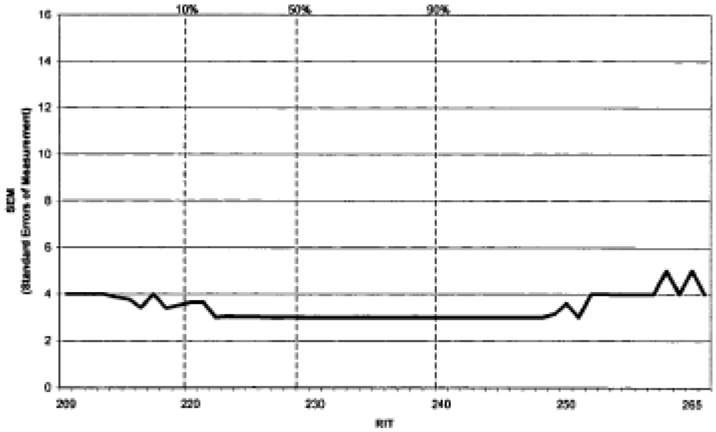
Native American



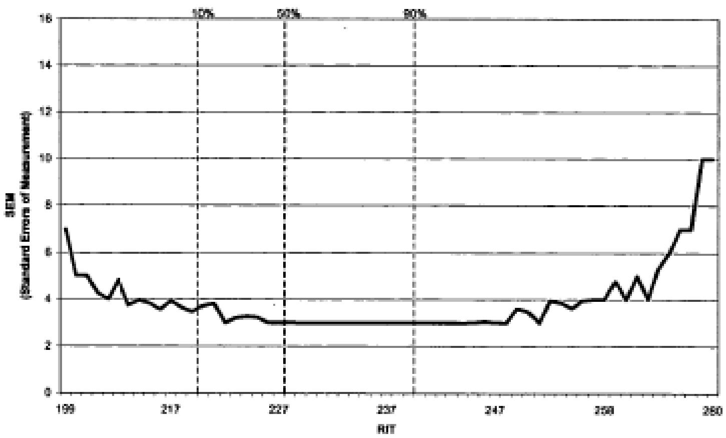
Asian



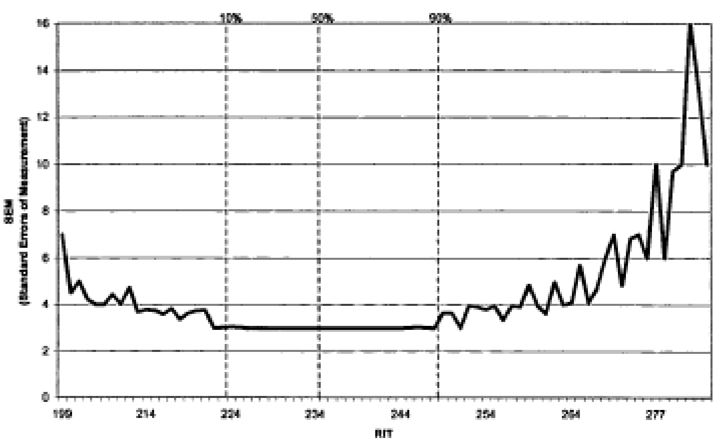
African American



Hispanic



White



APPENDIX B
TESA STRAND RELIABILITY TABLES, READING (GRADES 3 AND 8) AND MATH (GRADES 5 AND 10), 2005–06

Strand Validity Tables

B.1. Grade 3, Reading

Strand	RIT Score	SE	Proficiency Determination
Word Meaning (SRC1)	171	15	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	11	Below
	182	11	Below
	183	11	Below
	184	11	Below
	185	11	Below
	186	10	Below
	187	9	Below
	188	9	Below
	189	9	Below
	190	9	Below
	191	9	Below
	192	9	Near
	193	8	Near
	194	8	Near
	195	8	Near
	196	8	Near
	197	8	Near
	198	8	Near
	199	8	Near
	200	8	Near
	201	8	Near
	202	8	Near
	203	8	Near
	204	8	Near
	205	8	Near
	206	8	Near
	207	8	Near
	208	8	Near
	209	8	Near
	210	8	Above
	211	8	Above
	212	8	Above

Strand	RIT Score	SE	Proficiency Determination
	213	9	Above
	214	9	Above
	215	9	Above
	216	9	Above
	217	9	Above
	218	9	Above
	219	9	Above
	220	10	Above
	221	11	Above
	222	11	Above
	223	11	Above
	224	11	Above
	225	11	Above
	226	11	Above
	227	11	Above
	228	11	Above
	229	11	Above
	230	11	Above
	234	15	Above
Locating Information (SRC2)	167	15	Below
	174	11	Below
	176	11	Below
	178	11	Below
	179	11	Below
	180	12	Below
	181	11	Below
	182	11	Below
	183	11	Below
	184	13	Below
	185	13	Below
	186	11	Below
	187	11	Below
	188	10	Below
	189	11	Below
	190	11	Near
	191	10	Near
	192	9	Near
	193	12	Near
	194	12	Near
	195	12	Near
	196	13	Near
	197	11	Near
	198	11	Near
	199	10	Near
	200	8	Near
	201	12	Near
	202	12	Near

Strand	RIT Score	SE	Proficiency Determination
	203	9	Near
	204	9	Near
	205	10	Near
	206	9	Near
	207	9	Near
	208	10	Near
	209	10	Near
	210	12	Near
	211	12	Near
	212	12	Near
	213	11	Above
	214	11	Above
	215	11	Above
	216	12	Above
	217	11	Above
	218	11	Above
	219	10	Above
	220	10	Above
	221	10	Above
	222	12	Above
	223	12	Above
	224	11	Above
	225	11	Above
	227	12	Above
	228	12	Above
	229	12	Above
	230	11	Above
	231	11	Above
	239	15	Above
Literal Comprehension (SRC3)	156	15	Below
	163	11	Below
	164	11	Below
	165	11	Below
	166	11	Below
	167	11	Below
	168	11	Below
	169	11	Below
	170	11	Below
	171	11	Below
	172	10	Below
	173	9	Below
	174	9	Below
	175	9	Below
	176	9	Below
	177	9	Below
	178	9	Below
	179	8	Below

Strand	RIT Score	SE	Proficiency Determination
	180	8	Below
	181	8	Below
	182	8	Below
	183	8	Below
	184	7	Below
	185	7	Below
	186	7	Below
	187	7	Below
	188	7	Below
	189	7	Below
	190	7	Below
	191	7	Below
	192	7	Below
	193	7	Below
	194	6	Below
	195	6	Near
	196	6	Near
	197	6	Near
	198	6	Near
	199	6	Near
	200	6	Near
	201	6	Near
	202	6	Near
	203	7	Near
	204	7	Near
	205	7	Near
	206	7	Near
	207	7	Near
	208	7	Near
	209	7	Above
	210	7	Above
	211	7	Above
	212	7	Above
	213	7	Above
	214	7	Above
	215	8	Above
	216	8	Above
	217	8	Above
	218	8	Above
	219	8	Above
	220	8	Above
	221	8	Above
	222	9	Above
	223	10	Above
	224	10	Above
	225	11	Above
	226	11	Above
	227	11	Above

Strand	RIT Score	SE	Proficiency Determination
	228	11	Above
	229	11	Above
	230	11	Above
	231	11	Above
	232	11	Above
	238	15	Above
Inferential Comprehension (SRC4)	166	15	Below
	172	11	Below
	173	11	Below
	174	11	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	10	Below
	182	9	Below
	183	9	Below
	184	9	Below
	185	9	Below
	186	8	Below
	187	8	Below
	188	8	Below
	189	8	Below
	190	8	Below
	191	8	Below
	192	8	Below
	193	8	Near
	194	8	Near
	195	7	Near
	196	7	Near
	197	7	Near
	198	7	Near
	199	7	Near
	200	7	Near
	201	7	Near
	202	7	Near
	203	7	Near
	204	7	Near
	205	7	Near
	206	7	Near
	207	7	Near
	208	7	Near
	209	7	Above
	210	8	Above
	211	7	Above

Strand	RIT Score	SE	Proficiency Determination
	212	7	Above
	213	8	Above
	214	8	Above
	215	8	Above
	216	8	Above
	217	8	Above
	218	8	Above
	219	8	Above
	220	8	Above
	221	9	Above
	222	9	Above
	223	9	Above
	224	8	Above
	225	8	Above
	226	9	Above
	227	9	Above
	228	11	Above
	229	11	Above
	230	11	Above
	231	11	Above
	232	11	Above
	233	11	Above
	234	11	Above
	235	11	Above
	241	15	Above
Evaluative Comprehension (SRC 5)	163	15	Below
	169	11	Below
	170	11	Below
	171	11	Below
	172	11	Below
	173	12	Below
	174	11	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	10	Below
	180	11	Below
	181	10	Below
	182	10	Below
	183	10	Below
	184	10	Below
	185	10	Below
	186	8	Below
	187	8	Below
	188	8	Below
	189	9	Below

Strand	RIT Score	SE	Proficiency Determination
	190	9	Below
	191	9	Below
	192	8	Below
	193	8	Near
	194	8	Near
	195	8	Near
	196	8	Near
	197	8	Near
	198	8	Near
	199	8	Near
	200	8	Near
	201	8	Near
	202	8	Near
	203	8	Near
	204	8	Near
	205	8	Near
	206	8	Near
	207	8	Near
	208	8	Near
	209	8	Near
	210	8	Above
	211	8	Above
	212	8	Above
	213	8	Above
	214	8	Above
	215	8	Above
	216	8	Above
	217	8	Above
	218	8	Above
	219	9	Above
	220	9	Above
	221	9	Above
	222	9	Above
	223	9	Above
	224	9	Above
	225	11	Above
	226	11	Above
	227	11	Above
	228	11	Above
	229	11	Above
	230	11	Above
	231	11	Above
	232	11	Above
	233	11	Above
	234	11	Above
	240	15	Above

Strand	RIT Score	SE	Proficiency Determination
Literary Forms (SRC 6)	180	15	Below
	186	11	Below
	187	11	Below
	188	11	Below
	189	11	Below
	190	11	Near
	191	11	Near
	192	12	Near
	193	12	Near
	194	12	Near
	195	12	Near
	196	11	Near
	197	11	Near
	198	10	Near
	199	10	Near
	200	10	Near
	201	9	Near
	202	10	Near
	203	10	Near
	204	10	Near
	205	10	Near
	206	10	Near
	207	9	Near
	208	10	Near
	209	10	Near
	210	11	Near
	211	10	Near
	212	10	Above
	213	11	Above
	214	10	Above
	215	11	Above
	216	10	Above
	217	11	Above
	218	11	Above
	219	11	Above
	220	12	Above
	221	13	Above
	222	13	Above
	223	13	Above
	224	13	Above
	225	12	Above
	226	12	Above
	227	11	Above
	228	11	Above
	229	11	Above
	240	15	Above

Strand	RIT Score	SE	Proficiency Determination
Literary Elements and Devices (SRC 7)	165	15	Below
	171	11	Below
	172	12	Below
	173	12	Below
	174	12	Below
	175	12	Below
	176	12	Below
	177	12	Below
	178	12	Below
	179	12	Below
	180	12	Below
	181	12	Below
	182	12	Below
	183	11	Below
	184	12	Below
	185	11	Below
	186	11	Below
	187	12	Below
	188	11	Below
	189	11	Below
	190	10	Below
	191	10	Near
	192	11	Near
	193	10	Near
	194	11	Near
	195	10	Near
	196	11	Near
	197	10	Near
	198	11	Near
	199	11	Near
	200	11	Near
	201	11	Near
	202	11	Near
	203	11	Near
	204	11	Near
	205	11	Near
	206	12	Near
	207	12	Near
	208	11	Near
	209	11	Near
	210	12	Near
	211	12	Near
	212	12	Near
	213	13	Near
	214	12	Above
	215	13	Above
	216	13	Above

Strand	RIT Score	SE	Proficiency Determination
	217	13	Above
	218	13	Above
	219	12	Above
	220	12	Above
	221	12	Above
	222	13	Above
	223	12	Above
	224	12	Above
	225	12	Above
	226	12	Above
	232	16	Above

B. 2. Grade 5, Math

Strand	RIT Score	SE	Proficiency Determination
Calculations and Estimations (SRC1)	166	15	Below
	175	11	Below
	176	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	8	Below
	182	8	Below
	184	8	Below
	185	8	Below
	186	8	Below
	187	7	Below
	188	7	Below
	189	8	Below
	190	7	Below
	191	6	Below
	192	6	Below
	193	6	Below
	194	6	Below
	195	6	Below
	196	7	Below
	197	6	Below
	198	6	Below
	199	6	Below
	200	6	Below
	201	6	Below
	202	6	Below
	203	6	Below
	204	6	Below
	205	6	Below
	206	6	Below
	207	6	Below
	208	6	Below
	209	6	Near
	210	6	Near
	211	6	Near
	212	6	Near
	213	6	Near
	214	6	Near
	215	6	Near
	216	6	Near
	217	6	Near
	218	6	Near
	219	6	Near
	220	6	Near

	221	6	Near
	222	6	Above
	223	6	Above
	224	6	Above
	225	6	Above
	226	6	Above
	227	6	Above
	228	6	Above
	229	6	Above
	230	7	Above
	231	6	Above
	232	6	Above
	233	6	Above
	234	8	Above
	235	7	Above
	236	7	Above
	237	7	Above
	238	8	Above
	239	8	Above
	240	8	Above
	241	8	Above
	242	8	Above
	243	11	Above
	245	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	250	11	Above
	251	11	Above
	258	15	Above
Measurement (SRC2)	164	15	Below
	174	11	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	181	8	Below
	182	9	Below
	183	9	Below
	184	9	Below
	185	9	Below
	186	9	Below
	187	7	Below
	188	9	Below
	189	8	Below
	190	8	Below
	191	8	Below

192	8	Below
193	8	Below
194	8	Below
195	8	Below
196	7	Below
197	8	Below
198	7	Below
199	7	Below
200	7	Below
201	8	Below
202	7	Below
203	7	Below
204	7	Below
205	7	Below
206	7	Below
207	7	Below
208	7	Near
209	7	Near
210	7	Near
211	7	Near
212	7	Near
213	7	Near
214	7	Near
215	7	Near
216	7	Near
217	7	Near
218	7	Near
219	7	Near
220	7	Near
221	7	Near
222	7	Near
223	7	Above
224	7	Above
225	7	Above
226	7	Above
227	7	Above
228	7	Above
229	7	Above
230	7	Above
231	8	Above
232	8	Above
233	7	Above
234	7	Above
235	7	Above
236	9	Above
237	9	Above
238	9	Above
239	8	Above
240	8	Above

	241	8	Above
	242	11	Above
	243	11	Above
	244	11	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	258	15	Above
Statistics and Probability (SRC3)	165	15	Below
	174	11	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	11	Below
	182	11	Below
	183	9	Below
	184	9	Below
	185	9	Below
	186	10	Below
	187	9	Below
	188	10	Below
	189	9	Below
	190	9	Below
	191	9	Below
	192	9	Below
	193	9	Below
	194	9	Below
	195	9	Below
	196	8	Below
	197	8	Below
	198	9	Below
	199	8	Below
	200	9	Below
	201	8	Below
	202	8	Below
	203	8	Below
	204	8	Below
	205	8	Below
	206	8	Below
	207	8	Near
	208	8	Near
	209	8	Near
	210	8	Near

	211	8	Near
	212	8	Near
	213	8	Near
	214	8	Near
	215	8	Near
	216	8	Near
	217	8	Near
	218	7	Near
	219	7	Near
	220	8	Near
	221	7	Near
	222	8	Near
	223	7	Near
	224	8	Above
	225	8	Above
	226	8	Above
	227	8	Above
	228	7	Above
	229	8	Above
	230	8	Above
	231	8	Above
	232	8	Above
	233	7	Above
	234	9	Above
	235	9	Above
	236	9	Above
	237	9	Above
	238	8	Above
	239	8	Above
	240	8	Above
	241	11	Above
	242	11	Above
	243	11	Above
	244	11	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	258	15	Above
Algebraic relationships (SRC4)	164	15	Below
	174	11	Below
	175	11	Below
	176	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	9	Below

182	9	Below
183	9	Below
184	10	Below
185	9	Below
186	9	Below
187	8	Below
188	8	Below
189	8	Below
190	8	Below
191	8	Below
192	8	Below
193	8	Below
194	8	Below
195	8	Below
196	8	Below
197	8	Below
198	8	Below
199	7	Below
200	8	Below
201	7	Below
202	7	Below
203	7	Below
204	7	Below
205	7	Below
206	7	Below
207	7	Below
208	7	Near
209	7	Near
210	7	Near
211	7	Near
212	7	Near
213	7	Near
214	7	Near
215	7	Near
216	7	Near
217	7	Near
218	7	Near
219	7	Near
220	7	Near
221	7	Near
222	7	Near
223	7	Above
224	7	Above
225	7	Above
226	7	Above
227	7	Above
228	7	Above
229	7	Above
230	7	Above

	231	7	Above
	232	7	Above
	233	7	Above
	234	7	Above
	235	7	Above
	236	8	Above
	237	9	Above
	238	8	Above
	239	9	Above
	240	8	Above
	241	8	Above
	242	11	Above
	243	11	Above
	244	11	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	258	15	Above
[redacted]			
Geometry (SRC5)	164	15	Below
	174	11	Below
	175	11	Below
	177	11	Below
	178	11	Below
	179	11	Below
	180	11	Below
	181	8	Below
	182	9	Below
	183	8	Below
	184	10	Below
	185	9	Below
	186	8	Below
	187	8	Below
	188	8	Below
	189	8	Below
	190	8	Below
	191	7	Below
	192	8	Below
	193	8	Below
	194	8	Below
	195	8	Below
	196	8	Below
	197	8	Below
	198	8	Below
	199	7	Below
	200	7	Below
	201	7	Below

202	7	Below
203	7	Below
204	7	Below
205	7	Below
206	7	Below
207	7	Below
208	7	Near
209	7	Near
210	7	Near
211	7	Near
212	7	Near
213	7	Near
214	7	Near
215	7	Near
216	7	Near
217	7	Near
218	7	Near
219	7	Near
220	7	Near
221	7	Near
222	7	Near
223	7	Above
224	7	Above
225	7	Above
226	7	Above
227	7	Above
228	7	Above
229	7	Above
230	7	Above
231	8	Above
232	8	Above
233	7	Above
234	8	Above
235	7	Above
236	8	Above
237	8	Above
238	9	Above
239	8	Above
240	8	Above
241	11	Above
242	11	Above
243	11	Above
244	11	Above
245	11	Above
246	11	Above
247	11	Above
248	11	Above
249	11	Above
257	15	Above

B.3. Grade 8, Reading

Strand	RIT Score	SE	Proficiency Determination
Word Meaning (SRC1)	186	15	Below
	190	11	Below
	191	11	Below
	192	11	Below
	193	11	Below
	194	11	Below
	195	11	Below
	196	11	Below
	197	11	Below
	198	9	Below
	199	9	Below
	200	9	Below
	201	8	Below
	202	8	Below
	203	8	Below
	204	8	Below
	205	8	Below
	206	8	Below
	207	7	Below
	208	7	Below
	209	7	Below
	210	7	Below
	211	7	Below
	212	7	Below
	213	7	Below
	214	7	Below
	215	7	Below
	216	7	Below
	217	7	Below
	218	7	Below
	219	7	Below
	220	7	Below
	221	7	Below
	222	7	Below
	223	7	Below
	224	7	Near
	225	7	Near
	226	7	Near
	227	7	Near
	228	7	Near
	229	7	Near
	230	7	Near
	231	7	Near
	232	7	Near
	233	7	Near
	234	7	Near

	235	7	Near
	236	7	Near
	237	7	Near
	238	8	Near
	239	8	Near
	240	8	Above
	241	8	Above
	242	8	Above
	243	8	Above
	244	8	Above
	245	8	Above
	246	8	Above
	247	9	Above
	248	10	Above
	249	11	Above
	250	11	Above
	251	11	Above
	252	11	Above
	253	11	Above
	254	11	Above
	255	11	Above
	256	11	Above
	257	11	Above
	261	15	Above
Locating Information (SRC2)	193	15	Below
	198	11	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	12	Below
	203	11	Below
	204	11	Below
	205	12	Below
	206	11	Below
	207	11	Below
	208	11	Below
	209	10	Below
	210	10	Below
	211	10	Below
	212	11	Below
	213	9	Below
	214	10	Below
	215	8	Below
	216	9	Below
	217	8	Below
	218	9	Below
	219	9	Below
	220	8	Below

	221	8	Below
	222	9	Below
	223	7	Below
	224	9	Near
	225	8	Near
	226	8	Near
	227	8	Near
	228	8	Near
	229	8	Near
	230	7	Near
	231	8	Near
	232	9	Near
	233	8	Near
	234	8	Near
	235	8	Near
	236	9	Near
	237	9	Near
	238	9	Near
	239	9	Near
	240	9	Near
	241	10	Near
	242	11	Near
	243	11	Above
	244	11	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	250	11	Above
	251	11	Above
	256	15	Above
Literal Comprehension (SRC3)	186	15	Below
	193	11	Below
	194	11	Below
	195	11	Below
	196	11	Below
	197	11	Below
	198	11	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	10	Below
	203	9	Below
	204	8	Below
	205	8	Below
	206	8	Below
	207	9	Below

208	8	Below
209	8	Below
210	8	Below
211	8	Below
212	8	Below
213	8	Below
214	7	Below
215	7	Below
216	7	Below
217	7	Below
218	7	Below
219	7	Below
220	7	Below
221	7	Below
222	7	Below
223	7	Below
224	7	Near
225	7	Near
226	7	Near
227	7	Near
228	7	Near
229	7	Near
230	7	Near
231	7	Near
232	7	Near
233	7	Near
234	7	Near
235	7	Near
236	7	Near
237	7	Near
238	8	Near
239	8	Above
240	8	Above
241	8	Above
242	8	Above
243	8	Above
244	9	Above
245	10	Above
246	11	Above
247	11	Above
248	11	Above
249	11	Above
250	11	Above
251	11	Above
252	11	Above
253	11	Above
254	11	Above
261	15	Above

Inferential Comprehension (SRC4)	185	15	Below
	192	11	Below
	193	11	Below
	194	11	Below
	195	11	Below
	196	11	Below
	197	11	Below
	198	11	Below
	199	11	Below
	200	11	Below
	201	10	Below
	202	9	Below
	203	9	Below
	204	9	Below
	205	9	Below
	206	8	Below
	207	8	Below
	208	8	Below
	209	8	Below
	210	8	Below
	211	8	Below
	212	8	Below
	213	8	Below
	214	8	Below
	215	7	Below
	216	7	Below
	217	7	Below
	218	7	Below
	219	7	Below
	220	7	Below
	221	7	Below
	222	7	Below
	223	7	Below
	224	7	Near
	225	7	Near
	226	7	Near
	227	7	Near
	228	7	Near
	229	7	Near
	230	7	Near
	231	7	Near
	232	7	Near
	233	7	Near
	234	7	Near
	235	7	Near
	236	8	Near
	237	8	Near
	238	8	Near
	239	8	Near

	240	8	Above
	241	8	Above
	242	8	Above
	243	8	Above
	244	9	Above
	245	8	Above
	246	8	Above
	247	8	Above
	248	8	Above
	249	10	Above
	250	11	Above
	251	11	Above
	252	11	Above
	253	11	Above
	254	11	Above
	255	11	Above
	256	11	Above
	257	11	Above
	258	11	Above
	263	15	Above
Evaluative Comprehension (SRC 5)	190	15	Below
	195	11	Below
	196	11	Below
	197	11	Below
	198	11	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	11	Below
	203	11	Below
	204	10	Below
	205	9	Below
	206	9	Below
	207	9	Below
	208	9	Below
	209	9	Below
	210	9	Below
	211	8	Below
	212	8	Below
	213	8	Below
	214	8	Below
	215	8	Below
	216	8	Below
	217	8	Below
	218	8	Below
	219	8	Below
	220	8	Below
	221	8	Below

	222	8	Below
	223	8	Near
	224	8	Near
	225	8	Near
	226	8	Near
	227	8	Near
	228	8	Near
	229	8	Near
	230	8	Near
	231	8	Near
	232	8	Near
	233	8	Near
	234	8	Near
	235	8	Near
	236	8	Near
	237	8	Near
	238	9	Near
	239	8	Near
	240	8	Above
	241	8	Above
	242	8	Above
	243	9	Above
	244	10	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	250	11	Above
	251	11	Above
	252	11	Above
	253	11	Above
	254	11	Above
	256	15	Above
Literary Forms (SRC 6)	195	15	Below
	200	11	Below
	201	11	Below
	202	11	Below
	203	11	Below
	204	11	Below
	205	11	Below
	206	11	Below
	207	11	Below
	208	11	Below
	209	12	Below
	210	11	Below
	211	11	Below
	212	10	Below

	213	10	Below
	214	10	Below
	215	9	Below
	216	9	Below
	217	9	Below
	218	9	Below
	219	9	Below
	220	9	Below
	221	9	Below
	222	9	Near
	223	9	Near
	224	9	Near
	225	9	Near
	226	9	Near
	227	9	Near
	228	9	Near
	229	9	Near
	230	9	Near
	231	9	Near
	232	9	Near
	233	9	Near
	234	9	Near
	235	9	Near
	236	9	Near
	237	10	Near
	238	10	Near
	239	11	Near
	240	11	Near
	241	11	Near
	242	11	Near
	243	11	Above
	244	11	Above
	245	11	Above
	246	11	Above
	247	11	Above
	248	11	Above
	249	11	Above
	250	11	Above
	256	15	Above
Literary Elements and Devices (SRC 7)	193	15	Below
	198	11	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	11	Below
	203	11	Below
	204	11	Below
	205	11	Below

206	11	Below
207	11	Below
208	9	Below
209	9	Below
210	9	Below
211	9	Below
212	9	Below
213	8	Below
214	8	Below
215	8	Below
216	8	Below
217	8	Below
218	8	Below
219	8	Below
220	7	Below
221	7	Below
222	7	Below
223	7	Below
224	7	Near
225	7	Near
226	7	Near
227	7	Near
228	7	Near
229	7	Near
230	7	Near
231	7	Near
232	7	Near
233	7	Near
234	7	Near
235	7	Near
236	7	Near
237	7	Near
238	7	Near
239	7	Above
240	7	Above
241	7	Above
242	7	Above
243	7	Above
244	7	Above
245	7	Above
246	8	Above
247	8	Above
248	8	Above
249	8	Above
250	8	Above
251	9	Above
252	9	Above
253	11	Above
254	11	Above

255	11	Above
256	11	Above
257	11	Above
258	11	Above
259	10	Above
260	10	Above
266	14	Above

B.4. Grade 10, Math

Strand	RIT Score	SE	Proficiency Determination
Calculations and Estimations (SRC1)	198	15	Below
	207	11	Below
	208	12	Below
	209	12	Below
	210	12	Below
	211	11	Below
	212	12	Below
	213	12	Below
	214	12	Below
	215	12	Below
	216	12	Below
	217	10	Below
	218	10	Below
	219	11	Below
	220	10	Below
	221	11	Below
	222	11	Below
	223	11	Below
	224	11	Below
	225	11	Below
	226	10	Below
	227	10	Below
	228	11	Near
	229	11	Near
	230	10	Near
	231	10	Near
	232	10	Near
	233	10	Near
	234	11	Near
	235	11	Near
	236	10	Near
	237	10	Near
	238	10	Near
	239	10	Near
	240	11	Near
	241	11	Near
	242	11	Near
	243	12	Near
	244	11	Near
	245	11	Near
	246	11	Near
	247	11	Near
	248	11	Near
	249	11	Near
	250	12	Near

	251	10	Above
	252	11	Above
	253	11	Above
	254	11	Above
	255	11	Above
	256	12	Above
	257	10	Above
	258	11	Above
	259	13	Above
	260	13	Above
	261	12	Above
	262	12	Above
	263	12	Above
	264	12	Above
	265	12	Above
	266	12	Above
	267	11	Above
	268	11	Above
	277	15	Above
Measurement (SRC2)	201	15	Below
	211	11	Below
	212	12	Below
	213	12	Below
	214	11	Below
	215	11	Below
	216	12	Below
	217	12	Below
	218	12	Below
	219	12	Below
	220	12	Below
	221	10	Below
	222	11	Below
	223	11	Below
	224	10	Below
	225	11	Below
	226	11	Below
	227	10	Below
	228	11	Near
	229	10	Near
	230	10	Near
	231	10	Near
	232	10	Near
	233	10	Near
	234	10	Near
	235	10	Near
	236	11	Near
	237	10	Near
	238	10	Near

	239	11	Near
	240	10	Near
	241	10	Near
	242	10	Near
	243	10	Near
	244	10	Near
	245	11	Near
	246	10	Near
	247	11	Near
	248	10	Near
	249	10	Near
	250	11	Near
	251	11	Above
	252	10	Above
	253	11	Above
	254	11	Above
	255	10	Above
	256	11	Above
	257	10	Above
	258	12	Above
	259	12	Above
	260	12	Above
	261	12	Above
	262	12	Above
	263	12	Above
	264	12	Above
	265	11	Above
	266	11	Above
	267	11	Above
	276	15	Above
Statistics and Probability (SRC3)	191	15	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	11	Below
	203	11	Below
	204	11	Below
	205	11	Below
	206	11	Below
	207	11	Below
	208	8	Below
	209	9	Below
	210	9	Below
	211	9	Below
	212	10	Below
	213	8	Below
	214	7	Below
	215	8	Below

216	8	Below
217	8	Below
218	7	Below
219	7	Below
220	8	Below
221	8	Below
222	8	Below
223	7	Below
224	7	Below
225	7	Below
226	7	Below
227	7	Below
228	7	Below
229	7	Below
230	7	Below
231	7	Below
232	7	Near
233	7	Near
234	7	Near
235	7	Near
236	7	Near
237	7	Near
238	7	Near
239	7	Near
240	7	Near
241	7	Near
242	7	Near
243	7	Near
244	7	Near
245	7	Near
246	7	Near
247	7	Above
248	7	Above
249	7	Above
250	7	Above
251	8	Above
252	7	Above
253	8	Above
254	8	Above
255	8	Above
256	7	Above
257	9	Above
258	9	Above
259	9	Above
260	7	Above
261	10	Above
262	9	Above
263	10	Above
264	9	Above

	265	8	Above
	266	8	Above
	268	11	Above
	269	11	Above
	272	11	Above
	273	11	Above
	274	11	Above
	283	15	Above
Algebraic relationships (SRC4)	184	14	Below
	200	8	Below
	201	8	Below
	202	8	Below
	203	8	Below
	204	9	Below
	205	7	Below
	206	7	Below
	207	6	Below
	208	6	Below
	209	6	Below
	210	6	Below
	211	6	Below
	212	5	Below
	213	6	Below
	214	5	Below
	215	5	Below
	216	5	Below
	217	5	Below
	218	5	Below
	219	5	Below
	220	5	Below
	221	5	Below
	222	5	Below
	223	5	Below
	224	5	Below
	225	5	Below
	226	5	Below
	227	5	Below
	228	5	Below
	229	5	Below
	230	5	Below
	231	5	Below
	232	5	Below
	233	5	Below
	234	5	Near
	235	5	Near
	236	5	Near
	237	5	Near
	238	5	Near

	239	5	Near
	240	5	Near
	241	5	Near
	242	5	Near
	243	5	Near
	244	5	Near
	245	5	Above
	246	5	Above
	247	5	Above
	248	5	Above
	249	5	Above
	250	5	Above
	251	5	Above
	252	5	Above
	253	5	Above
	254	5	Above
	255	5	Above
	256	5	Above
	257	5	Above
	258	5	Above
	259	5	Above
	260	5	Above
	261	5	Above
	262	5	Above
	263	6	Above
	264	6	Above
	265	6	Above
	266	6	Above
	267	6	Above
	268	7	Above
	269	7	Above
	271	8	Above
	272	8	Above
	273	8	Above
	274	8	Above
	278	10	Above
	279	10	Above
	281	10	Above
	282	10	Above
	289	14	Above
Geometry (SRC5)	189	15	Below
	199	11	Below
	200	11	Below
	201	11	Below
	202	11	Below
	203	11	Below
	204	11	Below
	205	8	Below

206	8	Below
207	8	Below
208	8	Below
209	8	Below
210	9	Below
211	7	Below
212	7	Below
213	7	Below
214	7	Below
215	6	Below
216	6	Below
217	6	Below
218	6	Below
219	6	Below
220	6	Below
221	6	Below
222	6	Below
223	6	Below
224	6	Below
225	6	Below
226	6	Below
227	6	Below
228	6	Below
229	6	Below
230	6	Below
231	6	Below
232	6	Below
233	6	Near
234	6	Near
235	6	Near
236	6	Near
237	6	Near
238	6	Near
239	6	Near
240	6	Near
241	6	Near
242	6	Near
243	6	Near
244	6	Near
245	6	Near
246	6	Above
247	6	Above
248	6	Above
249	6	Above
250	6	Above
251	6	Above
252	7	Above
253	6	Above
254	6	Above

255	7	Above
256	6	Above
257	6	Above
258	6	Above
259	7	Above
260	7	Above
261	7	Above
262	7	Above
263	8	Above
264	8	Above
265	8	Above
266	8	Above
267	8	Above
269	11	Above
271	10	Above
272	10	Above
273	10	Above
274	10	Above
275	10	Above
283	14	Above

APPENDIX C

CLASSIFICATION ANALYSIS, TESA (LONG AND SHORT) AND PAPER AND PENCIL FOR READING (GRADES 3 AND 8) AND MATH (GRADES 5 AND 10), 2005-06

C.1. Grade 3, Reading: TESA

RIT	Probability of False Negative	Probability of False Positive
171	6.88E-11	n/a
172	1.75E-10	n/a
173	4.15E-11	n/a
174	5.27E-11	n/a
175	3.65E-09	n/a
176	2.24E-09	n/a
177	4.47E-08	n/a
178	8.03E-08	n/a
179	9.63E-08	n/a
180	0.000000661	n/a
181	0.00000193	n/a
182	0.00000634	n/a
183	0.0000248	n/a
184	0.0000719	n/a
185	0.000294	n/a
186	0.000824	n/a
187	0.00254	n/a
188	0.0075	n/a
189	0.0204837	n/a
190	0.0540572	n/a
191	0.1459198	n/a
192	0.3561398	n/a
193	0.8245528	n/a
194	1.7732786	n/a
195	3.6825965	n/a
196	6.8725674	n/a
197	11.7689405	n/a
198	19.1047705	n/a
199	28.9403867	n/a
200	40.9763442	n/a
201	n/a	45.7300946
202	n/a	33.8513459
203	n/a	23.3550722
204	n/a	14.5514303
205	n/a	8.7190518
206	n/a	4.7131365
207	n/a	2.5483493
208	n/a	1.2574837
209	n/a	0.5488733
210	n/a	0.229087

RIT	Probability of False Negative	Probability of False Positive
211	n/a	0.0889186
212	n/a	0.0315279
213	n/a	0.010629
214	n/a	0.00333
215	n/a	0.00108
216	n/a	0.000298
217	n/a	0.0000775
218	n/a	0.0000166
219	n/a	0.00000534
220	n/a	0.00000136
221	n/a	0.00000035
222	n/a	7.63E-08
223	n/a	2.27E-08
224	n/a	8.86E-09
225	n/a	2.46E-09
226	n/a	4.88E-10
227	n/a	1.15E-10
228	n/a	4.71E-11
229	n/a	8.68E-12
230	n/a	2.83E-12
231	n/a	6.8E-13
232	n/a	1.18E-13
233	n/a	1.24E-13
234	n/a	1.69E-14
235	n/a	8.44E-14
236	n/a	1.29E-14
237	n/a	1.19E-15
238	n/a	2.07E-16
239	n/a	3.35E-15
240	n/a	1.32E-15
241	n/a	1.52E-16
242	n/a	2.28E-17
243	n/a	2.37E-19
248	n/a	2.3E-18
249	n/a	5.74E-19

C.2. Grade 3, Reading: TESA (Short)

RIT	Probability of False Negative	Probability of False Positive
171	0.0000047	n/a
172	0.0000163	n/a
173	0.0000112	n/a
174	5.15E-09	n/a
175	0.0000016	n/a
176	0.0000635	n/a
177	0.0000601	n/a
178	0.000688	n/a
179	0.00048	n/a
180	0.000582	n/a
181	0.00176	n/a
182	0.00297	n/a
183	0.0103441	n/a
184	0.0197707	n/a
185	0.040643	n/a
186	0.0882357	n/a
187	0.2792949	n/a
188	0.3946316	n/a
189	0.6556946	n/a
190	1.3241364	n/a
191	2.0720717	n/a
192	3.590123	n/a
193	6.2712214	n/a
194	8.5696191	n/a
195	14.4088018	n/a
196	18.7248326	n/a
197	26.3680317	n/a
198	35.0670489	n/a
199	43.0388273	n/a
200	52.2960983	n/a
201	n/a	37.994701
202	n/a	30.7139424
203	n/a	22.3831122
204	n/a	17.3359031
205	n/a	13.118032
206	n/a	8.6518015
207	n/a	5.4795635
208	n/a	3.4943039
209	n/a	1.9179564
210	n/a	1.3126585
211	n/a	0.7415497
212	n/a	0.3951977
213	n/a	0.2286838
214	n/a	0.1323331
215	n/a	0.0461306
216	n/a	0.0241647

RIT	Probability of False Negative	Probability of False Positive
217	n/a	0.00966
218	n/a	0.00321
219	n/a	0.00172
220	n/a	0.000665
221	n/a	0.000294
222	n/a	0.0000782
223	n/a	0.0000259
224	n/a	0.0000132
225	n/a	0.00000425
226	n/a	0.00000118
227	n/a	0.000000563
228	n/a	0.000000271
229	n/a	5.43E-08
230	n/a	0.00000002
231	n/a	5.87E-09
232	n/a	7.22E-10
233	n/a	8.94E-10
234	n/a	1.05E-09
235	n/a	3.61E-11
236	n/a	1.63E-10
237	n/a	9.46E-11
238	n/a	5.38E-11
239	n/a	3.16E-10
240	n/a	3.23E-11
241	n/a	6.09E-12
242	n/a	3.42E-12
243	n/a	1.12E-14
248	n/a	1.11E-13
249	n/a	4.49E-14

C.3. Grade 3, Reading: Paper and Pencil

RIT	Probability of False Negative	Probability of False Positive
155	1.4E-09	n/a
162	1.77E-08	n/a
169	0.000000179	n/a
174	0.00000148	n/a
177	0.0000103	n/a
179	0.0000612	n/a
182	0.000314	n/a
183	0.00141	n/a
185	0.00558	n/a
187	0.0196508	n/a
188	0.061986	n/a
189	0.1760357	n/a
191	0.4528887	n/a
192	1.0613435	n/a
193	2.2740559	n/a
194	4.4662911	n/a
195	8.0686419	n/a
196	13.4799434	n/a
197	20.951493	n/a
198	30.4368061	n/a
199	41.4505223	n/a
200	53.0895921	n/a
201	n/a	25.4372584
202	n/a	17.0497086
203	n/a	10.7057189
204	n/a	6.2659628
205	n/a	3.4024601
206	n/a	1.7099242
207	n/a	0.79501
208	n/a	0.3417833
209	n/a	0.1355254
210	n/a	0.0493598
211	n/a	0.0164484
212	n/a	0.005
213	n/a	0.00139
214	n/a	0.000349
215	n/a	0.0000796
216	n/a	0.0000163
217	n/a	0.000003
218	n/a	0.00000049
220	n/a	7.09E-08
221	n/a	0.000000009
223	n/a	9.91E-10
225	n/a	9.37E-11
227	n/a	7.5E-12
230	n/a	4.98E-13

RIT	Probability of False Negative	Probability of False Positive
233	n/a	2.68E-14
237	n/a	1.13E-15
244	n/a	3.62E-17
251	n/a	8.2E-19

C.4. Grade 5, Math: TESA

RIT	Probability of False Negative	Probability of False Positive
178	7.57E-23	n/a
179	1.17E-23	n/a
180	2.59E-22	n/a
181	1.04E-19	n/a
182	3.58E-19	n/a
183	1.26E-18	n/a
184	1.06E-17	n/a
185	4.42E-15	n/a
186	1.28E-15	n/a
187	2.55E-14	n/a
188	3.38E-13	n/a
189	4.39E-12	n/a
190	1.47E-11	n/a
191	2.74E-09	n/a
192	1.15E-09	n/a
193	1.08E-08	n/a
194	8.75E-08	n/a
195	0.000000366	n/a
196	0.00000246	n/a
197	0.0000111	n/a
198	0.0000496	n/a
199	0.000136	n/a
200	0.000502	n/a
201	0.00162	n/a
202	0.00564	n/a
203	0.0187641	n/a
204	0.0563173	n/a
205	0.1663884	n/a
206	0.4209329	n/a
207	0.9674363	n/a
208	2.1207679	n/a
209	4.2799679	n/a
210	7.7221983	n/a
211	13.3066397	n/a
212	20.9969741	n/a
213	30.9704541	n/a
214	42.5102569	n/a
215	n/a	45.0764408
216	n/a	32.5592679
217	n/a	22.1673517
218	n/a	13.9484534
219	n/a	8.0139787
220	n/a	4.2118615
221	n/a	1.9983636
222	n/a	0.8870769
223	n/a	0.345937

RIT	Probability of False Negative	Probability of False Positive
224	n/a	0.1163082
225	n/a	0.0365988
226	n/a	0.0103126
227	n/a	0.00326
228	n/a	0.000871
229	n/a	0.000152
230	n/a	0.0000312
231	n/a	0.00000787
232	n/a	0.0000012
233	n/a	0.000000193
234	n/a	0.000000028
235	n/a	4.57E-09
236	n/a	5.61E-10
237	n/a	1.12E-10
238	n/a	1.17E-11
239	n/a	1.67E-12
240	n/a	2.42E-13
241	n/a	4.75E-14
242	n/a	1.43E-14
243	n/a	1.85E-15
244	n/a	1.73E-16
245	n/a	5.79E-17
246	n/a	1.4E-17
247	n/a	2.98E-18
248	n/a	6.06E-19
249	n/a	1.4E-19
251	n/a	7.84E-21
254	n/a	1.79E-21
255	n/a	5.57E-23
258	n/a	1.86E-23
259	n/a	8.15E-24
265	n/a	1.2E-23
266	n/a	1.63E-24
274	n/a	6.34E-27

C.5. Grade 5, Math: TESA (Short)

RIT	Probability of False Negative	Probability of False Positive
178	1.73E-14	n/a
179	5.21E-14	n/a
180	2.67E-15	n/a
181	2.13E-12	n/a
182	9.03E-15	n/a
183	2.19E-11	n/a
184	3.52E-11	n/a
185	1.04E-07	n/a
186	1.09E-08	n/a
187	6.20E-08	n/a
188	3.42E-05	n/a
189	5.81E-06	n/a
190	6.25E-06	n/a
191	1.73E-03	n/a
192	3.74E-04	n/a
193	7.23E-05	n/a
194	3.77E-04	n/a
195	2.12E-03	n/a
196	1.87E-02	n/a
197	1.49E-02	n/a
198	3.29E-02	n/a
199	4.50E-02	n/a
200	9.29E-02	n/a
201	3.33E-01	n/a
202	4.81E-01	n/a
203	9.37E-01	n/a
204	1.22E+00	n/a
205	2.68E+00	n/a
206	4.52E+00	n/a
207	5.99E+00	n/a
208	9.70E+00	n/a
209	1.34E+01	n/a
210	1.92E+01	n/a
211	2.49E+01	n/a
212	3.18E+01	n/a
213	4.18E+01	n/a
214	4.95E+01	n/a
215	n/a	41.1349655
216	n/a	31.0088993
217	n/a	23.1623315
218	n/a	18.0266946
219	n/a	12.1910914
220	n/a	7.9883432
221	n/a	5.6345365
222	n/a	3.1148718
223	n/a	1.7833686

RIT	Probability of False Negative	Probability of False Positive
224	n/a	1.0734398
225	n/a	0.5490683
226	n/a	0.2628917
227	n/a	0.1203122
228	n/a	0.0756368
229	n/a	0.0254214
230	n/a	0.00991
231	n/a	0.00515
232	n/a	0.00135
233	n/a	0.000207
234	n/a	0.000155
235	n/a	0.0000897
236	n/a	0.00000696
237	n/a	0.00000429
238	n/a	0.000000278
239	n/a	3.38E-08
240	n/a	7.93E-09
241	n/a	7.02E-09
242	n/a	3.4E-09
243	n/a	1.1E-10
244	n/a	5.01E-11
245	n/a	1.11E-11
246	n/a	1.45E-12
247	n/a	1.39E-12
248	n/a	9.82E-13
249	n/a	7.16E-14
251	n/a	3.99E-15
254	n/a	6.95E-16
255	n/a	6.21E-16
258	n/a	5.2E-17
259	n/a	2.4E-17
265	n/a	2.17E-18
266	n/a	4.45E-19
274	n/a	5.53E-20

C.6. Grade 5, Math: Paper and Pencil

RIT	Probability of False Negative	Probability of False Positive
164	2.36E-12	n/a
171	3.66E-11	n/a
179	4.65E-10	n/a
183	4.96E-09	n/a
187	0.000000045	n/a
189	0.000000351	n/a
192	0.00000238	n/a
193	0.0000142	n/a
195	0.0000744	n/a
197	0.000347	n/a
198	0.00145	n/a
200	0.00543	n/a
201	0.0183391	n/a
202	0.0561628	n/a
204	0.1564159	n/a
205	0.3975059	n/a
206	0.9247114	n/a
207	1.975549	n/a
208	3.889333	n/a
209	7.0795172	n/a
210	11.9513846	n/a
211	18.7732126	n/a
212	27.5464159	n/a
213	37.9306489	n/a
214	49.2541651	n/a
215	n/a	39.402787
216	n/a	28.9066688
216	n/a	19.9552365
217	n/a	12.9252213
218	n/a	7.835346
219	n/a	4.4339783
220	n/a	2.3362258
221	n/a	1.143615
222	n/a	0.5192724
223	n/a	0.2183916
224	n/a	0.0849228
225	n/a	0.030466
226	n/a	0.0100615
227	n/a	0.00305
228	n/a	0.000849
229	n/a	0.000216
230	n/a	0.0000501
231	n/a	0.0000106
232	n/a	0.00000202
233	n/a	0.000000348
234	n/a	5.38E-08

RIT	Probability of False Negative	Probability of False Positive
236	n/a	7.42E-09
237	n/a	9.09E-10
239	n/a	9.8E-11
241	n/a	9.22E-12
243	n/a	7.49E-13
246	n/a	5.19E-14
249	n/a	3.01E-15
253	n/a	1.44E-16
260	n/a	5.52E-18
267	n/a	1.64E-19

C.7. Grade 8, Reading: TESA

RIT	Probability of False Negative	Probability of False Positive
196	1.48E-22	n/a
197	2.82E-20	n/a
198	6.81E-20	n/a
199	5.4E-21	n/a
200	1.63E-19	n/a
201	3.22E-19	n/a
202	6.13E-18	n/a
203	2.19E-16	n/a
204	7.17E-16	n/a
205	4.64E-15	n/a
206	3.32E-14	n/a
207	2.04E-13	n/a
208	2.29E-12	n/a
209	1.5E-11	n/a
210	2.04E-10	n/a
211	1.21E-09	n/a
212	9.82E-09	n/a
213	6.03E-08	n/a
214	0.000000419	n/a
215	0.0000027	n/a
216	0.0000132	n/a
217	0.0000799	n/a
218	0.0004	n/a
219	0.00184	n/a
220	0.00744	n/a
221	0.0270591	n/a
222	0.0909579	n/a
223	0.2582523	n/a
224	0.6883766	n/a
225	1.6887708	n/a
226	3.6207961	n/a
227	7.1611942	n/a
228	12.8812251	n/a
229	21.5155327	n/a
230	32.2754391	n/a
231	n/a	54.8395951
232	n/a	41.2553752
233	n/a	28.6262588
234	n/a	19.0326791
235	n/a	10.8510786
236	n/a	6.1919694
237	n/a	3.0782653
238	n/a	1.5050075
239	n/a	0.6467318
240	n/a	0.2622324
241	n/a	0.1025966

RIT	Probability of False Negative	Probability of False Positive
242	n/a	0.0389714
243	n/a	0.0142446
244	n/a	0.00421
245	n/a	0.00211
246	n/a	0.000806
247	n/a	0.000223
248	n/a	0.0000714
249	n/a	0.000032
250	n/a	0.0000113
251	n/a	0.00000498
252	n/a	0.00000147
253	n/a	0.000000365
254	n/a	0.000000365
255	n/a	7.53E-08
256	n/a	5.19E-08
257	n/a	2.32E-08
258	n/a	1.51E-09
259	n/a	5.71E-09
260	n/a	5.31E-09
261	n/a	5.91E-10
262	n/a	1.37E-10
264	n/a	2.7E-10
265	n/a	1.37E-11
271	n/a	1.09E-10

C.8. Grade 8, Reading: TESA (Short)

RIT	Probability of False Negative	Probability of False Positive
196	1.28E-14	n/a
197	1.67E-12	n/a
198	9.32E-11	n/a
199	3.3E-11	n/a
200	6.55E-10	n/a
201	2.94E-08	n/a
202	1.55E-08	n/a
203	6.06E-08	n/a
204	0.00000015	n/a
205	0.0000041	n/a
206	0.00000207	n/a
207	0.0000127	n/a
208	0.00109	n/a
209	0.000617	n/a
210	0.00342	n/a
211	0.00867	n/a
212	0.0228046	n/a
213	0.0463407	n/a
214	0.1058928	n/a
215	0.1806454	n/a
216	0.3510882	n/a
217	0.8028665	n/a
218	1.306316	n/a
219	1.5539585	n/a
220	2.9332959	n/a
221	3.9855997	n/a
222	5.6608404	n/a
223	7.4143036	n/a
224	10.5986146	n/a
225	13.7173432	n/a
226	19.0304373	n/a
227	24.3390222	n/a
228	31.7542945	n/a
229	38.1549669	n/a
230	46.1731328	n/a
231	n/a	46.0749447
232	n/a	38.1504286
233	n/a	30.5824821
234	n/a	24.8636265
235	n/a	18.9438877
236	n/a	14.8692779
237	n/a	10.1811381
238	n/a	7.1872957
239	n/a	5.0093975
240	n/a	3.502348
241	n/a	2.1103968

RIT	Probability of False Negative	Probability of False Positive
242	n/a	1.2374658
243	n/a	1.0181372
244	n/a	0.5374239
245	n/a	0.4732754
246	n/a	0.1871091
247	n/a	0.1145961
248	n/a	0.0705006
249	n/a	0.0450216
250	n/a	0.0296796
251	n/a	0.0279902
252	n/a	0.00649
253	n/a	0.00535
254	n/a	0.00435
255	n/a	0.00273
256	n/a	0.00304
257	n/a	0.00039
258	n/a	0.00062
259	n/a	0.000293
260	n/a	0.000537
261	n/a	0.000228
262	n/a	0.0000201
264	n/a	0.000311
265	n/a	0.00000568
271	n/a	0.000328

C.9. Grade 8, Reading: Paper and Pencil

RIT	Probability of False Negative	Probability of False Positive
174	1.54E-23	n/a
181	3.25E-22	n/a
188	6.01E-21	n/a
192	9.82E-20	n/a
196	1.43E-18	n/a
198	1.86E-17	n/a
200	2.19E-16	n/a
202	2.33E-15	n/a
204	2.26E-14	n/a
205	2E-13	n/a
206	1.62E-12	n/a
208	1.21E-11	n/a
209	8.28E-11	n/a
210	5.26E-10	n/a
211	3.09E-09	n/a
212	1.68E-08	n/a
213	8.54E-08	n/a
214	0.000000404	n/a
215	0.00000178	n/a
215	0.00000732	n/a
216	0.0000282	n/a
217	0.000102	n/a
218	0.000343	n/a
219	0.00109	n/a
219	0.00325	n/a
220	0.0091	n/a
221	0.0239932	n/a
222	0.0595896	n/a
222	0.1394905	n/a
223	0.3079843	n/a
224	0.6418534	n/a
224	1.263568	n/a
225	2.3516932	n/a
226	4.1417014	n/a
227	6.9091691	n/a
227	10.9297294	n/a
228	16.4173243	n/a
229	23.4525035	n/a
229	31.921349	n/a
230	41.4888468	n/a
231	n/a	48.385946
232	n/a	38.3165428
232	n/a	28.9474409
233	n/a	20.7894323
234	n/a	14.1464191
235	n/a	9.0922673

RIT	Probability of False Negative	Probability of False Positive
236	n/a	5.503545
236	n/a	3.1285036
237	n/a	1.6655845
238	n/a	0.8282308
239	n/a	0.3836142
240	n/a	0.1650314
241	n/a	0.0657479
242	n/a	0.0241811
243	n/a	0.00818
245	n/a	0.00254
246	n/a	0.000719
247	n/a	0.000185
249	n/a	0.0000431
251	n/a	0.00000902
253	n/a	0.00000169
255	n/a	0.00000028
259	n/a	4.09E-08
263	n/a	5.21E-09
270	n/a	5.72E-10
277	n/a	5.34E-11

C.10. Grade 10, Math: TESA

RIT	Probability of False Negative	Probability of False Positive
174	4.22E-34	n/a
200	3.36E-27	n/a
201	2.6E-24	n/a
202	7.35E-24	n/a
203	4.23E-23	n/a
204	1.64E-22	n/a
205	2.81E-22	n/a
206	8.47E-21	n/a
207	2.31E-19	n/a
208	2.18E-19	n/a
209	3.48E-17	n/a
210	1.95E-17	n/a
211	7.06E-16	n/a
212	3.26E-15	n/a
213	4.22E-14	n/a
214	5.94E-13	n/a
215	5.53E-12	n/a
216	7.69E-12	n/a
217	1.24E-10	n/a
218	5.42E-10	n/a
219	3.43E-09	n/a
220	1.67E-08	n/a
221	0.000000103	n/a
222	0.000000445	n/a
223	0.00000281	n/a
224	0.0000129	n/a
225	0.0000598	n/a
226	0.000256	n/a
227	0.000964	n/a
228	0.00326	n/a
229	0.011177	n/a
230	0.0329387	n/a
231	0.093787	n/a
232	0.2616742	n/a
233	0.6665741	n/a
234	1.5857826	n/a
235	3.426444	n/a
236	6.6219139	n/a
237	12.0080118	n/a
238	19.1771297	n/a
239	n/a	70.3941283
240	n/a	58.7551245
241	n/a	46.3137904
242	n/a	34.4675668
243	n/a	24.3599328
244	n/a	16.1268178

RIT	Probability of False Negative	Probability of False Positive
245	n/a	10.8712688
246	n/a	6.1968119
247	n/a	3.4258617
248	n/a	1.8215776
249	n/a	0.8707307
250	n/a	0.4723998
251	n/a	0.1966148
252	n/a	0.0751444
253	n/a	0.0289446
254	n/a	0.0109658
255	n/a	0.00277
256	n/a	0.000724
257	n/a	0.000206
258	n/a	0.0000443
259	n/a	0.00000938
260	n/a	0.00000188
261	n/a	0.000000375
262	n/a	6.31E-08
263	n/a	1.88E-08
264	n/a	1.84E-09
265	n/a	7.94E-11
266	n/a	1.77E-10
267	n/a	3.77E-11
268	n/a	1.32E-11
269	n/a	3.31E-13
270	n/a	1.54E-13
271	n/a	2.69E-12
272	n/a	2.6E-13
273	n/a	8.3E-16
274	n/a	2.23E-15
276	n/a	5.49E-17
279	n/a	7.16E-18
280	n/a	2.85E-18

C.12. Grade 10, Math: TESA Short

RIT	Probability of False Negative	Probability of False Positive
174	3.7E-25	n/a
200	5.19E-25	n/a
201	5.01E-18	n/a
202	6.92E-17	n/a
203	1.07E-15	n/a
204	2.36E-16	n/a
205	2.05E-15	n/a
206	8.84E-15	n/a
207	5.22E-11	n/a
208	9.71E-09	n/a
209	0.000000104	n/a
210	1.14E-10	n/a
211	0.000000121	n/a
212	3.74E-08	n/a
213	0.000000124	n/a
214	0.00000472	n/a
215	0.00000934	n/a
216	0.000012	n/a
217	0.0000755	n/a
218	0.0000509	n/a
219	0.000108	n/a
220	0.00039	n/a
221	0.00142	n/a
222	0.004	n/a
223	0.00954	n/a
224	0.0229699	n/a
225	0.0271103	n/a
226	0.0847821	n/a
227	0.1646088	n/a
228	0.3635755	n/a
229	0.7372409	n/a
230	1.1093555	n/a
231	1.6671097	n/a
232	3.0485615	n/a
233	5.4614914	n/a
234	7.3677404	n/a
235	11.4156626	n/a
236	16.5479663	n/a
237	23.0751765	n/a
238	29.8594984	n/a
239	n/a	63.3603166
240	n/a	53.590317
241	n/a	44.1734214
242	n/a	35.625661
243	n/a	24.8442455
244	n/a	20.3645759

RIT	Probability of False Negative	Probability of False Positive
245	n/a	13.7970721
246	n/a	10.0364448
247	n/a	6.4822385
248	n/a	4.0976288
249	n/a	2.4703234
250	n/a	1.4410339
251	n/a	1.3817359
252	n/a	0.4626177
253	n/a	0.3710412
254	n/a	0.1228315
255	n/a	0.1096992
256	n/a	0.0316853
257	n/a	0.0184473
258	n/a	0.00608
259	n/a	0.000982
260	n/a	0.00127
261	n/a	0.000227
262	n/a	0.0000533
263	n/a	0.0000722
264	n/a	0.00000311
265	n/a	0.000000844
266	n/a	0.000000427
267	n/a	0.000000523
268	n/a	8.35E-08
269	n/a	5.2E-09
270	n/a	1.01E-09
271	n/a	0.00000011
272	n/a	2.83E-09
273	n/a	3.04E-11
274	n/a	2.6E-10
276	n/a	2.91E-12
279	n/a	4.62E-13
280	n/a	1.86E-13

C.13. Grade 10, Math: Paper and Pencil

RIT	Probability of False Negative	Probability of False Positive
186	8.04E-14	n/a
193	5.73E-13	n/a
200	3.81E-12	n/a
204	2.38E-11	n/a
208	1.39E-10	n/a
210	7.62E-10	n/a
212	3.92E-09	n/a
214	0.000000019	n/a
216	8.62E-08	n/a
218	0.000000369	n/a
219	0.00000149	n/a
220	0.00000567	n/a
222	0.0000203	n/a
223	0.0000689	n/a
224	0.000221	n/a
225	0.000668	n/a
226	0.00191	n/a
227	0.00519	n/a
228	0.0133303	n/a
229	0.0324534	n/a
230	0.0749322	n/a
231	0.1641839	n/a
232	0.3415979	n/a
233	0.6753093	n/a
234	1.2693763	n/a
235	2.2703848	n/a
236	3.8670938	n/a
237	6.2783339	n/a
238	9.7259476	n/a
239	n/a	85.55201
240	n/a	72.364028
241	n/a	64.0218195
242	n/a	54.9315438
243	n/a	45.5638596
244	n/a	36.4314121
245	n/a	28.0068432
246	n/a	20.6526127
247	n/a	14.5782157
248	n/a	9.8320822
250	n/a	6.3252755
251	n/a	3.8758251
252	n/a	2.2590344
253	n/a	1.2509268
255	n/a	0.6573695
256	n/a	0.3274958
258	n/a	0.1545228

RIT	Probability of False Negative	Probability of False Positive
259	n/a	0.0689854
261	n/a	0.0291136
264	n/a	0.011604
266	n/a	0.00436
270	n/a	0.00155
274	n/a	0.000517
281	n/a	0.000162
289	n/a	0.0000479

APPENDIX D

DIF ANALYSIS OF TEST LENGTH, 2005-06

D.1. Grade 3 Reading DIF Results

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
1	R0000160	762	1267	-1.627543	-1.170113	-3.375926946
2	R0113750	3234	3224	-1.47164	-0.759252	-8.587169151
3	R0118250	1720	1867	-1.539922	-0.465575	-11.17405501
4	R0207610	2219	1557	-0.741925	-0.398939	-3.989136408
5	R0224210	1442	3244	-1.274523	-0.995604	-2.939094789
6	R0236330	3670	1567	-0.042333	-0.573065	5.147392203
7	R0238670	1100	2514	-1.160193	-0.561346	-6.787166323
8	R0254760	1917	661	0.181989	0.454876	-2.593809732
9	R0000180	5992	762	-1.685323	-0.771162	-7.203884592
10	R0113760	1227	3232	-1.124011	-0.964895	-1.96416439
11	R0118260	3187	1719	-2.126079	-1.345317	-6.430447624
12	R0207620	1833	2218	-1.207678	-0.649336	-5.548354273
13	R0224220	1391	1441	-0.300534	-0.51981	2.575037948
14	R0224730	2476	2899	-0.303496	-0.780115	4.511110877
15	R0225120	657	1563	-0.787813	-1.033525	2.853506938
16	R0236340	2696	1567	-1.704087	-1.294751	-3.58292416
17	R0238680	979	2513	0.39751	0.099515	3.680058525
18	R0254780	1908	661	-1.230927	-0.927464	-2.61469836
19	R0113780	3232	3140	-0.696352	-0.405157	-3.970369232
20	R0207660	2213	1327	-0.450009	0.217817	-6.850812194
21	R0224240	1440	2347	0.075341	0.345386	-3.298790868
22	R0236350	1565	922	-0.942948	-0.590359	-3.253917035
23	R0000210	761	1173	-1.38257	-0.466198	-6.977588758
24	R0113790	3232	3089	-0.870091	0.047095	-11.92285227
25	R0118300	1719	1803	-0.583557	-0.308779	-3.188588843
26	R0207680	2212	1281	-1.375254	-0.942358	-4.656046755
27	R0225160	1563	2341	-1.524593	-1.180133	-3.243735395
28	R0236370	1565	879	-1.222297	-0.188326	-9.344611047
29	R0238710	2512	1894	0.024948	-0.723449	10.13968587
30	R0254800	661	5875	-0.680276	-0.188371	-4.583277797
31	R0008480	6474	599	-1.046311	-0.643084	-3.876034128
32	R0113270	1784	2047	-1.38255	-0.881148	-5.637949081
33	R0208440	8113	1042	1.184638	1.620089	-5.192688604
34	R0232590	5217	2638	-0.294454	-0.152677	-2.20370074
35	R0236400	1564	831	-1.616159	-0.929486	-5.885772091
36	R0238720	2513	1858	-1.627442	-0.592407	-11.38911414
37	R0238750	2211	1202	-1.394363	-1.155326	-2.325436662
38	R0255050	3452	2509	0.250945	0.451581	-2.619157537
39	R0268210	4651	1281	0.481081	0.66337	-2.112474198
40	R0274960	2900	553	-0.333784	-0.011616	-2.899306045
41	R0007680	3231	1272	-2.493166	-2.226145	-2.218905817
42	R0113300	1751	1693	-0.886407	-0.551021	-3.742349538

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
43	R0113820	3228	2512	-0.706075	-0.127564	-7.994216182
44	R0127000	1719	1416	-0.355186	-0.072377	-3.037351105
45	R0131120	2255	1760	-0.667067	-0.334364	-4.172825533
46	R0203520	5588	6514	-0.702656	-0.902399	2.829116635
47	R0206520	2265	5940	-1.052283	-0.797128	-3.74817418
48	R0208470	8046	1065	1.027973	1.485741	-5.320841574
49	R0232610	5012	2786	-1.193597	-0.894543	-3.93882925
50	R0238730	2512	1821	-0.915612	-0.519203	-4.860792384
51	R0255080	3132	2806	-0.803891	-0.429427	-5.296593802
52	R0275000	2899	511	-1.539041	-1.151578	-3.158982921
53	R0109910	4136	1452	-1.713755	-1.481159	-2.556120678
54	R0131140	2162	1810	-0.466318	-0.07773	-4.642777068
55	R0255090	2815	3119	-0.729948	-0.46859	-3.708437375
56	R0259700	7926	1049	0.564721	0.728599	-2.159677977
57	R0109940	3362	2214	-1.715408	-1.494362	-2.438216132
58	R0131150	2105	1737	-0.22903	-0.047994	-2.27006518
59	R0205050	4104	7856	-0.706586	-0.541718	-2.557632843
60	R0232660	4009	3745	0.215437	0.465856	-3.819368932
61	R0238260	4558	4957	-0.32415	0.06199	-6.415520267
62	R0255100	2544	3370	0.196695	0.391422	-2.618849467
63	R0255780	2323	1617	-1.393213	-1.140048	-2.612618421
64	R0008690	3681	3661	-0.80001	-0.641379	-2.323687416
65	R0109970	2583	2959	-0.965982	-0.784595	-2.251351873
66	R0113350	1542	1638	-0.629577	-0.349116	-2.907529663
67	R0206590	1554	6408	-1.564186	-1.065888	-5.659844391
68	R0208510	7161	1900	0.662736	0.316694	5.228028141
69	R0232680	3838	3844	-0.075764	0.089754	-2.599793808
70	R0238760	4301	5200	-1.97596	-1.794189	-2.170240774
71	R0259720	6706	2252	0.694346	0.824063	-2.037408102
72	R0268250	3589	2307	0.915327	1.109272	-3.030485955
73	R0131740	2061	1573	-0.54025	-0.374625	-2.012404454
74	R0206610	1420	6370	-0.015896	0.145017	-2.224856858
75	R0235860	2990	2382	-0.262118	0.11193	-4.557999976
76	R0247380	3803	3787	-0.534153	-0.371763	-2.581220777
77	R0259730	6497	2403	0.291693	0.553058	-4.355305733
78	R0008740	3253	3997	-1.974408	-1.676227	-3.569016968
79	R0238180	812	5657	-1.067816	-0.784545	-2.886396661

D.2. Grade 8 Reading DIF Results

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
1	R0258670	3090	1280	-0.347603	-0.110927	-2.279361737

D.3. Grade 5 Math DIF Results

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
1	M0011240	3635	2371	-1.0753	-0.91881	-2.186528707
2	M0012050	3712	5226	1.055589	0.770783	3.915578794
3	M0209030	4669	7712	-0.80621	-0.639	-2.890379334
4	M0210210	5884	3322	0.078685	0.323675	-4.122383921
5	M0211060	7054	5466	-0.09939	0.043667	-2.589614793
6	M0211910	6575	2676	-0.63845	-0.45371	-2.978183905
7	M0212150	3988	4505	-0.99111	-0.84843	-2.308037358
8	M0220340	6505	8025	-0.36534	-0.18331	-3.308777255
9	M0242830	4181	4218	-0.67623	-0.49947	-3.112382935
10	M0243370	4807	3208	-0.6713	-0.4984	-2.680879278
11	M0244750	6758	7762	0.262579	0.430017	-3.63536077
12	M0244980	5990	8017	0.014758	0.175592	-2.851557012
13	M0245860	6974	9338	0.288551	0.495659	-4.784998907
14	M0247530	6669	8617	-0.24701	-0.10362	-2.778999878
15	M0274230	6944	7531	0.701095	0.884227	-2.717643897
16	M0211260	7536	873	1.093749	1.284731	-2.392681624
17	M0011350	2150	108	-2.73159	-1.85618	-3.914715616
18	M0107950	6510	7241	0.357888	0.536712	-3.330042126
19	M0111500	5539	8562	0.501991	0.682404	-2.552163195
20	M0209760	5675	6462	0.326856	0.471526	-2.58117475
21	M0211180	5525	8044	0.615978	0.747819	-2.391656868
22	M0212100	4987	5323	0.898864	0.685108	2.778014915
23	M0221080	6007	3645	0.781108	0.960589	-2.471993973
24	M0242500	5621	6048	0.606424	0.780628	-2.818565219
25	M0242750	1511	180	-2.82456	-1.66626	-5.533608627
26	M0252370	5134	7039	0.756391	0.938784	-3.260013513
27	M0264620	6286	7676	0.363419	0.562565	-3.943121226
28	M0011590	1407	1080	-2.43292	-2.09819	-2.999157144
29	M0101850	5865	5643	1.278246	0.823112	5.437952435
30	M0205920	864	575	-1.09849	-0.55033	-3.76134253
31	M0210710	3824	4262	-0.10102	0.129986	-3.349508821
32	M0211000	945	1434	-2.19193	-1.96582	-1.994810727

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
33	M0211990	1667	770	-1.84032	-1.5473	-2.621461753
34	M0241240	4458	3118	-0.10583	0.0349	-2.253735873
35	M0245050	6673	8552	-0.08807	0.243498	-7.185959562
36	M0246080	3678	4777	-1.50784	-1.35367	-2.291241338
37	M0246540	1583	1039	-2.61155	-2.3055	-2.746310528
38	M0264350	1409	1969	-1.50654	-1.11338	-4.409214823
39	M0264570	3454	2953	-1.34698	-1.08085	-3.709885416
40	M0274220	4766	1693	1.374262	1.625435	-2.926376746
41	M02E2340	3485	1705	-0.91083	-0.5611	-4.624257113
42	M02E3790	1537	1418	-1.64636	-1.4199	-2.427455104
43	M0011270	661	220	-1.91114	-1.52291	-2.105222637
44	M0102890	2799	3600	-1.72853	-1.56687	-2.060857134
45	M0211480	5547	3158	1.022502	1.217466	-2.919273215
46	M0211590	2513	2648	-1.16852	-0.90935	-3.421038046
47	M0211700	857	327	-2.22511	-1.54254	-4.13056509
48	M0220360	2309	4964	-1.46364	-1.27752	-2.638268373
49	M0242390	1661	2439	-1.34734	-1.03396	-3.428504596
50	M0243190	2407	3550	-1.68161	-1.47646	-3.042028545
51	M0244670	2505	2176	-1.38967	-1.22472	-2.113987286
52	M0249490	2447	2496	-1.60488	-1.39609	-2.752909834
53	M0011360	959	1209	-1.52262	-1.05072	-4.304746998
54	M0242510	1236	1353	-1.63772	-1.36475	-2.618195108
55	M0244940	753	335	-1.93785	-1.18257	-4.358257686
56	M0274340	4391	1081	1.146572	1.332704	-2.234613181
57	M0242150	4287	3564	1.03798	1.288604	-3.971052891
58	M0248690	1176	1886	-0.99461	-0.81466	-2.025871661
59	M0210750	3552	2426	2.1659	1.641115	6.398655143
60	M0212680	475	1275	-1.77053	-1.47195	-2.294134977
61	M0241720	4180	4156	1.069318	0.88501	2.224104571
62	M0270080	1267	2195	-1.0995	-0.83403	-3.221681771

D.4. Grade 10 Math DIF Results

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
1	M0013560	4183	2578	2.278939	1.862498	3.407729698
2	M0206390	2066	8158	0.406287	0.549038	-2.07020457
3	M0206600	5004	3863	1.233972	0.803794	4.00916437
4	M0213330	1904	5095	0.889456	0.613159	2.97642949
5	M0213830	4760	5358	2.074775	2.233973	-2.128963685
6	M0214690	4738	4253	0.427866	0.601245	-3.490517805
7	M0224740	6136	5012	0.400215	0.535186	-2.467817869
8	M0224900	3301	2968	0.854706	1.437268	-8.952358316
9	M0225000	4901	3449	0.675484	0.849336	-2.591491306
10	M0225020	4550	4795	0.274826	0.478583	-3.600739656
11	M0225750	10073	4903	0.315579	0.553278	-3.429417309
12	M0226230	9056	5236	0.150205	0.446876	-6.785173428
13	M0226580	4680	5326	0.545369	0.782429	-3.795714157
14	M0227080	2899	5330	0.745349	1.051335	-5.391339568
15	M0227310	2019	9061	0.603974	0.869031	-3.358435111
16	M0231130	6330	4414	1.006689	1.244378	-3.905675659
17	M0232980	2693	4976	0.398361	0.611679	-3.27076989
18	M0427840	5649	6640	0.634782	0.828215	-3.149275279
19	M0251120	5005	174	-0.69675	0.028702	-2.87664351
20	M0251750	6065	144	-0.29602	-1.74234	2.399803865
21	M0102310	6074	1258	-1.72093	-1.24405	-2.917480739
22	M0114650	3339	664	1.247286	0.620579	5.936290453
23	M0209420	2108	213	2.76803	2.331327	1.96465098
24	M0213390	1245	2168	1.312869	1.526981	-2.191174938
25	M0215210	3112	268	1.595453	0.948023	3.621474891
26	M0221610	3497	727	1.166124	0.778072	3.231538823
27	M0224780	3930	1437	1.5261	1.787687	-3.207405171
28	M0226000	6467	1181	-1.25402	-0.8784	-2.274133265
29	M0226090	2668	2694	1.170819	1.002721	1.984089413
30	M0226480	3432	702	2.798373	3.359881	-3.365347954
31	M0228230	746	2438	1.024834	1.393609	-3.361200298
32	M0251690	4387	1388	-1.87649	-1.50495	-2.49873554
33	M02E0230	3846	432	2.140125	2.428729	-2.378048244
34	M0106540	4127	6762	-0.23122	0.22146	-5.406914182
35	M0106690	7678	6885	0.105845	0.367299	-4.420791234
36	M0214760	3809	385	1.989512	1.515831	2.815249621
37	M0215180	2768	234	2.272893	2.646091	-2.556313594
38	M0224050	8049	6385	-0.04544	0.104296	-1.969216949
39	M0226120	6928	5299	0.361167	0.532655	-2.357127419
40	M0226240	5720	5125	0.310363	0.665485	-4.918190103
41	M0226670	3503	9344	0.498909	0.676487	-2.35334978
42	M0230350	2439	5896	-1.27824	-0.89625	-2.910962844
43	M0270330	3795	3214	-0.97854	-0.77696	-2.285308237
44	M0276270	2486	407	3.267292	4.555701	-5.126432087
45	M0006250	5839	7463	-0.48389	-0.21263	-2.760023843
46	M0206640	4337	1954	-2.61098	-2.20809	-2.605547048

Item Count	Item Name	Number of Times Item Appears in Block A	Number of Times Item Appears in Block B	b-Value Block A	b-Value Block B	T-Test
47	M0214370	1005	1066	2.005205	2.479114	-3.907789944
48	M0225040	7226	7861	-0.60302	-0.4099	-2.499700593
49	M0229730	5812	8076	-0.43064	-0.16523	-2.226241049
50	M0214600	7300	3886	-0.89151	-0.63176	-2.545614409
51	M0227270	4430	9947	0.179701	0.381828	-2.272129633

APPENDIX E
SUPPORTING TABLES FOR FIRST YEAR STUDY, 2003

Number of Participants

E.1. Math

OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	2032	2532	1443	1509	2555
1st Year College GPA	2039	2542	1473	1759	2799
1st Year Math GPA	941	2586	154	928	1748
Returned 2nd Fall	2077	2573	1480	1826	2839
SAT I	1932	2492	1448	1393	2839

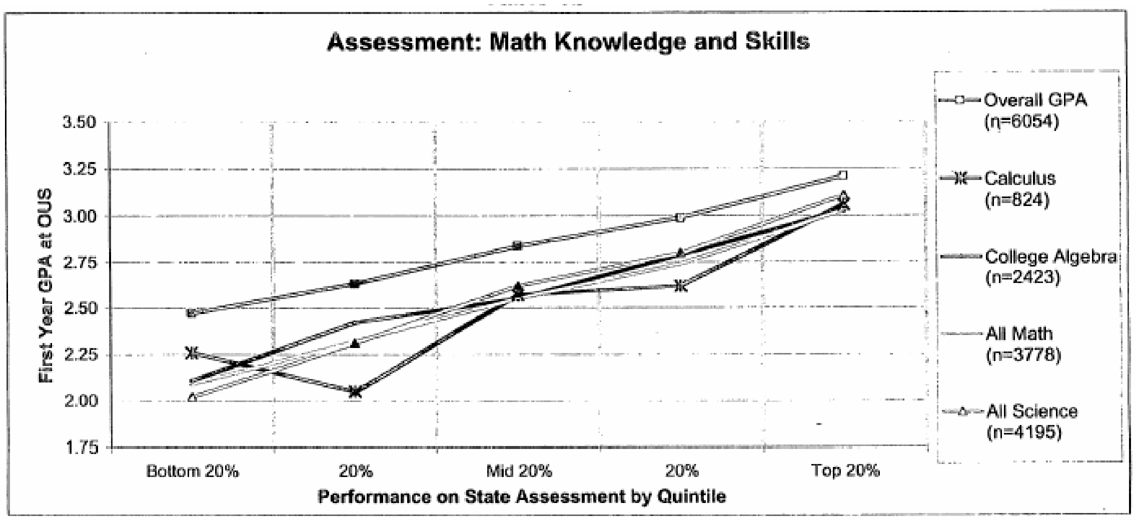
E.2. Reading

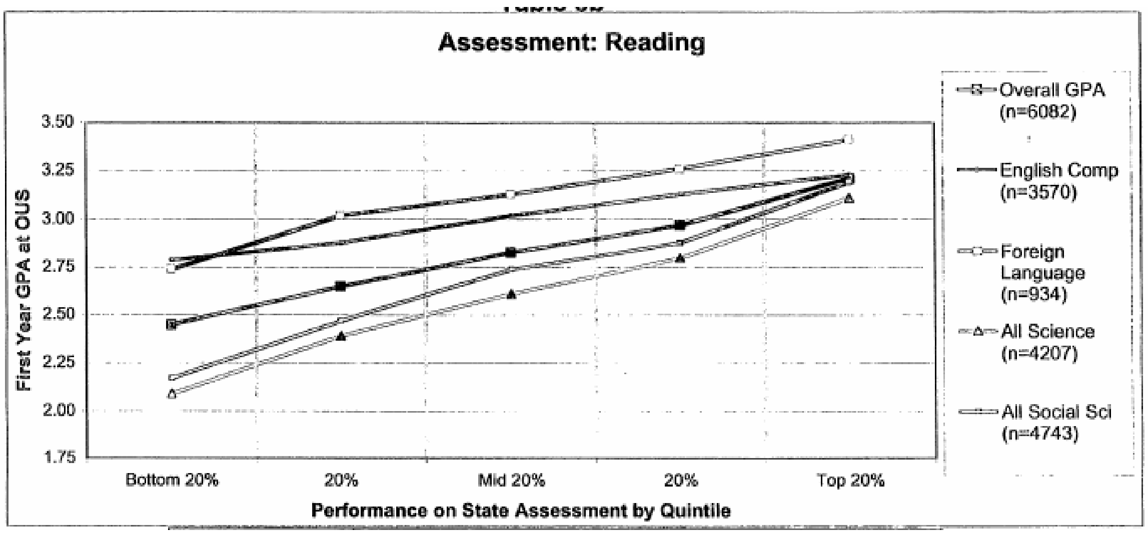
OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	1081	3070	1873	1492	2555
1st Year College GPA	1086	3100	1896	1731	2799
1st Year Arts/Letters GPA	662	2051	1339	1066	1892
Returned 2nd Fall	1107	3133	1916	1800	2839
SAT I	1021	3004	1867	1373	2472

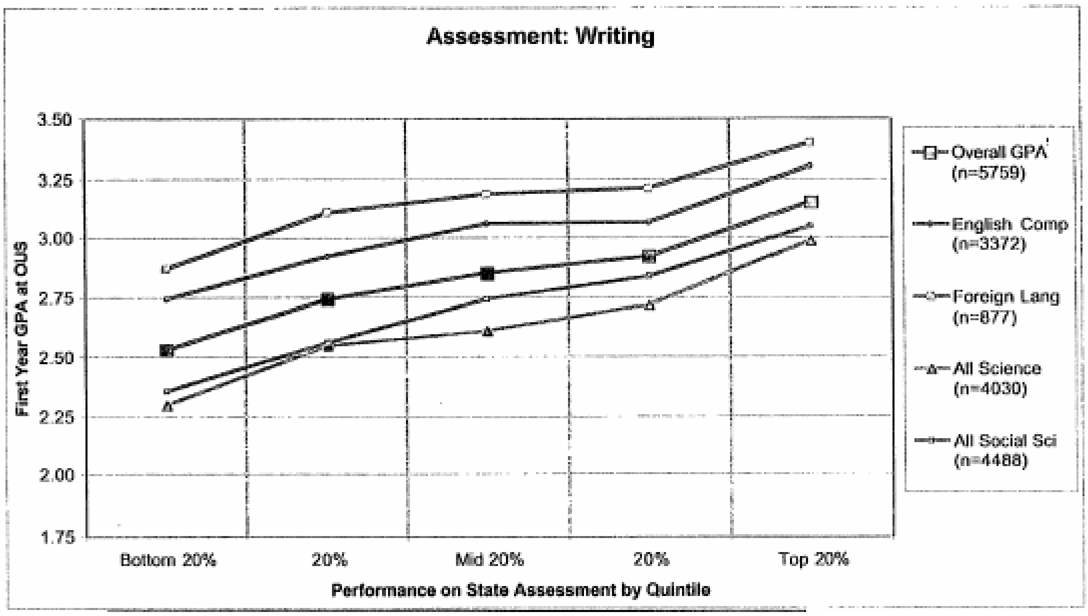
E.3. Writing

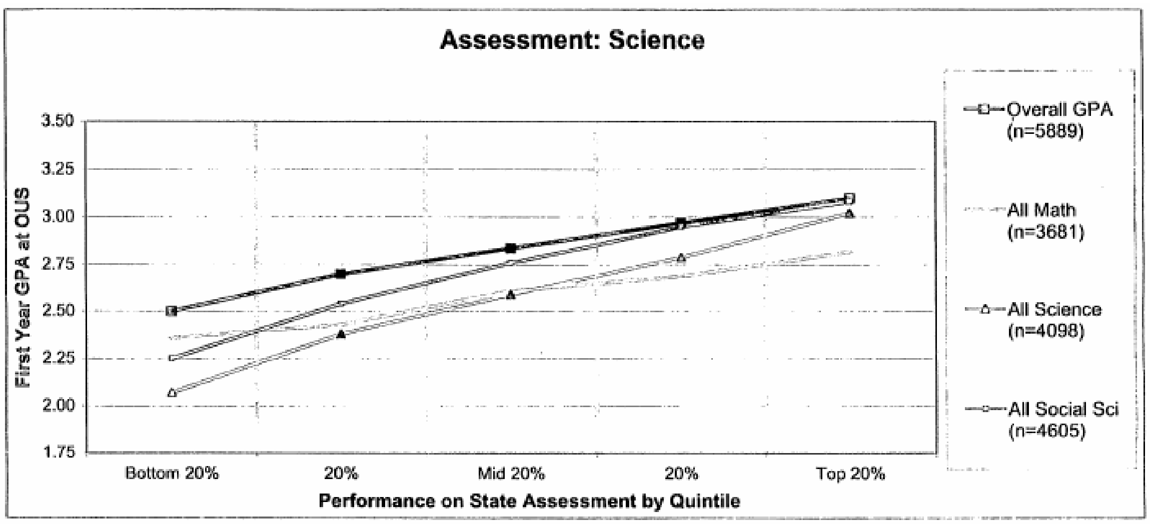
OUS Measure					
	Not Met*	Meets	Exceeds	No Score**	Non-Oregon
HS GPA	1833	3460	417	1806	2555
1st Year College GPA	1844	3492	423	2054	2799
1st Year Writing GPA	1124	2035	213	884	1633
Returned 2nd Fall	1871	3537	423	2125	2839
SAT I	1761	3405	411	1688	2472

E.4. Performance on State Assessment and First-Year College GPA at OUS

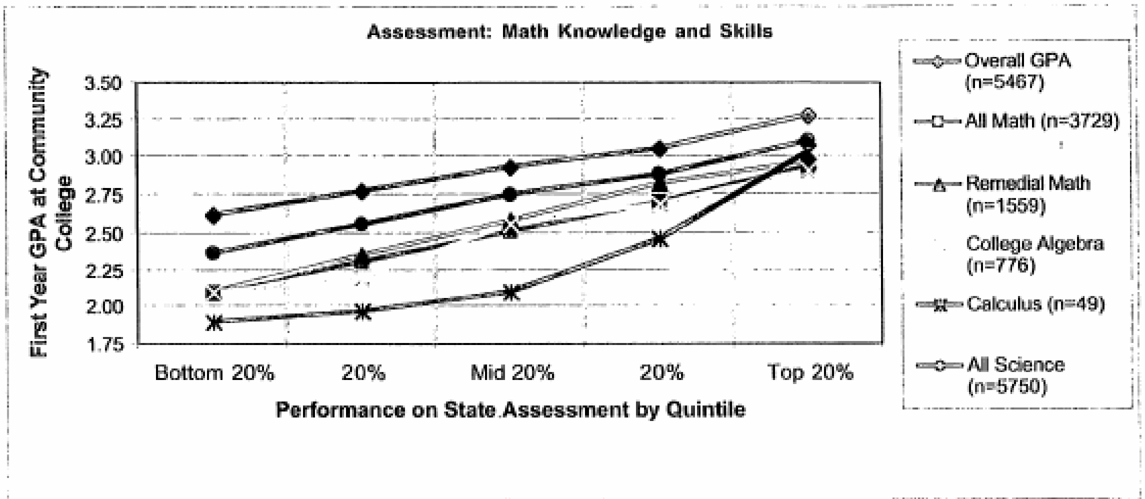


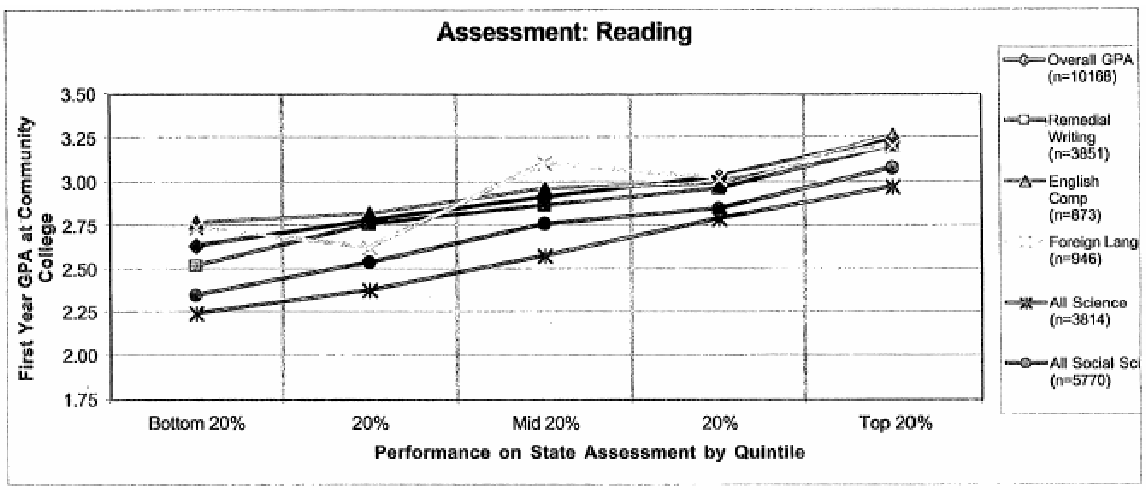


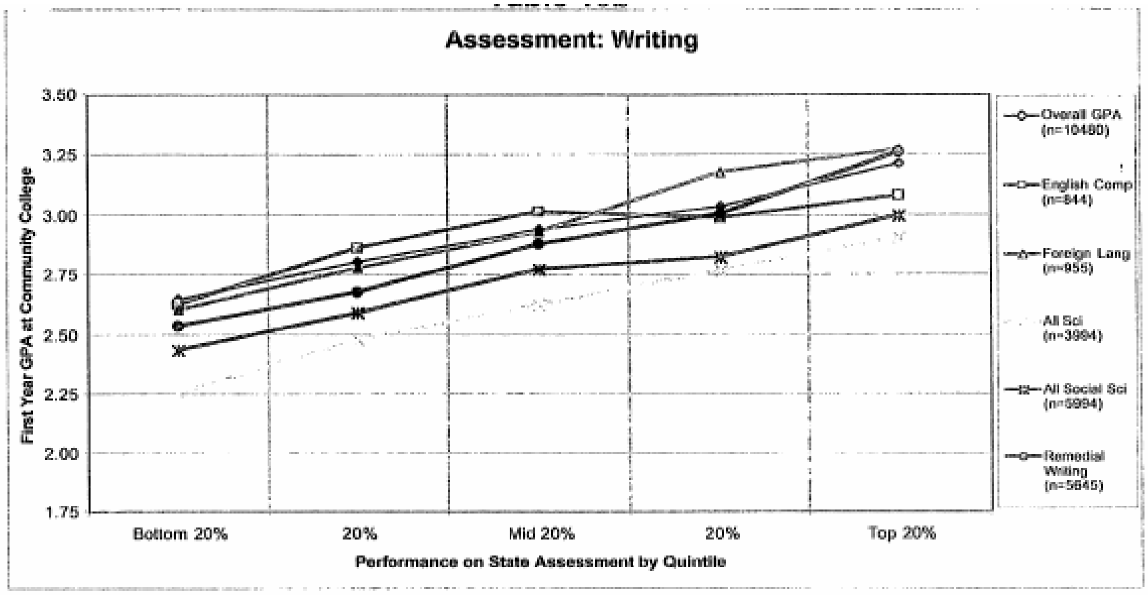

Assessment: Writing
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
First Year GPA at OUS
Bottom 20%
20%
Mid 20%
20%
Top 20%
Performance on State Assessment by Quintile
Overall GPA (n=5759)
English Comp (n=3372)
Foreign Lang (n=877)
All Science (n=4030)
All Social Sci (n=4488)

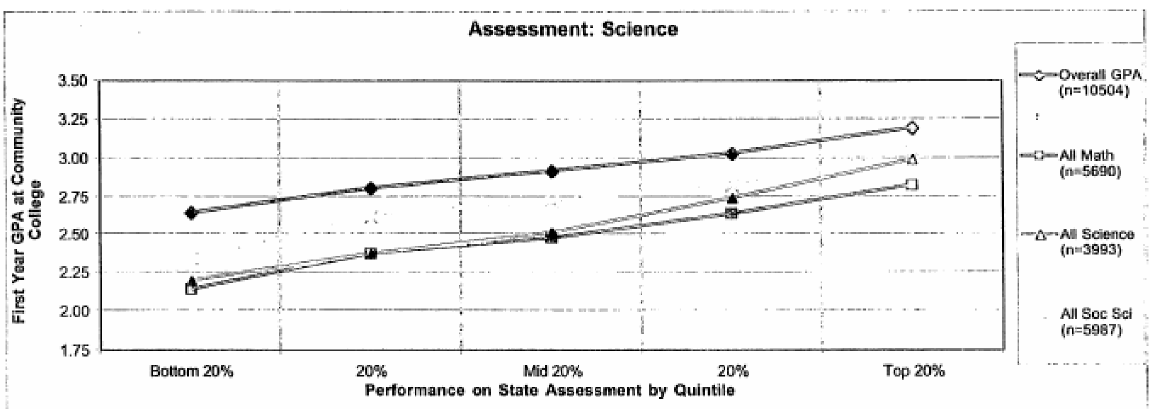

Assessment: Science
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
First Year GPA at OUS
Bottom 20%
20%
Mid 20%
20%
Top 20%
Performance on State Assessment by Quintile
Overall GPA (n=5889)
All Math (n=3681)
All Science (n=4098)
All Social Sci (n=4605)

E.5. Performance on State Assessment and First-Year College GPA at Community Colleges






Assessment: Writing
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
First Year GPA at Community College
Bottom 20%
20%
Mid 20%
20%
Top 20%
Performance on State Assessment by Quintile
Overall GPA (n=10480)
English Comp (n=844)
Foreign Lang (n=955)
All Sci (n=3994)
All Social Sci (n=5994)
Remedial Writing (n=5645)


Assessment: Science
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
First Year GPA at Community College
Bottom 20%
20%
Mid 20%
20%
Top 20%
Performance on State Assessment by Quintile
Overall GPA (n=10504)
All Math (n=5690)
All Science (n=3993)
All Soc Sci (n=5987)

APPENDIX F
CLASSIFICATION ANALYSIS – TESA AND PAPER-AND-PENCIL, 2003–04

F.1. Mathematics

Mathematics Grade 3 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	51.0	4.5	0.0	55.6		
Meets	5.2	31.6	2.8	39.6		
Exceeds	0.0	1.1	3.7	4.8	86.4	overall accuracy
Total	56.2	37.3	6.5		90.3	proficiency accuracy

Mathematics Grade 3 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	18.7	2.7	0.0	21.3		
Meets	3.8	44.9	6.1	54.9		
Exceeds	0.0	3.3	20.6	23.8	84.1	overall accuracy
Total	22.5	50.8	26.7		93.5	proficiency accuracy

Mathematics Grade 3 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	1.0	0.3	0.0	1.2		
Meets	0.7	20.6	3.6	24.9		
Exceeds	0.0	5.6	68.3	73.9	89.8	overall accuracy
Total	1.6	26.4	71.9		99.0	proficiency accuracy

Mathematics Grade 3 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	51.3	4.6	0.0	55.8		
Meets	4.5	32.9	1.7	39.1		
Exceeds	0.0	1.3	3.8	5.1	87.9	overall accuracy
Total	55.8	38.8	5.4		90.9	proficiency accuracy

Mathematics Grade 3 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	26.8	2.8	0.0	29.6		
Meets	3.4	36.8	3.3	43.5		
Exceeds	0.0	3.1	23.8	26.9	87.4	overall accuracy
Total	30.2	42.7	27.1		93.8	proficiency accuracy

Mathematics Grade 5 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	60.4	4.9	0.0	65.3		
Meets	4.5	28.1	0.9	33.4		
Exceeds	0.0	0.4	0.8	1.3	89.3	overall accuracy
Total	64.9	33.4	1.7		90.6	proficiency accuracy

Mathematics Grade 5 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	19.7	2.8	0.0	22.5		
Meets	3.8	56.1	4.4	64.2		
Exceeds	0.0	2.3	11.1	13.3	86.8	overall accuracy
Total	23.4	61.1	15.4		93.5	proficiency accuracy

Mathematics Grade 5 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	1.9	0.4	0.0	2.3		
Meets	0.8	29.7	3.9	34.4		
Exceeds	0.0	5.3	58.0	63.4	89.6	overall accuracy
Total	2.7	35.4	61.9		98.8	proficiency accuracy

Mathematics Grade 5 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	51.8	5.0	0.0	56.9		
Meets	4.7	36.2	0.8	41.8		
Exceeds	0.0	0.4	1.0	1.4	89.0	overall accuracy
Total	56.6	41.7	1.7		90.2	proficiency accuracy

Mathematics Grade 5 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	25.9	2.7	0.0	28.6		
Meets	3.3	45.4	3.0	51.7		
Exceeds	0.0	2.6	17.0	19.7	88.3	overall accuracy
Total	29.2	50.7	20.1		94.0	proficiency accuracy

Mathematics Grade 8 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	85.9	3.4	0.0	89.3		
Meets	1.9	6.7	1.0	9.6		
Exceeds	0.0	0.2	0.8	1.1	93.5	overall accuracy
Total	87.8	10.3	1.8	*	94.7	proficiency accuracy

Mathematics Grade 8 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	41.2	4.0	0.0	45.2		
Meets	4.3	22.5	4.1	30.9		
Exceeds	0.0	3.2	20.6	23.9	84.4	overall accuracy
Total	45.6	29.7	24.7		91.7	proficiency accuracy

Mathematics Grade 8 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	6.8	1.1	0.0	7.9		
Meets	1.8	12.3	2.8	17.0		
Exceeds	0.0	3.7	71.5	75.1	90.6	overall accuracy
Total	8.6	17.1	74.3		97.1	proficiency accuracy

Mathematics Grade 8 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	75.3	4.1	0.0	79.5		
Meets	3.0	12.1	1.6	16.7		
Exceeds	0.0	0.8	3.0	3.8	90.4	overall accuracy
Total	78.3	17.1	4.6		92.8	proficiency accuracy

Mathematics Grade 8 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	46.1	3.4	0.0	49.5		
Meets	3.4	17.6	3.1	24.1		
Exceeds	0.0	2.6	23.8	26.4	87.5	overall accuracy
Total	49.5	23.5	26.9		93.2	proficiency accuracy

Mathematics Grade 10 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	88.8	2.9	0.0	91.7		
Meets	1.7	5.9	0.3	7.9		
Exceeds	0.0	0.1	0.3	0.4	95.0	overall accuracy
Total	90.5	8.9	0.6		95.5	proficiency accuracy

Mathematics Grade 10 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	51.0	4.2	0.0	55.2		
Meets	4.1	26.2	2.9	33.2		
Exceeds	0.0	2.1	9.5	11.6	86.7	overall accuracy
Total	55.1	32.5	12.4		91.7	proficiency accuracy

Mathematics Grade 10 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	14.7	2.3	0.0	17.0		
Meets	3.1	27.1	4.1	34.2		
Exceeds	0.0	4.1	44.7	48.8	86.4	overall accuracy
Total	17.8	33.5	48.7		94.6	proficiency accuracy

Mathematics Grade 10 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	90.7	2.6	0.0	93.3		
Meets	1.3	5.0	0.2	6.5		
Exceeds	0.0	0.0	0.1	0.1	95.8	overall accuracy
Total	92.1	7.6	0.3		96.0	proficiency accuracy

Mathematics Grade 10 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	63.6	3.9	0.0	67.5		
Meets	3.4	19.5	1.9	24.7		
Exceeds	0.0	1.4	6.4	7.8	89.5	overall accuracy
Total	67.0	24.7	8.3		92.8	proficiency accuracy

F.2. Reading

Reading Grade 3 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	51.0	4.0	0.0	55.0		
Meets	4.4	32.2	2.8	39.4		
Exceeds	0.0	1.2	4.4	5.6	87.6	overall accuracy
Total	55.4	37.4	7.2		91.6	proficiency accuracy

Reading Grade 3 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	13.5	1.9	0.0	15.4		
Meets	2.8	41.3	4.1	48.2		
Exceeds	0.0	5.3	31.1	36.4	85.9	overall accuracy
Total	16.3	48.5	35.2		95.3	proficiency accuracy

Reading Grade 3 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	0.1	0.1	0.0	0.2		
Meets	0.2	13.3	2.4	16.0		
Exceeds	0.0	3.9	79.9	83.8	93.4	overall accuracy
Total	0.3	17.3	82.4		99.7	proficiency accuracy

Reading Grade 3 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	69.9	3.5	0.0	73.4		
Meets	4.3	19.1	1.4	24.8		
Exceeds	0.0	0.5	1.3	1.8	90.3	overall accuracy
Total	74.2	23.1	2.8		92.2	proficiency accuracy

Reading Grade 3 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	20.0	2.4	0.0	22.4		
Meets	2.9	39.5	3.5	45.9		
Exceeds	0.0	3.5	28.1	31.7	87.6	overall accuracy
Total	22.9	45.4	31.7		94.6	proficiency accuracy

Reading Grade 5 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	62.5	4.2	0.0	66.7		
Meets	3.8	27.2	1.2	32.2		
Exceeds	0.0	0.2	0.8	1.1	90.6	overall accuracy
Total	66.3	31.6	2.1		92.0	proficiency accuracy

Reading Grade 5 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	17.3	2.6	0.0	19.9		
Meets	3.6	58.4	4.0	66.0		
Exceeds	0.0	2.3	11.8	14.1	87.6	overall accuracy
Total	20.9	63.3	15.8		93.9	proficiency accuracy

Reading Grade 5 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	0.6	0.2	0.0	0.8		
Meets	0.6	33.8	4.8	39.2		
Exceeds	0.0	5.7	54.3	60.0	88.7	overall accuracy
Total	1.2	39.7	59.2		99.2	proficiency accuracy

Reading Grade 5 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	75.1	3.6	0.0	78.7		
Meets	3.6	16.6	0.6	20.7		
Exceeds	0.0	0.2	0.4	0.5	92.1	overall accuracy
Total	78.7	20.3	0.9		92.8	proficiency accuracy

Reading Grade 5 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	28.0	3.2	0.0	31.1		
Meets	3.7	49.5	2.9	56.2		
Exceeds	0.0	2.6	10.1	12.7	87.6	overall accuracy
Total	31.7	55.3	13.0		93.1	proficiency accuracy

Reading Grade 8 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	82.9	3.2	0.0	86.1		
Meets	2.3	8.5	0.9	11.7		
Exceeds	0.0	0.6	1.6	2.2	93.0	overall accuracy
Total	85.2	12.3	2.5		94.5	proficiency accuracy

Reading Grade 8 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	35.9	4.4	0.0	40.3		
Meets	5.1	26.6	4.7	36.5		
Exceeds	0.0	3.7	19.5	23.2	82.1	overall accuracy
Total	41.0	34.7	24.3		90.5	proficiency accuracy

Reading Grade 8 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	2.0	0.6	0.0	2.6		
Meets	1.6	13.7	3.7	18.9		
Exceeds	0.0	5.3	73.2	78.5	88.8	overall accuracy
Total	3.6	19.6	76.8		97.8	proficiency accuracy

Reading Grade 8 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	81.3	3.7	0.0	85.0		
Meets	2.2	7.4	2.1	11.6		
Exceeds	0.0	0.7	2.6	3.3	91.3	overall accuracy
Total	83.5	11.8	4.7		94.0	proficiency accuracy

Reading Grade 8 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	50.3	4.4	0.0	54.8		
Meets	4.4	21.3	3.5	29.2		
Exceeds	0.0	2.6	13.5	16.1	85.1	overall accuracy
Total	54.7	28.3	17.0		91.2	proficiency accuracy

Reading Grade 10 Paper/Pencil Form B

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	54.1	4.5	0.0	58.6		
Meets	4.3	26.0	2.6	33.0		
Exceeds	0.0	1.7	6.7	8.4	86.8	overall accuracy
Total	58.3	32.3	9.3		91.2	proficiency accuracy

Reading Grade 10 Paper/Pencil Form C

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	12.2	2.3	0.0	14.5		
Meets	3.6	32.9	5.1	41.5		
Exceeds	0.0	4.8	39.2	44.0	84.3	overall accuracy
Total	15.8	40.0	44.3		94.2	proficiency accuracy

Reading Grade 10 Paper/Pencil Form A

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	87.9	2.5	0.0	90.4		
Meets	1.7	6.5	0.6	8.9		
Exceeds	0.0	0.2	0.5	0.7	95.0	overall accuracy
Total	89.7	9.2	1.2		95.8	proficiency accuracy

Reading Grade 10 Paper/Pencil Plain Language

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	88.3	2.4	0.0	90.7		
Meets	2.1	5.4	0.6	8.1		
Exceeds	0.0	0.4	0.8	1.2	94.5	overall accuracy
Total	90.4	8.2	1.4		95.5	proficiency accuracy

Reading Grade 10 TESA Wide Range

True Score Category	Expected Score Category: Not Yet Met	Meets	Exceeds	Total		
Not Yet Met	58.0	4.5	0.0	62.5		
Meets	4.1	23.9	2.3	30.2		
Exceeds	0.0	1.6	5.6	7.2	87.5	overall accuracy
Total	62.1	29.9	7.9		91.4	proficiency accuracy

APPENDIX G
CORRELATIONS OF ALL SUBJECT TESTS WITH NWEA TESTS AT GRADES 5, 8, AND 10, 2004

Grade of Enrollment = 05		ODE Math K/S	ODE Read/Lit	NWEA Math	NWEA Reading	ODE Science	ODE Mth ProbSolv	ODE Writing
ODE Math K/S	Pearson Correlation	1	0.754	0.755	0.667	0.718	0.553	0.537
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000
	N	1,333	1,303	1,231	1,217	1,310	1,301	1,302
ODE Read/Lit	Pearson Correlation	0.754	1	0.670	0.811	0.770	0.480	0.601
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000	0.000
	N	1,303	1,318	1,218	1,204	1,298	1,285	1,284
NWEA Math	Pearson Correlation	0.755	0.670	1	0.753	0.648	0.496	0.513
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000	0.000
	N	1,231	1,218	1,275	1,157	1,249	1,235	1,238
NWEA Reading	Pearson Correlation	0.667	0.811	0.753	1	0.685	0.418	0.621
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000	0.000
	N	1,217	1,204	1,157	1,260	1,237	1,218	1,221
ODE Science	Pearson Correlation	0.718	0.770	0.648	0.685	1	0.462	0.495
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.000	0.000
	N	1,310	1,298	1,249	1,237	1,353	1,307	1,311
ODE Mth ProbSolv	Pearson Correlation	0.553	0.480	0.496	0.418	0.462	1	0.451
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000		0.000
	N	1,301	1,285	1,235	1,218	1,307	1,331	1,315
ODE Writing	Pearson Correlation	0.537	0.601	0.513	0.621	0.495	0.451	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	
	N	1,302	1,284	1,238	1,221	1,311	1,315	1,335

**. Correlation is significant at the 0.01 level (2-tailed).

Grade of Enrollment = 08		ODE Math K/S	ODE Read/Lit	NWEA Math	NWEA Reading	ODE Science	ODE Mth ProbSolv	ODE Writing
ODE Math K/S	Pearson Correlation	1	0.727	0.801	0.677	0.711	0.640	0.568
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000
	N	1,552	1,494	1,429	1,344	1,520	1,512	1,516
ODE Read/Lit	Pearson Correlation	0.727	1	0.679	0.800	0.759	0.522	0.651
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000	0.000
	N	1,494	1,536	1,414	1,327	1,504	1,504	1,514
NWEA Math	Pearson Correlation	0.801	0.679	1	0.761	0.692	0.618	0.571
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000	0.000
	N	1,429	1,414	1,492	1,323	1,453	1,423	1,436
NWEA Reading	Pearson Correlation	0.677	0.800	0.761	1	0.746	0.478	0.637
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000	0.000
	N	1,344	1,327	1,323	1,395	1,368	1,328	1,341
ODE Science	Pearson Correlation	0.711	0.759	0.692	0.746	1	0.496	0.541
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.000	0.000
	N	1,520	1,504	1,453	1,368	1,585	1,523	1,531
ODE Mth ProbSolv	Pearson Correlation	0.640	0.522	0.618	0.478	0.496	1	0.450
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000		0.000
	N	1,512	1,504	1,423	1,328	1,523	1,550	1,530
ODE Writing	Pearson Correlation	0.568	0.651	0.571	0.637	0.541	0.450	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	
	N	1,516	1,514	1,436	1,341	1,531	1,530	1,564

**. Correlation is significant at the 0.01 level (2-tailed).

Grade of Enrollment = 10		ODE Math K/S	ODE Read/Lit	NWEA Math	NWEA Reading	ODE Science	ODE Mth ProbSolv	ODE Writing
ODE Math K/S	Pearson Correlation	1	0.695	0.823	0.643	0.753	0.518	0.446
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000
	N	136	132	134	134	131	119	133
ODE Read/Lit	Pearson Correlation	0.695	1	0.720	0.765	0.739	0.522	0.604
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000	0.000
	N	132	132	130	130	128	117	131
NWEA Math	Pearson Correlation	0.823	0.720	1	0.747	0.682	0.520	0.562
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000	0.000
	N	134	130	137	135	132	120	134
NWEA Reading	Pearson Correlation	0.643	0.765	0.747	1	0.683	0.451	0.724
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000	0.000
	N	134	130	135	137	133	121	134
ODE Science	Pearson Correlation	0.753	0.739	0.682	0.683	1	0.538	0.477
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.000	0.000
	N	131	128	132	133	134	119	132
ODE Mth ProbSolv	Pearson Correlation	0.518	0.522	0.520	0.451	0.538	1	0.405
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000		0.000
	N	119	117	120	121	119	122	122
ODE Writing	Pearson Correlation	0.446	0.604	0.562	0.724	0.477	0.405	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	
	N	133	131	134	134	132	122	136

**. Correlation is significant at the 0.01 level (2-tailed).