



NEW MEXICO
Public Education Department

New Mexico High School Standards Assessment
Technical Report
for the Spring 2005 Administration

Prepared for
the New Mexico Public Education Department
by Pearson Educational Measurement



February 2, 2006

TABLE OF CONTENTS

INTRODUCTION AND OVERVIEW	7
CHAPTER 1 Test Administration	9
1.1 OVERVIEW	9
1.2 ADMINISTRATOR TRAINING	9
1.3 HANDLING MATERIALS	9
1.4 TEST SECURITY	10
1.5 STUDENT PARTICIPATION	11
1.6 ACCOMMODATIONS AND MODIFICATIONS	11
CHAPTER 2 Scoring	12
2.1 ITEM TYPES	12
2.2 SCORING PROCESS	13
2.3 QUALITY CONTROL	17
CHAPTER 3 Technical Characteristics of the 2005 NMHSSA Items	19
3.1 OVERVIEW OF TEST STRUCTURE AND CONTENT	19
3.2 TEST STATISTICS	22
3.3 ITEM STATISTICS	22
3.4 SUMMARY FOR MULTIPLE-CHOICE ITEMS	26
3.5 SUMMARY FOR CONSTRUCTED-RESPONSE ITEMS	26
CHAPTER 4 Item Calibration and Vertical Scaling	27
4.1 METHODOLOGY	27
4.2 ABILITY SCORE ESTIMATES	31
4.3 USING NMHSSA SCALE SCORES	31
CHAPTER 5 Standard Setting and Verification Overview	36
5.1 STANDARD SETTING	36
5.2 STANDARD VERIFICATION	42
5.3 EFFECTS OF STANDARD VERIFICATION ON PERFORMANCE DISTRIBUTION	52

CHAPTER 6	Reliability	60
6.1	TOTAL TEST RELIABILITY AND OVERALL SEM	60
6.2	CSEM	61
6.3	STRATIFICATION BY CONTENT	61
6.4	STRATIFICATION BY ITEM TYPE	62
6.5	DECISION CONSISTENCY	63
CHAPTER 7	Validity	65
7.1	EVIDENCE RELATED TO TEST CONTENT	65
7.2	EVIDENCE BASED ON INTERNAL STRUCTURE	69
REFERENCES, RESOURCES, AND RELATED DOCUMENTS		75

List of Tables

Table 2.21 Scoring Rules for Multiple Choice Items	13
Table 2.22 Training Set Compilation and Qualification Standards	14
Table 2.23 Nonscorable Condition Codes	16
Table 2.24 Blank Condition Code Assignment Versus Zero Score Assignment	16
Table 3.11 2005 NMHSSA Test Structure for Reading	19
Table 3.12 2005 NMHSSA Test Structure for Mathematics	20
Table 3.13 Content Strand Coverage for the Reading Test	20
Table 3.14 Content Strand Coverage for the Mathematics Test	21
Table 3.21 Summary of Test Statistics	22
Table 3.31 Item Summary Statistics for the Reading Test	24
Table 3.32 Item Summary Statistics for the Mathematics Test	25
Table 3.41 Summary of Item Statistics for Multiple-Choice Items	26
Table 3.51 Summary of Item Statistics for Short- and Extended Response Items	26
Table 4.11 Configuration of the 2005 NMHSSA Reading	28
Table 4.12 Configuration of the 2005 NMHSSA Mathematics	29
Table 4.13 Scale Transformation between 9 th and 11 th Grades	30
Table 4.21 Raw to Scale Scores Conversion Table for Reading	32
Table 4.22 Raw to Scale Scores Conversion Table for Mathematics	34
Table 4.31 Mean and Standard Error for the 2005 NMHSSA on the Reported Scale	31
Table 5.11 Summary of Schedule and Activities	38
Table 5.12 Panel Recommendations for the Reading Test after Round 3	39
Table 5.13 Panel Recommendations for the Mathematics Test after Round 3	40
Table 5.14 Panel Recommendations for the Mathematics Test after Round 2	41
Table 5.21 Standards Verification Meeting Agenda	43
Table 5.22 Performance Level Distribution for Mathematics	45
Table 5.23 Mathematics Performance Level Distribution by Gender	45
Table 5.24 Mathematics Performance Level Distribution by Ethnicity	46
Table 5.25 Mathematics Performance Level Distribution by County (Selection)	46
Table 5.26 Item Information and Data (Selection)	47
Table 5.27 Performance Level Distribution for Reading	49
Table 5.28 Reading Performance Level Distribution by County (Selection)	51
Table 5.29 Item Information and Data (Selection)	52
Table 5.31 Modified Performance Level Standards	57
Table 5.32 Overall Mathematics Performance Level Distribution Based on the Modified Cut Scores	57
Table 5.33 Overall Reading Performance Level Distribution Based on the Modified Cut Scores	58
Table 5.34 Mathematics Performance Level Distribution by Gender Based on the Modified Cut Scores	58
Table 5.35 Mathematics Performance Level Distribution by Ethnicity Based on the Modified Cut Scores	59
Table 6.11 Reliability Coefficients and Standard Errors of Measurement	61
Table 6.21 Conditional Standard Errors of Measurement at Cut Scores	61

Table 6.31 Coefficient Alpha Stratified by Content Strand: Reading	62
Table 6.32 Coefficient Alpha Stratified by Content Strand: Mathematics	62
Table 6.41 Coefficient Alpha Stratified by Item Type: Reading	62
Table 6.42 Coefficient Alpha Stratified by Item Type: Mathematics	63
Table 6.5 Consistency Indices for Performance Levels for the 2005 NMHSSA	64
Table 7.11 Spring 2005 NMHSSA Reading Item Distribution by Strands	67
Table 7.12 Spring 2005 NMHSSA Mathematics Item Distribution by Strands	68
Table 7.21 Content Strand Correlation Matrix for the 2005 NMHSSA	69
Table 7.22 Factor Loadings of the Five Strand Scores	70
Table 7.23 Total Variance Explained by the Five Components	71
Table 7.24 Factor Loadings of the Five Strand Scores Based on Two-Factor Solution	71
Table 7.25 Differential Item Functioning Summary for the Reading Test	73
Table 7.26 Differential Item Functioning Summary for the Mathematics Test	74

List of Figures

Figure 3.21 2005 NMHSSA Reading Test Raw Score Distribution	23
Figure 3.22 2005 NMHSSA Mathematics Test Raw Score Distribution	23
Figure 5.21 Expected and Actual Performance for 2005 NMHSSA Mathematics	44
Figure 5.22 2005 NMHSSA Mathematics Raw Score Distribution	44
Figure 5.23 Mathematics Performance Distribution for 2004 and 2005	47
Figure 5.24 Expected and Actual Performance for 2005 Reading	48
Figure 5.25 2005 NMHSSA Reading Raw Score Distribution	49
Figure 5.26 Reading Performance Distribution by Gender	50
Figure 5.27 Reading Performance Distribution by Ethnicity	50
Figure 5.28 Reading Performance Distribution for 2004 and 2005	51
Figure 5.31 Effect of Cut Score Change on the Distribution (Math: Beginning Step/Nearing Proficient)	53
Figure 5.32 Effect of Cut Score Change on the Distribution (Math: Nearing Proficient/Proficient)	53
Figure 5.33 Effect of Cut Score Change on the Distribution (Math: Proficient/Advanced)	54
Figure 5.34 Effect of Cut Score Change on the Distribution (Reading: Beginning Step/Nearing Proficient)	54
Figure 5.35 Effect of Cut Score Change on the Distribution (Reading: Nearing Proficient/Proficient)	55
Figure 5.36 Effect of Cut Score Change on the Distribution (Reading: Proficient/Advanced)	55
Figure 5.37 Reading Performance Distribution by Gender Based on the Modified Cut Scores	58
Figure 5.38 Reading Performance Distribution by Ethnicity Based on the Modified Cut Scores	59
Figure 7.23 Scree Plot from the Factor Analysis of the 2005 NMHSSA	71

INTRODUCTION AND OVERVIEW

The New Mexico High School Standards Assessment (NMHSSA) is a state-level criterion-referenced examination aligned with the high-school curriculum standards. The NMHSSA is designed to measure the achievement of public high-school students in Reading and Mathematics. The primary purpose of the assessment is to evaluate the Adequate Yearly Progress (AYP) of schools as part of the federal No Child Left Behind (NCLB) legislation. Potential secondary objectives and outcomes of the assessment include:

- *increasing the academic achievement of students*
- *raising high school graduation rates*
- *implementing rigorous academic standards aligned with the New Mexico curriculum*
- *improving instruction through the implementation of higher standards*
- *using assessment results that measure performance relative to standards*
- *informing various audiences—including teachers, school and district administrators, New Mexico Public Education Department (NMPED) staff, parents, and the public—about the current achievement status of and the progress toward meeting academic achievement standards of New Mexico’s public school students*

The NMHSSA Reading and Mathematics subtests were created over a span of three years commencing in late 2000. In November 2003, the NMHSSA in Reading and Mathematics was administered operationally for the first time, preceded by a full census field-testing event conducted during the spring of 2003. Pearson Educational Measurement (PEM) served as the prime contractor for the administration and scoring of the tests. The Riverside Publishing Company, in partnership with PEM, provided item and test development as well as psychometric services for the program.

Statewide educational standards are the basis of the NMHSSA assessment. The NMHSSA is made up of two subtests; Reading and Mathematics. The New Mexico Reading standards were adopted in June 2000. The Reading subtest measures standards from the New Mexico content standards, and includes two reporting content strands 1-A, 1-C, 1-D, and 3-B: Reading and Listening for Comprehension and Literature and Media. Reading performance standards that were judged to be assessable on a paper and pencil exam made up the content that was assessed by the Reading subtest. The New Mexico Mathematics standards were adopted in June 2002. The Mathematics subtest measures three mathematics content strands 2.1-4, 3.1-4, and 5.1-4: Algebra, Geometry; and Data Analysis and Probability. The *Mathematics Test and Item Specifications* are made up of benchmarks from the three area high school level content area strands. Both the Reading and Mathematics subtests employed three item types - multiple-choice, short response,

and extended response. See the *NMHSSA Reading and Mathematics Test and Item Specifications* for additional information.

This technical report focuses on a description of the operational assessments in Reading and Mathematics and associated technical information. Chapters 1 and 2 provide an overview of the administration and scoring of the 2005 tests. Chapters 3 through 7 document the technical characteristics of the test items, scaling and cut scores, reliability and standard errors of measurement, differential item functioning, and validity evidence for the subtests.

CHAPTER 1

TEST ADMINISTRATION

1.1 OVERVIEW

The 2005 spring operational administration of the New Mexico High School Standards Assessment (NMHSSA) occurred in March 2005. The test administration window was from February 28 to March 11 with numerous state, district, and school personnel participating in the NMHSSA administration. Each school district appointed one district test coordinator. At least one school test coordinator was appointed for each school, and one test administrator was assigned to each classroom. Pearson Educational Measurement (PEM) distributed all required testing materials to each district test coordinator, who in turn distributed the materials as indicated above. In the case of Bureau of Indian Affairs (BIA), charter, and special state supported schools, materials were distributed directly to the schools.

1.2 ADMINISTRATOR TRAINING

District test coordinators (DTCs) were trained through sessions coordinated and conducted by the NMPED. The *NMHSSA Test Administrator Manual: Reading and Mathematics* (NMPED, 2003C) and the *NMHSSA Test Coordinator Manual: Reading and Mathematics* (NMPED, 2003D), which cover general testing information, test security, test material receipt, distribution and return procedures, and guidelines for student participation as well as other topics were distributed to districts and schools and posted on the NMPED website. Districts and schools also received copies of each manual with their test materials.

1.3 HANDLING MATERIALS

Distribution

Test materials were sent to the DTCs in shrink-wrapped packages within boxes that included district and school inventories. All students received a test booklet that was used for both the Reading and Mathematics subtests. Once the materials were accounted for and any missing materials reported to PEM, the DTCs delivered the materials to the appropriate schools. PEM provided procedures for documenting discrepancies and were notified of any missing materials.

School test coordinators were responsible for distributing materials to test administrators. Each secure test book contained a security bar-code, which was used to track each document. Each day of the test administration, secure test materials were signed in and out and school security checklists were used to help keep track of the exchange of materials. Materials distributed each day were limited to those needed for testing on that particular day. When not in use, secure materials were locked in storage.

Returns

When testing was completed, materials were collected and tabulated. In addition, demographic information on the answer documents was edited for correct coding and, if necessary, the “Teacher Use Only” portion of the answer document was completed (e.g., special education programs and/or accommodations). The documents were then packaged together and locked in the storage until they were shipped to PEM.

1.4 TEST SECURITY

Test security was an important issue before, during, and after testing. The specific procedures to be used during the test administration and used in the handling of documentation were outlined in the *NMHSSA Test Administrator Manual* and the *NMHSSA Test Coordinator Manual*.

The manuals provided specific guidelines regarding the storage and handling of test materials, the responsibility of administrators to monitor students during testing, the removal of supplemental materials from the testing room, and the requirement that the administrators refrain from interference with student responses.

Following the test administration and the return of materials to PEM, PEM generated a missing documents report listing the identification numbers of non-returned secure materials. The report was used to notify districts of missing materials. Subsequently, the districts located and returned the materials or sent explanations as to why materials were not found. Toll-free telephone support was provided to answer questions regarding missing documents. After these initial attempts at collecting all missing materials, a list of districts and the materials still unaccounted for was sent to the New Mexico Public Education Department for additional follow-up.

Secure Materials

Secure materials—each assigned human- and machine-readable security identification numbers—included both the Reading and the Mathematics test booklets. Secure materials were locked in storage until the day of the test administration and were signed in and out when in use. These materials were not to be left unattended at any time. Additionally, the NMPED and PEM established security policies requiring secure storage, limited access to items, and secure disposal of shredded documents containing secure information.

1.5 STUDENT PARTICIPATION

All students attending New Mexico public high schools were required to participate in testing for the NMHSSA. The following accommodations and/or modifications were allowed: alternate presentation, alternate response, and alternate schedule. Test booklets were also produced in large print and Braille formats. Federal and state law (i.e., Individuals with Disabilities Education Act of 1997 and the New Mexico Education Accountability Act of 1998) does not exempt any student from participating in the statewide assessment.

Students with Disabilities and Limited English Proficiency

Students with Disabilities were included with appropriate accommodations and/or modifications based upon recommendations by each student's Individualized Education Program (IEP) committee. Students with 504 Accommodation Plans and limited-English-proficient (LEP) students were included in the NMHSSA administration. Note: guidelines for the inclusion of IEP and LEP students were given in the *NMHSSA Test Administrator Manual*. The responses of visually impaired students were transcribed into answer documents by authorized school and district level personnel.

1.6 ACCOMMODATIONS AND MODIFICATIONS

Supplemental information regarding the administration of the NMHSSA to students with disabilities was provided in the *NMHSSA Test Administrator Manual*, which provides guidelines for IEP teams in making decisions regarding testing students with disabilities and outlines specific information regarding testing accommodations, testing modifications, test forms and materials, and administration procedures.

Accommodations

For New Mexico standards-based assessment programs, the term *accommodation* refers to any change in the testing environment, procedures, or presentation that does not alter in any significant way the construct that the test measures. Accordingly, an accommodation has no effect on the comparability of scores. The purpose of accommodations is to enable students to participate in an assessment in a way that allows the knowledge and skills of interest, rather than disabilities, to be assessed. Validity is improved because a source of construct-irrelevant variance is removed. Testing accommodations should be those typically used during routine instruction. Accommodations address areas such as setting, timing, scheduling, alternate response options, and presentation. Besides the specific accommodations listed in the *NMHSSA Test Administrator Manual*, others that are determined necessary by the IEP team can be used.

CHAPTER 2

SCORING

This chapter describes the types of items used on the NMHSSA as well as the scoring procedures. Scoring of items was completed using keys for multiple-choice (MC) items and scoring rubrics for short- and extended-response (SR and ER) items.

2.1 ITEM TYPES

The NMHSSA contained three types of items: multiple choice, short response, and extended response. Additional resources that describe the three item types used on the NMHSSA and their respective scoring are provided in the *NMHSSA Test and Item Specifications Part One: Reading Test Specifications* (2001b) documents. These descriptions are briefly summarized below.

Multiple-Choice Items

These items required students to select a correct answer from four alternatives. Each multiple-choice item was scored as right or wrong and had a value of one (1) point. Missing responses (items that a student did not answer) and multiple responses (double grids) were scored as incorrect (0).

Short-Response Items

These items required a student to respond with a few words or sentences or to demonstrate a process. Both the Reading and the Mathematics test contained some SR items worth two (2) to three (3) points each. Students were given approximately one-half page for each response.

Extended-Response Items

These items require a written response from students. Both the Reading and Mathematics subtests contained some ER items worth four (4) to six (6) points each. The quality of student's response, rather than the length, was the determining factor for the score. Students were given approximately a full page for each response.

Rubrics

The rubrics for all the content areas were initially developed by Pearson Educational Measurement (PEM) and then reviewed by content and fairness committees of New

Mexico Educators prior to field-testing. Prior to scoring, the rubrics were reviewed along with samples of actual student responses by range finding committees. Any changes to the rubrics by the range finding committees were made by PEM’s Scoring Directors and then sent to the New Mexico Public Education Department.

2.2 SCORING PROCESS

PEM was responsible for scoring, analyzing, and reporting on the NMHSSA. Although separate test booklets contained the Reading and Mathematics test items, students responded in a single scannable answer document. Multiple-choice items were scanned and scored electronically through response capture by PEM’s scanning system, with scoring performed at PEM facilities. Student responses to multiple-choice items were scanned in with special coding used for omitted and multiple-marked items. These items were then scored as correct (when the selected option matched the key) or incorrect (all other responses including omits and multiple marks).

Performance Scoring of the Multiple Choice Items

The NMHSSA was administered using separate test booklets for the Reading and Mathematics portions of the test; however, each student recorded their scores in a single scannable answer document. PEM’s response capture system, located in Iowa City, Iowa, facilitated the performance scoring of the multiple choice items on the NMHSSA. Student responses to multiple choice items were scanned into the response capture system. Each response was then electronically compared to the answer key and marked as correct or incorrect. The scoring rules are documented in Table 2.21.

Table 2.21 Scoring Rules for Multiple Choice Items

Status	Rule
Correct	<ul style="list-style-type: none"> • Student’s response matched the answer key exactly
Incorrect	<ul style="list-style-type: none"> • Student’s response does not match the answer key exactly <ul style="list-style-type: none"> ○ Includes omits ○ Includes multiple marks

Performance Scoring of the Constructed Response Items

The NMHSSA constructed response items were comprised of both short response and extended response items. Scoring of these items was carried out at two of PEM’s Performance Scoring Centers (PSC). The Brooklyn Center, Minnesota, PSC conducted the performance scoring for the Reading constructed response items and the Lansing, Michigan, PSC conducted the performance scoring for the Math constructed response items.

Rangefinding and Rubric Review

The goal of rangefinding was to obtain, through consensus scoring, enough scored responses to build strong and effective anchor, practice, and qualifying sets which were to be used for scorer training. Student responses were assembled by PEM's scoring directors and assembled into sets by item. Papers were selected so that a full range of scores for each item would be represented and clearly illustrated. They also contained unusual responses in order for the scoring directors to get direction from the customer on how to score these responses. These sets were then submitted to the NMPED for initial review, followed by a joint review session with PEM via teleconference. During joint review, student responses were evaluated and given a consensus score based on state-approved rubrics. In some instances rubric modification was done during rangefinding to allow maximum benefit of credit to be given to students' responses. Following joint review, PEM assembled final scorer training sets and obtained written approval from the NMPED regarding the assigned scores and scoring decisions that were documented during rangefinding.

Training and Qualification Procedures

Training was administered so that scoring directors trained scoring supervisors first, followed by scorers' training. Delivery of materials was presented in a classroom environment, whereby the scoring director conducted the review and explanation of item-specific materials approved by the NMPED. All scorers were required to qualify before being allowed to score live student responses. Training set compilation and qualification standards are listed in Table 2.22.

Table 2.22 Training Set Compilation and Qualification Standards

Set	Compilation
Anchor Set	10-15 papers, 2-3 examples per score point. Arranged from low to high.
Practice Set One	10-15 papers, 2-3 examples per score point. Arranged randomly.
Practice Set Two	10-15 papers, 2-3 examples per score point. Arranged randomly.
Qualification Set One	10-15 papers arranged randomly. <i>Qualification standard:</i> <ul style="list-style-type: none">• 80% exact agreement with approved rangefinding consensus scores for 0-2 pt items.• 65% exact agreement with approved rangefinding consensus scores for 3-5 pt items.
Qualification Set Two	10-15 papers. Arranged randomly. <i>Qualification standard:</i> <ul style="list-style-type: none">• 80% exact agreement with approved rangefinding consensus scores for 0-2 pt items.

	<ul style="list-style-type: none"> • 65% exact agreement with approved range-finding consensus scores for 3-5 pt items.
--	--

Scoring Procedures

Student responses were scanned into PEM’s image-based (ePEN) system whereby trained and qualified scorers were able to access electronic student response images for scoring. Scorers evaluated student responses and assigned scores (as determined from the scoring guide and criterion which were delivered in training). Images were presented individually and a scorer was not allowed to move on to another image until a score for the current image had been assigned. This process continued until the scorer logged off or until no more student responses were available for scoring.

All student responses were automatically routed for a required first reading. Ten percent of the student responses were randomly selected and routed for a required second reading. The ePEN system generated performance reports that document inter-rater agreement and validity agreement on a daily and cumulative basis in order to be able to monitor scorers’ performance and to help ensure quality scoring.

A process referred to as “backreading” was also utilized during constructed response scoring. In backreading scoring directors and supervisors had the ability to review scorers’ work by being able to view images which had already been assigned a score in the ePEN system. Scoring directors and supervisors backread a sample of responses per scorer to monitor scorer performance. Backreading scores were captured in the ePEN system and used to generate daily and cumulative reports. These reports documented the number of responses read by a scorer, the number of those responses that had gone through backreading, and the agreement percentages between a scorer and scoring director/supervisor. Through this process scoring directors were able to monitor reliability and identify scorers who were in need of retraining.

Student responses that could not be scored were assigned a condition code to indicate the reason why they were not scored. Non-scorable responses were sent to a review queue by scorers. They were then reviewed by a scoring supervisor or scoring director and assigned a non-scorable condition code. These values are defined in Table 2.23.

Table 2.23 Nonscorable Condition Codes

Condition Code	Definition
OT	<ul style="list-style-type: none"> • Answer completely unrelated to the prompt • Direct refusal to participate • Obscenities (and nothing else) • Inappropriate statements, such as “I don’t care” • Symbols, scribbles, or doodles that don’t represent a response • Putting an intentional dash or X in the answer space • Pictures that seem unrelated to the prompt
IL	<ul style="list-style-type: none"> • Not legible • Letters are legible, but not intelligible, e.g., “apten ertkcd” • Words are arranged in such a way that no meaning is conveyed, e.g., “dog cat boy picture”
NE	<ul style="list-style-type: none"> • Language other than English. • If the response is a combination of English and another language, and there is enough written in English to determine an accurate score, go ahead and score it.

The NMPED provided PEM with rules pertaining to the scoring of constructed responses where a condition code of “BL” (blank) should be assigned as opposed to a score of zero. These rules are documented in Table 2.24.

Table 2.24 Blank Condition Code Assignment Versus Zero Score Assignment

Blank Condition Code Assignment	Zero Score Assignment
<ul style="list-style-type: none"> • Blank • Complete erasure • Completely crossed out • Answer space is blank, but the student has circled, underlined, or placed an X on the item number 	<ul style="list-style-type: none"> • Attempting to answer, but the answer is incorrect • Recopying or paraphrasing the prompt (and nothing else) • Statements such as “I don’t know” or “?”(question mark) or “N/A”

Validity Papers

The use of prescored responses strategically interspersed into the pool of live responses to be scored was utilized as a tool for monitoring the accuracy of scoring. It was an objective procedure to help ensure scorers were applying the same standards throughout the scoring process. This was a procedure that offered immediate feedback on individual scorers and the scoring room as a whole in relation to the accuracy and consistency of scoring.

Responses used as validity papers were responses that exemplified solid score points. They were not questionable or “line” responses nor were they unusual responses that would generate questions or discussion. Responses were selected by scoring directors, based on the anchor papers approved by the NMPED, and scanned into the ePEN system’s validity pool. Validity papers were delivered to scorers at a presentation rate of 1:50; meaning that one out of every 50 responses viewed by a scorer was a validity paper. While being viewed by scorers, validity papers were not distinguishable from live student responses.

Validity agreement reports, provided via the ePEN system, were reviewed regularly by scoring directors. The reports reflected if a scorer was scoring accurately, too high or too low. Scorers who did not meet the required validity standards were locked out of the ePEN system until they were retrained. Their performance continued to be monitored until they demonstrated that they were scoring at the required accuracy level.

“Alert” Papers

As a routine part of any scoring project, scorers were instructed to be aware of and to identify student responses that may have been indicative of potentially serious situations. These responses could have been where language was relating to or indicating child abuse, suicide or threat of harm to oneself or another person. Other responses that would have been flagged were indications of cheating, such as more than one student having identical responses. These “alerts” are brought to the attention of the scoring supervisors and forwarded to the NMPED for review and follow up action.

2.3 QUALITY CONTROL

Constructed Responses

Quality Control for Image Scoring: A variety of reports are produced throughout the scoring process to allow scoring supervisory staff to monitor the progress of the project, the reliability of scores assigned and individual scorer’s work. For image-based scoring, these reports were produced by the scoring system and used by scoring supervisors.

NMPED was provided access to the following reports for their ongoing review and oversight of the scoring process. These reports include:

Daily and Cumulative Inter-Rater Reliability Reports by Item and Scorer. These reports provided information about how many times scorers were in exact agreement, assigned adjacent scores or required resolutions. The reliability was computed and monitored daily and cumulatively for the project.

Daily and Cumulative Frequency Distributions. These reports showed how many times each score point was assigned to the item being scored by the reader. They were produced both on a daily basis and cumulatively for the entire scoring project. This report allowed scoring supervisors and subject leaders to see whether scorers had a tendency to score consistently high or low.

PEM reviewed the NMHSSA multiple-choice item performance using traditional item analysis procedures. All items having suspicious statistics as determined by examination by PEM psychometricians were flagged for further review by content specialists to ensure accuracy of keying. For example, an item was flagged if the item appeared difficult with a low percent correct, had a low or negative correlation with total test score, or had an incorrect response with a higher percent selection to provide reliable item performance statistics, and therefore, all the items and keys associated with this form were reviewed directly.

A statistical review does not guarantee that all incorrectly keyed items will be identified or that the identified items are, in fact, incorrectly keyed. The purpose of such a review is to identify items with relatively suspicious statistics. These items may be ineffectively written, have no correct answer, or have multiple correct answers. Such items may also be too difficult to perform correctly in this context. This procedure is intended to support the verification of the scoring keys. It should not be used as the sole replacement for a formal review of the keys by a content expert or any other internal key check procedures.

In this regard, no inaccuracies were found with the keying of the operational forms for the NMHSSA administered in spring 2005.

CHAPTER 3

TECHNICAL CHARACTERISTICS OF THE 2005 NMHSSA OPERATIONAL ITEMS

A total of 20,907 students in spring 2005 participated in the NMHSSA. Item response data for the entire student population were collected. The following cases were eliminated from the data set for further analyses:

- Students who took the Braille form Reading and Mathematics
- Students who did not attempt all three sections of a particular subtest
- Students who did not attempt at least 10 items in each subtest (Reading and Mathematics)

The exemption rules above led to a new analysis data file for mathematics comprised of 19,843 valid cases and a new data file for Reading comprised of 19,816 valid cases. All results in this section are based on the analyses of these new data files. The item and test statistics presented in this chapter were generated using SAS 8.2 and Winsteps 3.32.

3.1 OVERVIEW OF TEST STRUCTURE AND CONTENT

The basic structure for both the Reading and Mathematics subtests is presented in Tables 3.11 (Reading) and 3.12 (Mathematics). Note that both subtests include regular operational items and 20 linking items used for vertical scaling. Linking items were embedded in the tests. Scores for the linking items were not counted and reported toward the total student scores. As noted previously, the 2005 NMHSSA assessments are comprised of multiple-choice (MC), short- and extended-response (SR and ER) items. This is true for both operational and linking sets. The score point distribution across item types is shown below.

Table 3.11 2005 NMHSSA Test Structure for Reading

Test Set	Item Type	Number of Items	Number of Points	Percentage of Total Points
Operational	MC	46	46	65.8
	SR	6	12	17.1
	ER	3	12	17.1
	Total	55	70	100.0
Linking	MC	14	14	46.6
	SR	4	8	26.7
	ER	2	8	26.7
	Total	20	30	100.0

Table 3.12 2005 NMHSSA Test Structure for Mathematics

Test Set	Item Type	Number of Items	Number of Points	Percentage of Total Points
Operational	MC	39	39	43.4
	SR	8	20	22.2
	ER	7	31	34.4
	Total	54	90	100.0
Linking	MC	14	14	46.6
	SR	4	8	26.7
	ER	2	8	26.7
	Total	20	30	100.0

An overview of the 2005 test content for Reading and Mathematics is shown in Tables 3.13 (Reading) and 3.14 (Mathematics). Please see Chapter 7 for more detailed information regarding content coverage for both subtests.

Table 3.13 Content Strand Coverage for the Reading Test

Content Strands		Content Standards	MC	SR	ER
1	Reading and Listening for Comprehension	A. Read, react to, and analyze information C. Critical thinking to evaluate information and solve problems D. Evaluate print, non-print, and technology-based information	29	5	3
3	Literature and Media	B. Understand literary elements, concepts, and genres	17	1	0
Total			46	6	3

Note: MC = Multiple Choice; SR = Short Response; ER = Extended Response

Table 3.14 Content Strand Coverage for the Mathematics Test

Content Strands		Content Standards	MC	SR	ER
2	Algebra	1. Understand patterns, relations, functions, and graphs 2. Represent and analyze mathematical situations and structures using algebraic symbols 3. Use mathematical models to represent and understand quantitative relationships 4. Analyze changes in various contexts	8	5	2
3	Geometry	1. Analyze characteristics and properties two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships 2. Specify locations and describe spatial relationships using coordinate geometry and other representational systems 3. Apply transformations and use symmetry to analyze mathematical situations 4. Use visualization, spatial reasoning, and geometric modeling to solve problems	15	1	3
5	Data Analysis and Probability	1. Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them 2. Select and use appropriate statistical methods to analyze data	16	2	2

		3. Develop and evaluate inferences and predictions that are based on data 4. Understand and apply basic concepts of probability			
Total			39	8	7

Note: MC = Multiple Choice; SR = Short Response; ER = Extended Response

3.2 TEST STATISTICS

Test statistics for Reading and Mathematics are presented in Table 3.21. The raw score distributions for the two subjects are shown in Figure 3.21 and Figure 3.22. The mean raw score for Reading is associated with a percent correct score of 62.7, while the mean raw score for Mathematics represents 38.4 percent correct. Overall, the Reading subtest was relatively easier than the Mathematics subtest. Mathematics scores also showed a relatively larger variance than Reading scores did. Descriptive statistics based on the scale scores are presented in the next chapter.

Table 3.21 Summary of Test Statistics

	Raw Score Mean	Standard Deviation	Maximum	Minimum
Reading (N=19,816)	43.86	13.40	69	1
Mathematics (N=19,843)	34.59	19.19	90	0

3.3 ITEM STATISTICS

Analyses were also performed at an item level. Item mean scores were generated based on the entire population as well as individual ability subgroups. Item-total (point-biserial) correlation was used as an index of discrimination power. Tables 3.31 and 3.32 show all the item descriptive statistics for the operational items.

Figure 3.21 2005 NMHSSA Reading Test Raw Score Distribution

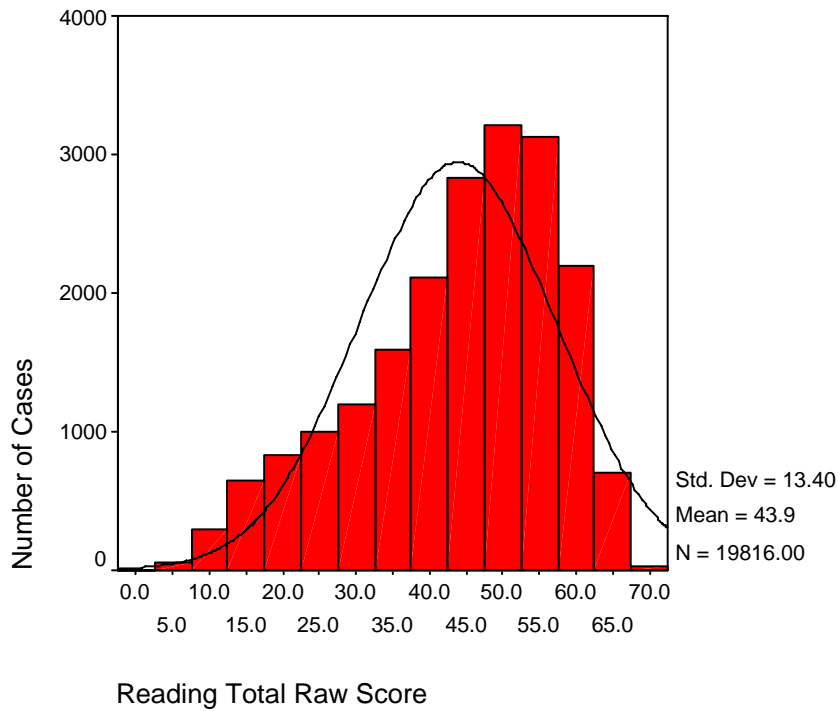


Figure 3.22 2005 NMHSSA Mathematics Test Raw Score Distribution

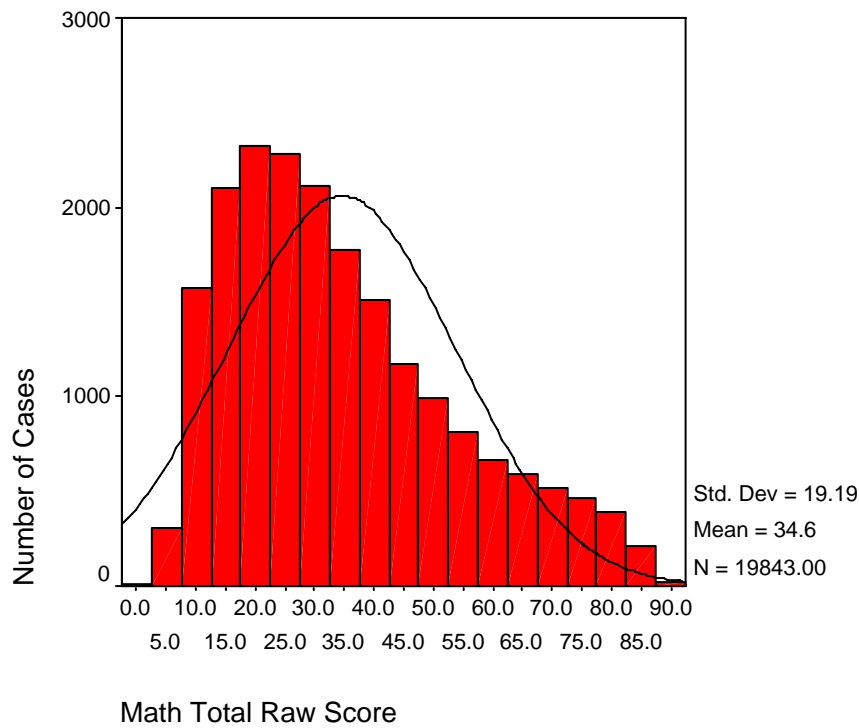


Table 3.31 Item Summary Statistics for the Reading Test

Item Position	Domain/ Strand	Objective/ Benchmark	Item Type	Pts	Item Average Score	Item-Total Point-Biserial	Item Mean Advanced	Item Mean Proficient	Item Mean Nearing Proficient	Item Mean Beginning Step
1	3	B	MC	1	0.5	0.31	0.87	0.56	0.35	0.29
2	3	B	MC	1	0.73	0.38	0.94	0.82	0.64	0.35
3	3	B	MC	1	0.76	0.41	0.98	0.86	0.65	0.39
4	1	C	MC	1	0.4	0.18	0.66	0.42	0.3	0.28
5	1	C	MC	1	0.83	0.42	0.99	0.93	0.77	0.46
6	1	C	MC	1	0.81	0.4	0.96	0.91	0.75	0.43
7	1	C	MC	1	0.89	0.47	1	0.98	0.87	0.48
8	3	B	MC	1	0.78	0.33	0.96	0.86	0.71	0.48
9	1	C	MC	1	0.81	0.65	0.95	0.89	0.75	0.46
10	1	C	MC	1	0.59	0.31	0.84	0.68	0.47	0.3
11	1	C	MC	1	0.49	0.29	0.81	0.55	0.34	0.27
12	1	C	ER	4	2.03	0.55	3.14	2.39	1.55	0.71
13	1	A	MC	1	0.68	0.39	0.95	0.78	0.56	0.34
14	1	D	MC	1	0.69	0.25	0.88	0.76	0.62	0.45
15	1	D	MC	1	0.52	0.27	0.82	0.57	0.39	0.32
16	1	D	SR	2	0.56	0.43	1.21	0.67	0.31	0.13
17	1	D	SR	2	1.52	0.47	1.92	1.75	1.36	0.65
18	1	A	MC	1	0.46	0.48	0.93	0.58	0.23	0.13
19	1	A	MC	1	0.8	0.48	0.99	0.93	0.71	0.36
20	1	A	MC	1	0.79	0.45	0.98	0.91	0.69	0.37
21	1	D	SR	2	0.92	0.42	1.54	1.09	0.65	0.28
22	1	A	SR	2	0.97	0.54	1.51	1.17	0.73	0.3
31	1	C	MC	1	0.58	0.22	0.74	0.64	0.51	0.33
32	1	C	MC	1	0.42	0.18	0.71	0.42	0.33	0.29
33	1	C	MC	1	0.48	0.27	0.79	0.54	0.36	0.27
34	3	B	MC	1	0.41	0.21	0.65	0.44	0.33	0.22
35	1	C	MC	1	0.59	0.5	0.97	0.75	0.35	0.22
36	1	D	MC	1	0.74	0.46	0.98	0.87	0.6	0.35
37	1	A	MC	1	0.67	0.44	0.96	0.8	0.5	0.3
38	1	D	MC	1	0.35	0.28	0.75	0.38	0.2	0.22
39	1	D	MC	1	0.54	0.39	0.86	0.64	0.38	0.19
40	1	D	MC	1	0.72	0.55	0.99	0.89	0.53	0.25
41	1	C	MC	1	0.72	0.49	0.97	0.87	0.56	0.29
42	1	C	ER	4	1.65	0.64	3.32	2.13	0.8	0.19
55	1	C	MC	1	0.73	0.47	0.96	0.87	0.6	0.27
56	1	C	MC	1	0.89	0.47	0.99	0.98	0.88	0.49
57	1	C	MC	1	0.83	0.48	0.99	0.95	0.75	0.38
58	1	C	MC	1	0.78	0.51	0.99	0.92	0.67	0.29
59	1	C	MC	1	0.88	0.53	1	0.99	0.87	0.39
60	1	C	MC	1	0.86	0.56	0.99	0.98	0.82	0.32
61	3	B	MC	1	0.77	0.45	0.97	0.89	0.67	0.33
62	3	B	MC	1	0.81	0.53	0.99	0.95	0.72	0.3
63	3	B	MC	1	0.63	0.46	0.92	0.77	0.48	0.17
64	3	B	MC	1	0.64	0.36	0.88	0.74	0.55	0.27
65	3	B	MC	1	0.75	0.48	0.98	0.89	0.64	0.3
66	3	B	MC	1	0.84	0.56	1	0.98	0.77	0.31
67	3	B	MC	1	0.82	0.55	1	0.96	0.73	0.3
68	3	B	MC	1	0.74	0.32	0.9	0.82	0.7	0.37
69	3	B	MC	1	0.85	0.55	1	0.97	0.81	0.32
70	3	B	MC	1	0.73	0.43	0.95	0.85	0.62	0.31
71	3	B	MC	1	0.83	0.57	1	0.97	0.77	0.28
72	3	B	MC	1	0.81	0.55	0.99	0.95	0.73	0.26
73	3	B	SR	2	1.27	0.58	1.75	1.47	1.11	0.45
74	1	C	SR	2	1.17	0.64	1.87	1.52	0.77	0.14
75	1	C	ER	4	1.86	0.63	2.94	2.17	1.47	0.53

Item positions not listed in this table were the linking items.

Table 3.32 Item Summary Statistics for the Mathematics Test

Item Position	Domain/ Strand	Objective/ Benchmark	Item Type	Pts	Item Average Score	Item-Total Point-Bi-serial	Item Mean Advanced	Item Mean Proficient	Item Mean Nearing Proficient	Item Mean Beginning Step
1	1	1	MC	1	0.84	0.29	0.98	0.95	0.87	0.63
2	2	3	MC	1	0.56	0.43	0.94	0.8	0.54	0.25
3	3	3	MC	1	0.76	0.4	0.98	0.95	0.78	0.44
4	2	4	MC	1	0.28	0.45	0.83	0.5	0.18	0.11
5	1	2	MC	1	0.44	0.5	0.95	0.74	0.33	0.19
6	3	3	MC	1	0.44	0.27	0.84	0.57	0.37	0.32
7	1	4	MC	1	0.58	0.4	0.87	0.82	0.57	0.25
8	2	1	MC	1	0.59	0.39	0.95	0.82	0.56	0.32
9	3	2	MC	1	0.41	0.23	0.72	0.52	0.36	0.3
10	3	2	MC	1	0.57	0.25	0.83	0.69	0.56	0.38
11	2	4	MC	1	0.41	0.35	0.88	0.58	0.33	0.25
12	2	1	MC	1	0.39	0.43	0.87	0.61	0.32	0.14
13	3	1	MC	1	0.75	0.28	0.93	0.88	0.77	0.52
14	1	3	ER	4	1.06	0.69	3.63	2.28	0.55	0.06
15	2	2	SR	3	0.77	0.55	2.03	1.33	0.62	0.11
16	2	2	ER	4	0.69	0.61	3.49	1.4	0.25	0.04
17	3	2	ER	4	2.03	0.63	3.46	2.91	2.02	0.71
18	3	3	SR	2	0.75	0.56	1.64	1.32	0.65	0.11
19	3	2	MC	1	0.46	0.25	0.73	0.61	0.42	0.31
20	2	4	MC	1	0.42	0.21	0.68	0.53	0.39	0.28
22	3	1	MC	1	0.62	0.4	0.95	0.82	0.62	0.3
23	2	3	MC	1	0.28	0.44	0.81	0.5	0.18	0.11
24	3	3	MC	1	0.61	0.24	0.93	0.71	0.59	0.45
25	1	2	MC	1	0.33	0.23	0.72	0.42	0.27	0.24
26	2	2	MC	1	0.36	0.47	0.91	0.61	0.26	0.14
27	2	1	MC	1	0.5	0.44	0.96	0.74	0.44	0.23
28	3	3	MC	1	0.44	0.48	0.93	0.72	0.37	0.17
30	1	4	MC	1	0.26	0.32	0.69	0.38	0.18	0.14
32	3	2	MC	1	0.43	0.44	0.93	0.69	0.33	0.22
34	2	1	MC	1	0.34	0.23	0.7	0.44	0.27	0.26
35	3	4	MC	1	0.59	0.54	0.98	0.9	0.55	0.2
38	1	4	MC	1	0.56	0.48	0.98	0.85	0.51	0.25
39	1	2	SR	3	2.11	0.52	2.91	2.78	2.32	0.74
40	1	3	SR	3	1.85	0.5	2.84	2.6	1.99	0.46
41	1	1	SR	2	0.27	0.59	1.69	0.57	0.05	0
42	1	1	SR	2	0.8	0.56	1.91	1.36	0.7	0.07
43	1	1	SR	2	0.62	0.59	1.82	1.15	0.45	0.03
45	2	4	ER	4	0.69	0.65	3.45	1.52	0.19	0.01
46	2	4	MC	1	0.61	0.48	0.97	0.87	0.59	0.24
47	3	4	MC	1	0.64	0.36	0.91	0.84	0.62	0.37
49	1	2	MC	1	0.34	0.16	0.68	0.38	0.3	0.29
50	3	3	MC	1	0.47	0.46	0.96	0.75	0.37	0.24
51	2	4	MC	1	0.26	0.37	0.76	0.42	0.16	0.15
52	2	1	MC	1	0.36	0.57	0.94	0.69	0.24	0.1
54	3	1	MC	1	0.35	0.26	0.75	0.43	0.32	0.21
59	1	2	MC	1	0.48	0.49	0.96	0.78	0.39	0.22
62	3	3	MC	1	0.54	0.28	0.86	0.69	0.51	0.35
63	2	2	MC	1	0.39	0.05	0.6	0.36	0.41	0.35
64	2	3	MC	1	0.4	0.38	0.88	0.59	0.32	0.21
65	3	1	MC	1	0.71	0.45	0.99	0.93	0.72	0.36
69	3	4	ER	5	1.92	0.64	4.83	3.78	1.4	0.15
70	3	4	SR	3	0.48	0.61	2.5	1	0.15	0.03
72	2	1	ER	6	0.81	0.68	4.63	1.58	0.25	0.02
74	1	4	ER	4	0.95	0.74	3.17	1.95	0.56	0.05

Item positions not listed in this table were the linking items.

3.4 SUMMARY FOR MULTIPLE-CHOICE ITEMS

Table 3.41 provides a summary of the major indices reported in Tables 3.31 and 3.32 for the multiple-choice items. The average item difficulty (mean p-value, proportion correct) was higher for Reading than Mathematics. In other words, the Reading MC items were generally easier for students than the Mathematics MC items. The mean items-total (point-biserial) correlation, an index of item discrimination, for the MC items is also presented in Table 3.41. Discrimination indices measure item performance with respect to whether the item is difficult for low ability students and easier for high ability students. The value ranges from -1.0 to +1.0. A negative point-biserial would indicate that an item was difficult for high ability students and easier for low ability students, an undesirable condition for achievement test items.

Table 3.41 Summary of Item Statistics for Multiple-Choice Items

Test	Number of Items	Mean	
		p-value	Point-biserial
Reading	46	.69	.42
Mathematics	39	.48	.36

3.5 SUMMARY FOR CONSTRUCTED-RESPONSE ITEMS

Table 3.51 provides a summary of the major indices reported in Tables 3.31 and 3.32 for the SR and ER items. Item means were converted to pseudo p-values by dividing the item mean by the maximum possible score for the item type. The pseudo p-values put the SR and ER items on the same scale as the MC items, allowing the relative difficulty of each item type to be compared. As in the analysis of the MC items, Reading SR and ER items were found to be much easier than those on the mathematics test. Also, the average pseudo p-values for the SR and ER items are generally lower than the p-values for the MC items. However, discrimination power for the SR and ER items seems to be higher than the MC items.

Table 3.51 Summary of Item Statistics for Short- and Extended- Response Items

Test	Item Type	Number of Items	Item Mean	Mean Pseudo p-value	Item-Total Correlation
Reading	SR	6	1.07	.53	.51
	ER	3	1.85	.46	.61
Mathematics	SR	8	.96	.37	.56
	ER	7	1.16	.27	.66

CHAPTER 4

ITEM CALIBRATION AND VERTICAL SCALING

This chapter provides an outline of the procedures used for calibrating and scaling the operational subsets from the March 2005 administration of the New Mexico High School Standard Assessment. Major results are also presented.

Item calibration is the process of assigning a difficulty estimate to each item on an assessment so that they are placed onto a common scale. The purpose of a scaling analysis is to create a score scale. Scaling is used to transform test score values onto a scale that can be more easily interpreted by users. The other purpose of the vertical scale is to evaluate the AYP of schools as part of the federal NCLB legislation. Specific to the case of high school achievement, we are interested in knowing if 11th graders are doing better academically than 9th graders in the State of New Mexico. Since different tests were administered to the two grade levels, the scores of the two grades can not be directly compared to show the progress. One convention is to establish a vertical scale where scores generated from various tests are placed upon a common scale so that they can be properly compared. This chapter presents the methods applied to construct the vertical scale and how scores from the 2005 NMHSSA relate to scores from the 9th grade assessments.

For the NMHSSA, the resulting scale scores will be used for score reporting and performance level classification. The NMHSSA classifies students into four achievement levels: Beginning Step, Nearing Proficient, Proficient, and Advanced.

4.1 METHODOLOGY

Data

As described in the beginning of the previous chapter, the data for the calibration and scaling analyses were the item responses to the 2005 NMHSSA Reading and Mathematics for grades 9 and 11. Same data exemption rules were applied to the two grades analyzed.

The 11th grade Mathematics subtest was designed to assess three major content strands: Algebra, Geometry, and Data Analysis and Probability. The 11th grade Reading subtest measured two major content strands: Reading and Listening for Comprehension and Literature and Media.

Test Construction –Common Item Design

To link the performance between 9th and 11th grades and to construct a vertical scale, a set of 20 items were selected and embedded in the operational test for both Mathematics and Reading subtests in both grades. The source of the common items (Harcourt Publication), which was also responsible for the 9th grade test construction, was not able to provide PEM with scoring materials for 6 open-ended items (3 mathematics and 3 reading items). Without common scoring rubrics and training procedures, those 6 items were no longer considered appropriate as linking items. Thus the number of common items was reduced to 17 for both Mathematics and Reading subtests. The 11th grade test contained an operational test that is specific to the grade level and 17 common items that also appeared in the 9th grade test. The 17 common items were selected to be representative of the overall test in terms of standards representation, range of difficulty, and when possible, item format. The configuration of the reading and mathematics operational tests and the 17 common items are shown in Table 4.11 and 4.12 below. As illustrated in the tables, both multiple-choice and open-ended items were used as common items. To minimize systematic error due to context effect, the 17 common items were placed in approximately the same positions in the 9th and 11th grade tests. There were a total of 71 active items (54+17) in the Mathematics subtest and 72 active items (55+17) in the Reading subtest.

Table 4.11 Configuration of the 2005 NMHSSA Reading

Operational Test				Common Items			
Content	Item Format		Subtotal	Content	Item Format		Subtotal
Strand	Multiple-Choice	Open-Ended		Strand	Multiple-Choice	Open-Ended	
1	29	8	37 (67%)	1	9	2	11 (65%)
3	17	1	18 (33%)	3	5	1	6 (35%)
Subtotal	46 (84%)	9 (16%)	Total=55 (100%)	Subtotal	14 (82%)	3 (18%)	Total=17 (100%)

Table 4.12 Configuration of the 2005 NMHSSA Mathematics

Operational Test				Common Items			
Content	Item Format		Subtotal	Content	Item Format		Subtotal
Strand	Multiple-Choice	Open-Ended		Strand	Multiple-Choice	Open-Ended	
2	8	7	15 (28%)	2	3	1	4 (24%)
3	15	4	19 (35%)	3	5	1	6 (35%)
5	16	4	20 (37%)	5	6	1	7 (41%)
Subtotal	39 (72%)	15 (28%)	Total=54 (100%)	Subtotal	14 (82%)	3 (18%)	Total=17 (100%)

Item Calibration

Two IRT models were used to calibrate the test items. The Rasch model was used to estimate parameters for the MC items. Master’s Partial Credit Model (Masters, 1982) was used to estimate parameters for the constructed-response items. These models parameterize the steps needed to reach a complete solution to each item so that there is a difficulty parameter for each possible item score after zero. MC items can be considered as partial-credit items with only one step. Item parameter estimation was performed through Winsteps 3.32.

Kim and Cohen (1998) found that “when the number of common items is small (less than 20), linking of separate calibration runs may be preferable to concurrent calibration.” In addition, item parameter estimates can be compared from one grade to another to identify items that are behaving differently in different grades with separate calibration. For these reasons, item calibration was conducted separately for grades 9 and 11.

Scale Transformation

The separate calibrations produced a unique metric for each grade, which was not comparable to each other. Therefore after the separate item calibration runs, a grade-by-grade chained linking was performed to place all item parameter estimates for the two grades on a common scale. A common scale was achieved through the use of the common items. The scale for 9th grade was identified as the base scale. The two sets of item parameter estimates for the items in common between the 9th and 11th grades were used to estimate a scale transformation that placed the item parameter estimates from the 11th grade onto the scale of the 9th grade.

The scale transformation constant was generated by finding the difference between the average b parameters (difficulty) of the common items for the two grades. As shown in Table 4.13, for the Reading test, the common item “b” value average for the 11th grade is -0.0234 and the average for the 9th grade is 1.6508, with a difference of 1.6742. For the Mathematics test, the average “b” value for the 11th grade is -0.4748, and the average for the 9th grade is 3.0315. The difference is 3.5062. The difference of the average was used as the transformation constant. By adding the constant to the originally calibrated “b” value based on the 11th grade data, the item difficulty (b) estimates for the 11th grade operational items were then transformed onto the 9th grade scale. These new item difficulty parameters were fixed and applied to estimate θ , or individual ability. The final step transforms the θ values to a reportable scale. The transformation equation is consistent with the elementary and secondary grades.

Table 4.13 Scale Transformation between 9th and 11th Grades

Reading			Mathematics		
Linking Item Position	Estimate of Difficulty (b) for 11 TH Grade	Estimate of Difficulty (b) for 9 TH Grade	Linking Item Position	Estimate of Difficulty (b) for 11 TH Grade	Estimate of Difficulty (b) for 9 TH Grade
23	0.3339	2.1983	29	1.1372	4.8493
24	0.0068	1.7413	31	-0.9432	2.9108
26	-0.2610	1.1323	33	-0.0934	3.2036
28	-0.6922	1.1781	36	0.1346	3.6178
29	-0.8123	1.2432	37	-1.1449	2.1239
30	0.5762	2.2971	44	-0.7438	2.7802
43	-1.0360	0.6304	48	0.6797	4.3566
44	-0.6627	1.0994	53	-1.5392	2.0167
46	-0.9102	0.8070	56	-1.0859	2.7966
47	0.8780	1.9582	57	-1.1594	1.9353
48	-1.3463	0.2140	58	-0.2401	3.6174
49	0.6157	2.5529	60	1.2032	4.5942
50	1.4345	3.1538	61	-0.5338	2.8131
51	1.0805	2.7327	66	-0.8759	3.0533
52	-0.0218	1.5243	67	-0.8444	2.5833
53	-1.5155	0.0917	68	-1.2730	2.2279
54	1.9352	3.5094	73	-0.7487	2.0549
Average	-0.0234	1.6508	Average	-0.4748	3.0315
Difference	1.6742		Difference	3.5062	

4.2 ABILITY SCORE ESTIMATES

After parameter transformation, ability estimates were generated separately with the item difficulty parameters fixed using Winsteps 3.32. Only operational items that contributed to student scores were included in the analysis. Winsteps was also used to derive the raw score to Rasch ability (θ) score estimates as well as the standard errors of those estimates. Theta estimates have properties that might be difficult for some users to understand (e.g., negative ability scores). Consequently, these scores were converted to a new score scale using a linear transformation function. The scale scores were calculated based on the equation below.

$$\text{Scale Score} = 35 * \theta + 600$$

Table 4.21 shows the scale score conversions for Reading while Table 4.22 shows the scale score conversions for Mathematics. Also included in the tables are the information of the scores, percentile, and the performance level based on the standard setting results (see Chapter 5).

4.3 USING NMHSSA SCALE SCORES

The 2005 NMHSSA achievement data are reported at the student level on the Individual Student Report. The student report contains raw score, scale score, and performance level achieved. The mean and standard errors on the newly transformed scale are presented below in Table 4.31.

Table 4.31 Mean and Standard Error for the 2005 NMHSSA on the Reported Scale

Content	Mean	Standard Error
Reading	688	75
Mathematics	725	60

The NMHSSA scale scores should be interpreted only within each content area. That is, scale scores are not status indicators, like a percentile rank would be, and therefore cannot be used to profile strengths and weaknesses. Scale scores that are based on IRT models are typically assumed to be of the interval type. It is therefore appropriate for comparisons to be made on differences in scale scores. For example, it is safe to infer that the ability difference between 500 and 520 represents the same ability difference that separates 700 and 720. Also, because vertical scale between 9th grade and 11th grade was established, comparisons or growth statements between these two grade levels are possible.

Table 4.21 Raw to Scale Scores Conversion Table for Reading

Raw Score	Theta	S.E. of Theta	Information	Scale Score	S.E. of Scale Score	Percentile	Performance Level
0	-3.9675	1.833	0.2975	461	64	0	Beginning Step
1	-2.7441	1.013	0.9729	504	35	1	Beginning Step
2	-2.0231	0.726	1.8941	529	25	1	Beginning Step
3	-1.5895	0.601	2.7669	544	21	1	Beginning Step
4	-1.2737	0.527	3.594	555	18	1	Beginning Step
5	-1.0223	0.477	4.3783	564	17	1	Beginning Step
6	-0.8115	0.441	5.1229	572	15	1	Beginning Step
7	-0.6287	0.414	5.8297	578	14	1	Beginning Step
8	-0.4664	0.392	6.5011	584	14	1	Beginning Step
9	-0.3197	0.374	7.1389	589	13	1	Beginning Step
10	-0.1854	0.359	7.7448	594	13	1	Beginning Step
11	-0.0609	0.346	8.3209	598	12	1	Beginning Step
12	0.0555	0.335	8.8678	602	12	2	Beginning Step
13	0.1651	0.326	9.387	606	11	2	Beginning Step
14	0.2689	0.318	9.8799	609	11	3	Beginning Step
15	0.3678	0.31	10.3468	613	11	3	Beginning Step
16	0.4624	0.304	10.7888	616	11	4	Beginning Step
17	0.5533	0.298	11.2054	619	10	5	Beginning Step
18	0.641	0.293	11.5982	622	10	5	Beginning Step
19	0.7259	0.289	11.9662	625	10	6	Beginning Step
20	0.8083	0.285	12.3103	628	10	7	Beginning Step
21	0.8884	0.281	12.6301	631	10	8	Beginning Step
22	0.9667	0.278	12.9262	634	10	9	Beginning Step
23	1.0432	0.275	13.1976	637	10	10	Beginning Step
24	1.1183	0.272	13.445	639	10	11	Beginning Step
25	1.192	0.27	13.6677	642	9	12	Beginning Step
26	1.2647	0.268	13.8667	644	9	13	Nearing Proficient
27	1.3363	0.266	14.0415	647	9	14	Nearing Proficient
28	1.4071	0.265	14.1922	649	9	15	Nearing Proficient
29	1.4773	0.264	14.3199	652	9	16	Nearing Proficient
30	1.5469	0.263	14.4242	654	9	17	Nearing Proficient
31	1.616	0.262	14.5058	657	9	18	Nearing Proficient
32	1.6847	0.262	14.5653	659	9	20	Nearing Proficient
33	1.7533	0.261	14.6032	661	9	21	Nearing Proficient
34	1.8218	0.261	14.62	664	9	22	Nearing Proficient
35	1.8902	0.261	14.6162	666	9	24	Nearing Proficient
36	1.9586	0.261	14.5923	669	9	26	Nearing Proficient
37	2.0272	0.262	14.5487	671	9	27	Nearing Proficient
38	2.0961	0.262	14.4855	673	9	29	Nearing Proficient
39	2.1654	0.263	14.4031	676	9	31	Nearing Proficient
40	2.2351	0.264	14.3012	678	9	33	Nearing Proficient
41	2.3052	0.265	14.1802	681	9	36	Nearing Proficient
42	2.3761	0.266	14.0391	683	9	38	Nearing Proficient
43	2.4477	0.268	13.8782	686	9	40	Nearing Proficient
44	2.5203	0.27	13.6965	688	9	43	Proficient
45	2.5938	0.272	13.494	691	10	46	Proficient
46	2.6685	0.274	13.2692	693	10	49	Proficient
47	2.7446	0.277	13.0224	696	10	52	Proficient
48	2.8222	0.28	12.7528	699	10	55	Proficient

Raw Score	Theta	S.E. of Theta	Information	Scale Score	S.E. of Scale Score	Percentile	Performance Level
49	2.9015	0.283	12.4601	702	10	58	Proficient
50	2.9828	0.287	12.1433	704	10	61	Proficient
51	3.0663	0.291	11.8031	707	10	64	Proficient
52	3.1523	0.295	11.4391	710	10	68	Proficient
53	3.2412	0.3	11.0515	713	11	71	Proficient
54	3.3334	0.306	10.6403	717	11	74	Proficient
55	3.4293	0.313	10.2061	720	11	78	Proficient
56	3.5296	0.32	9.7485	724	11	81	Proficient
57	3.6348	0.328	9.2684	727	11	84	Proficient
58	3.7457	0.337	8.7647	731	12	87	Proficient
59	3.8634	0.348	8.2368	735	12	89	Advanced
60	3.989	0.36	7.6836	740	13	92	Advanced
61	4.1242	0.375	7.1026	744	13	94	Advanced
62	4.2714	0.392	6.4907	749	14	96	Advanced
63	4.4337	0.413	5.8435	755	14	97	Advanced
64	4.6156	0.44	5.1564	762	15	98	Advanced
65	4.8246	0.475	4.4243	769	17	99	Advanced
66	5.0731	0.524	3.6422	778	18	99	Advanced
67	5.3845	0.596	2.8071	788	21	99	Advanced
68	5.8119	0.721	1.9189	803	25	99	Advanced
69	6.5253	1.009	0.981	828	35	99	Advanced
70	7.7425	1.83	0.2984	871	64	100	Advanced

Table 4.22 Raw to Scale Scores Conversion Table for Mathematics

Raw Score	Theta	S.E. of Theta	Information	Scale Score	S.E. of Scale Score	Percentile	Performance Level
0	-1.7812	1.835	0.2968	538	64	1	Beginning Step
1	-0.5528	1.016	0.9674	581	36	1	Beginning Step
2	0.1725	0.728	1.8844	606	25	1	Beginning Step
3	0.6072	0.6	2.7692	621	21	1	Beginning Step
4	0.9213	0.524	3.6353	632	18	1	Beginning Step
5	1.1683	0.471	4.4917	641	16	1	Beginning Step
6	1.3722	0.432	5.3441	648	15	1	Beginning Step
7	1.5458	0.401	6.1935	654	14	1	Beginning Step
8	1.6971	0.376	7.0385	659	13	2	Beginning Step
9	1.8313	0.356	7.8755	664	12	3	Beginning Step
10	1.9521	0.339	8.6999	668	12	5	Beginning Step
11	2.062	0.324	9.504	672	11	6	Beginning Step
12	2.1631	0.3111	0.2853	676	11	8	Beginning Step
13	2.2569	0.3011	1.0369	679	11	11	Beginning Step
14	2.3447	0.2911	1.7554	682	10	13	Beginning Step
15	2.4273	0.2831	2.4404	685	10	15	Beginning Step
16	2.5057	0.2761	3.0917	688	10	17	Beginning Step
17	2.5803	0.2701	3.7087	690	9	19	Beginning Step
18	2.6517	0.2641	4.2943	693	9	21	Beginning Step
19	2.7203	0.2591	4.853	695	9	24	Nearing Proficient
20	2.7865	0.2541	5.3885	698	9	26	Nearing Proficient
21	2.8504	0.2501	5.9067	700	9	28	Nearing Proficient
22	2.9122	0.2461	6.4132	702	9	31	Nearing Proficient
23	2.9723	0.2431	6.9164	704	9	33	Nearing Proficient
24	3.0305	0.2391	7.4221	706	8	35	Nearing Proficient
25	3.0871	0.2361	7.9354	708	8	38	Nearing Proficient
26	3.142	0.2321	8.4628	710	8	40	Nearing Proficient
27	3.1954	0.2291	9.0074	712	8	42	Nearing Proficient
28	3.2473	0.2261	9.5696	714	8	45	Nearing Proficient
29	3.2976	0.2222	0.1489	715	8	47	Nearing Proficient
30	3.3465	0.2192	0.7409	717	8	49	Nearing Proficient
31	3.3941	0.2162	1.3396	719	8	51	Nearing Proficient
32	3.4403	0.2132	1.937	720	7	53	Nearing Proficient
33	3.4853	0.2102	2.5218	722	7	55	Nearing Proficient
34	3.5291	0.2082	3.0859	724	7	57	Nearing Proficient
35	3.572	0.2052	3.6181	725	7	59	Nearing Proficient
36	3.6139	0.2032	4.1125	726	7	60	Nearing Proficient
37	3.6549	0.2012	4.5619	728	7	62	Nearing Proficient
38	3.6953	0.2002	4.9614	729	7	64	Nearing Proficient
39	3.7351	0.1982	5.3117	731	7	65	Nearing Proficient
40	3.7743	0.1972	5.6121	732	7	67	Nearing Proficient
41	3.8132	0.1962	5.8665	733	7	68	Nearing Proficient
42	3.8517	0.1952	6.0801	735	7	70	Proficient
43	3.8899	0.1952	6.2587	736	7	71	Proficient
44	3.9279	0.1942	6.4092	737	7	73	Proficient
45	3.9656	0.1942	6.5375	739	7	74	Proficient
46	4.0032	0.1932	6.6499	740	7	75	Proficient
47	4.0407	0.1932	6.7512	741	7	76	Proficient
48	4.078	0.1932	6.8444	743	7	77	Proficient
49	4.1152	0.1922	6.9317	744	7	78	Proficient
50	4.1523	0.1922	7.013	745	7	79	Proficient
51	4.1893	0.1922	7.0874	747	7	80	Proficient
52	4.2261	0.1912	7.1522	748	7	81	Proficient
53	4.2629	0.1912	7.2037	749	7	82	Proficient
54	4.2996	0.1912	7.2377	750	7	83	Proficient

Raw Score	Theta	S.E. of Theta	Information	Scale Score	S.E. of Scale Score	Percentile	Performance Level
55	4.3364	0.1912	7.249	752	7	83	Proficient
56	4.373	0.1912	7.2325	753	7	84	Proficient
57	4.4098	0.1912	7.1828	754	7	85	Proficient
58	4.4467	0.1922	7.0949	756	7	86	Proficient
59	4.4836	0.1922	6.9636	757	7	87	Proficient
60	4.5209	0.1932	6.7845	758	7	87	Proficient
61	4.5583	0.1942	6.5537	760	7	88	Proficient
62	4.5962	0.1952	6.2679	761	7	89	Proficient
63	4.6345	0.1962	5.9241	762	7	89	Proficient
64	4.6734	0.1982	5.5189	764	7	90	Proficient
65	4.7129	0.1992	5.0514	765	7	90	Proficient
66	4.7533	0.2012	4.5198	766	7	91	Proficient
67	4.7945	0.2042	3.9223	768	7	92	Proficient
68	4.8369	0.2072	3.2568	769	7	92	Proficient
69	4.8806	0.2102	2.5243	771	7	93	Proficient
70	4.9258	0.2142	1.7245	772	7	93	Advanced
71	4.9728	0.2192	0.8569	774	8	94	Advanced
72	5.0218	0.2241	9.924	776	8	94	Advanced
73	5.0733	0.2291	8.9246	778	8	95	Advanced
74	5.1277	0.2361	7.8637	779	8	95	Advanced
75	5.1854	0.2441	6.7497	781	9	96	Advanced
76	5.2473	0.2531	5.5844	784	9	96	Advanced
77	5.314	0.2631	4.382	786	9	97	Advanced
78	5.3867	0.2751	3.1496	789	10	97	Advanced
79	5.4667	0.2891	1.904	791	10	97	Advanced
80	5.5554	0.3061	0.6582	794	11	98	Advanced
81	5.6551	0.325	9.4284	798	11	98	Advanced
82	5.7686	0.348	8.2272	802	12	99	Advanced
83	5.8997	0.376	7.0663	806	13	99	Advanced
84	6.0538	0.409	5.9524	812	14	99	Advanced
85	6.2391	0.452	4.8874	818	16	99	Advanced
86	6.4687	0.508	3.8692	826	18	99	Advanced
87	6.7669	0.588	2.8883	837	21	99	Advanced
88	7.1878	0.719	1.9304	852	25	99	Advanced
89	7.9019	1.011	0.9766	877	35	99	Advanced
90	9.1243	1.833	0.2974	919	64	99	Advanced

CHAPTER 5

STANDARD SETTING AND VERIFICATION OVERVIEW

5.1 STANDARD SETTING

Introduction

Committees of New Mexico educators were convened on February 22-23, 2005, in order to set standards on the grade 11 Reading and Mathematics assessments. A total of 35 educators participated in the 2-day conference. The Modified Angoff procedure was applied to set standards. The outcomes of the conference are described in this summary and more detailed information will be provided in the subsequent technical report.

Panelist Information

Of the 35 panelists, 15 from Reading and 17 from Mathematics participated in the entire conference.¹ All 35 educators provided voluntary demographic information.

Of the 15 panelists in the reading group, 11 reported their ethnicity as Anglo, Caucasian, or White, 3 reported as Hispanic, and 1 as Navajo. Fourteen of the panelists were female. Years of teaching ranged from 1 to 38, with 23 as the median. All of them reported high school teaching experience in English or Language Arts. Educators represented a variety of school sizes and regions.

Of the 20 panelists in the mathematics group, 13 reported their ethnicity as Anglo, Caucasian, or White, and 7 reported as Hispanic. Eleven of the panelists were male. Years of teaching ranged from 9 to 34, with 21 as the median. The current faculty position was mostly High School Math Teacher (18), but also included two university instructors. All panelists reported direct teaching experience in the high school curriculum. As in the Reading group, educators represented a variety of school sizes and regions.

Method and Procedure

Prior to beginning the standard setting activities, the New Mexico Public Education Department (NMPED) and Pearson Educational Measurement (PEM) staff briefed the committees on the purpose of the conference and use of the conference outcomes.

¹ Three panelists in the mathematics group did not participate in the final round of judgments.

Specifically, panelists were advised that the outcomes of the conference were a set of cutscore recommendations to the NMPED. The panelists were informed that the educator committees were an important part of the activities in the complete policy setting procedure of standard setting. That is, their final cutscore recommendations would be considered by the NMPED along with other relevant information as it made final standard-setting decisions.

A modification of the Angoff procedure was the judgmental process used. In this procedure, panelists are asked to predict the percentage of examinees at the borderline of each level of achievement who will respond correctly to an item. The Angoff procedure was modified in two ways. First, panelists were asked to provide their predictions to the nearest 5 percent (e.g., they should report their percentage estimates as 35 or 50, rather than 33 or 52). Second, for polytomous items panelists provided ratings for each score point. In the case of the NMHSSA, polytomous items are scored with variable numbers of score points. Short response questions can have 3 to 4 score points (0 to 2 or 0 to 3), while extended response items can have up to 7 score points (0 to 6). The sum of the ratings for any particular polytomous item must be 100.

The cutscore at each achievement level is determined by summing the predictions across items and dividing the total by 100. This represents the minimum raw score that an examinee must attain to be classified at the particular level. Cuts are usually computed to be between raw scores. In the final report, all cutscores will be rounded to the next higher point if the decimal value is larger than 4 (e.g., 15.5 would become 16).

A summary of the conference schedule is provided in Table 5.11. On the first day, panelists were asked to operationally define each level of achievement using the standards and test specifications established by the NMPED, their experience with students, and the scope and sequence of curriculum in high school. The operational definitions of the performance levels were intended to help panelists consider the knowledge and skills involved at each level, to work toward (though not necessarily obtain) consensus, and then to describe the content knowledge and skills most likely to be needed by students at the borderline of adjacent levels. The operational definitions also provided a context for panelists to determine specific primary differences in knowledge and skills required between adjacent levels of achievement. On Day Two, panelists used the level descriptions in their application of the Modified Angoff procedure.

The panelists worked effectively with one another in both large and small content groups. The conference was collegial and engaging. There were many instances in which panelists engaged in meaningful debate and were able to come to a working agreement on critical tasks, such as the achievement level descriptors on Day One. Consensus was not required for any activity: panelists were instructed to consider the discussions around each activity as a broader context within which they were to make their individual ratings.

When applying the judgmental process, panelists were directed to work independently. Between each round, panelists were asked to discuss their understanding of the content and the rationale behind their decisions. A panelist from each small group was asked to serve as “Table Leader” for purposes of keeping the group on schedule and maintaining task focus.

Table 5.11 Summary of Schedule and Activities

	Day 1 (2/22)	Day 2 (2/23)
Morning	Check in Introductions & Agenda Getting to know the HSSA	Large group training of the Modified Angoff method Round 1 Standard Setting
Lunch	Table leaders training	
Afternoon	Generate performance level (PL) descriptors Discussion and Modification of PL descriptors	Round 2 Standard Setting Round 3 Standard Setting

At the beginning of Rounds 2 and 3 panelists were provided with results generated from the previous round to inform their decision making. In Round 2, panelists were informed of their individual cutscores and how they compared to the cutscores of other panelists in their small group. At the beginning of Round 3, panelists were provided the updated information based on Round 2 results. The means of all panelists were used to describe the content group cutscores.

Results

The Reading subtest is comprised of 55 items with a total possible score of 70. Cut scores after the Round 3 final rating for the Reading test are summarized in Table 5.12.

For the Mathematics subtest, the results are summarized in two ways. The Mathematics test has 54 items with a total possible score of 90. Table 5.13 shows the cut scores based on 17 panelists' ratings after Round 3 since three panelists did not complete the Round 3 final rating. Table 5.14 shows the cut scores based on all 20 panelists' ratings after Round 2. Because only minor changes were made after Round 2, and because only 3 panelists dropped out of the process, the cut scores based on Round 3 ratings are close to those based on Round 2 ratings.

Table 5.12 Panel Recommendations for the Reading Test after Round 3

Table		Round 3 Cutscores		
		Level		
		Nearing Proficient	Proficient	Advanced
1	Max	29.9	45.6	60.9
	Min	16.7	37.0	49.4
	Mean	23.9	40.3	55.4
	Median	24.1	38.7	54.8
	Std	5.1	3.7	4.8
2	Max	40.0	65.9	69.3
	Min	19.9	35.9	50.0
	Mean	27.6	46.9	58.6
	Median	25.5	44.1	59.9
	Std	7.8	11.2	7.3
3	Max	35.1	54.0	63.8
	Min	22.4	39.6	50.6
	Mean	29.1	46.2	59.0
	Median	31.0	45.7	60.7
	Std	5.7	5.8	5.1
Total	Max	40.0	65.9	69.3
	Min	16.7	35.9	49.4
	Mean	26.9	44.4	57.7
	Median	25.5	43.6	59.6
	Std	6.3	7.7	5.7

Table 5.13 Panel Recommendations for the Mathematics Test after Round 3

Table		Round 3 Cutscores		
		Level		
		Nearing Proficient	Proficient	Advanced
1	Max	17.8	46.9	86.4
	Min	10.3	41.3	68.4
	Mean	14.0	44.4	75.7
	Median	14.1	44.9	74.1
	Std	2.7	2.0	7.4
2	Max	20.9	44.7	73.2
	Min	17.8	43.2	71.9
	Mean	19.7	44.0	72.4
	Median	20.5	44.1	72.2
	Std	1.7	0.8	0.7
3	Max	48.3	61.7	74.5
	Min	15.2	44.5	64.5
	Mean	30.2	51.6	69.6
	Median	28.7	50.1	69.7
	Std	13.7	7.2	4.2
4	Max	24.8	61.7	81.9
	Min	11.5	52.9	74.4
	Mean	20.4	56.2	77.9
	Median	23.2	56.1	78.0
	Std	5.6	3.5	2.7
Total	Max	48.3	61.7	86.4
	Min	10.3	41.3	64.5
	Mean	20.7	49.5	74.3
	Median	18.5	46.9	74.1
	Std	9.0	6.5	5.5

Table 5.14 Panel Recommendations for the Mathematics Test after Round 2

Table		Round 2 Cutscores		
		Level		
		Nearing Proficient	Proficient	Advanced
1	Max	17.6	46.6	86.4
	Min	10.3	41.3	68.4
	Mean	14.1	44.4	75.6
	Median	14.1	44.5	72.9
	Std	2.7	2.0	7.7
2	Max	24.1	47.8	76.4
	Min	16.5	39.9	66.2
	Mean	20.0	44.0	71.2
	Median	21.3	45.2	71.4
	Std	3.2	3.1	3.8
3	Max	49.4	61.7	74.5
	Min	13.9	44.5	64.5
	Mean	30.0	50.9	69.7
	Median	30.1	49.8	70.0
	Std	12.7	6.4	3.6
4	Max	24.9	62.7	81.9
	Min	11.4	51.8	74.4
	Mean	20.6	56.2	77.9
	Median	23.9	56.1	78.0
	Std	5.7	4.3	2.7
Total	Max	49.4	62.7	86.4
	Min	10.3	39.9	64.5
	Mean	21.2	48.9	73.6
	Median	19.9	47.2	72.7
	Std	8.9	6.5	5.6

5.2 STANDARD VERIFICATION

Introduction

Panels of New Mexico educators were convened on June 2, 2005, in order to verify the performance standards of the grade 11 Reading and Mathematics assessments that comprise the New Mexico High School Standards Assessment (NMHSSA). These standards had been established previously on February 22-23, 2005. Sixteen out of the 35 educators from the February 2005 standard-setting panels participated in the 1-day standards verification conference. The impact data from the spring 2005 administration of the NMHSSA were presented to the Committee and the standards were reexamined. The outcomes of the conference are described in this summary and more detailed information will be provided in the subsequent technical report.

Purposes of the Conference

The major purpose of the conference was to provide the Standard Verification Panels and the New Mexico Public Education Department (NMPED) the opportunity to review the impact data that were not available during the initial standard setting. By reviewing the impact data, the Committee and the NMPED were able to examine the actual student performance and its difference from their expectation. In addition, they were able to take into account the effect of standards (or cut scores) on the performance distribution in determining the final recommendation of the standards for the 2005 NMHSSA.

Participants

Invitations to the Standards Verification Conference were sent to all of the 35 panelists (15 from Reading and 20 from Mathematics) who participated in the February standard setting. Seven panelists from the Reading group and 9 from the Mathematics group responded and attended the June 2 standard verification conference.

Of the 7 panelists in the reading group, 6 reported their ethnicity as Anglo, Caucasian, or White, 1 reported as Hispanic. Five of the panelists were female.

Of the 9 panelists in the mathematics group, 5 reported their ethnicity as Anglo, Caucasian, or White, and 4 reported as Hispanic. Eight of the panelists were female.

Conference Agenda

Standards verification meetings were held separately for the two subjects. The Mathematics panel convened in the morning and the Reading panel in the afternoon. The meeting agenda is shown in Table 5.21 below.

Table 5.21 Standards Verification Meeting Agenda

Time		Section	Activities
Math	Reading		
8:30 – 9:00	1:00 – 1:30	Introduction	<ul style="list-style-type: none">• Agenda Overview• Purpose of Meeting
9:00 – 9:30	1:30 – 2:00	Review Procedures	<ul style="list-style-type: none">• Review of Performance Level Descriptors• Review of Impact Data
9:30 – 10:30	2:00 – 3:00	Standard Verification	Rating by Performance Levels
10:30 – 10:50	3:00 – 3:20	Break	
10:50 – 11:50	3:20 – 4:20	Conclusion	Evaluation of Cut Scores

Impact Data for the Mathematics Assessment

During the meeting of the mathematics group, both test-level and item-level impact data were presented to the panelists.

The test level impact data included:

- Raw score mean based on the entire 11th grade student population (Figure 5.21)
- Overall performance distribution: proportion of students in each performance level based on previously set cut scores (Figure 5.22, and Table 5.22)
- Performance distribution by gender, ethnicity, and geographical location (Table 5.23-Table 5.25)
- Performance distribution from last year (Figure 5.23)

Figure 5.21 Expected and Actual Performance for 2005 NMHSSA Mathematics

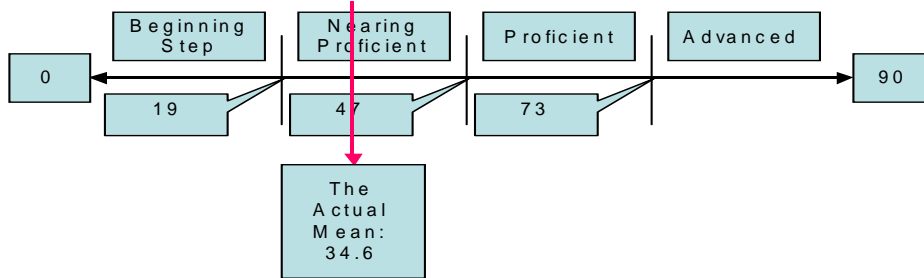


Figure 5.22 2005 NMHSSA Mathematics Raw Score Distribution

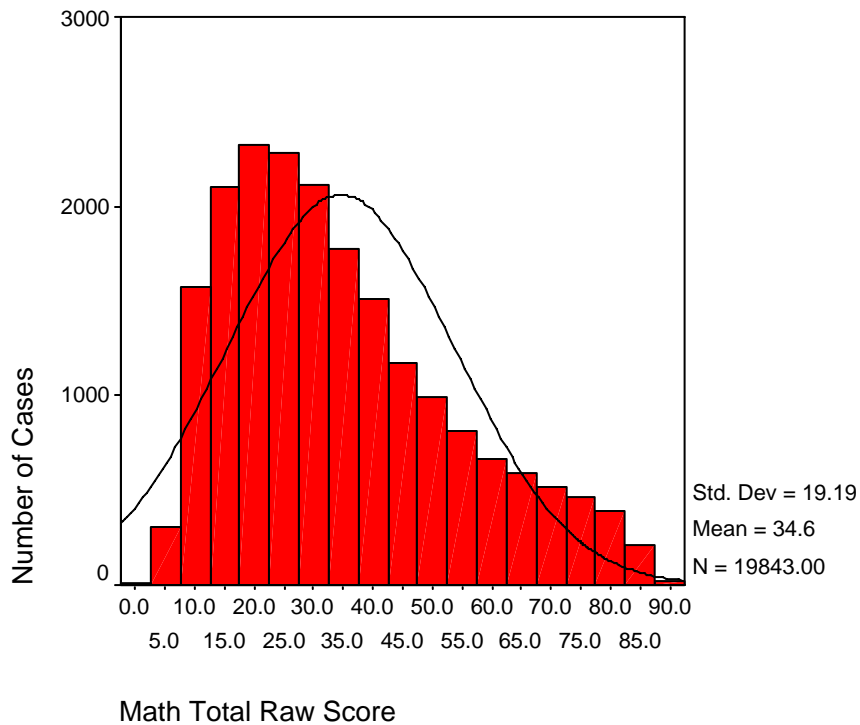


Table 5.22 Performance Level Distribution for Mathematics

	Frequency	Percent	Cumulative Percent
Beginning Step	4459	22.5	22.5
Nearing Proficient	10492	52.9	75.3
Proficient	3801	19.2	94.5
Advanced	1091	5.5	100.0
Total	19843	100.0	

4892 out of 19843 (24.7%) students reached the level of "Proficient" or above.

Table 5.23 Mathematics Performance Level Distribution by Gender

	Gender			
	Female		Male	
	Count	%	Count	%
Beginning Step	2006	20.5%	2453	24.4%
Nearing Proficient	5448	55.7%	5044	50.1%
Proficient	1840	18.8%	1961	19.5%
Advanced	483	4.9%	608	6.0%

Similar distributions were observed for male and female.

Table 5.24 Mathematics Performance Distribution by Ethnicity

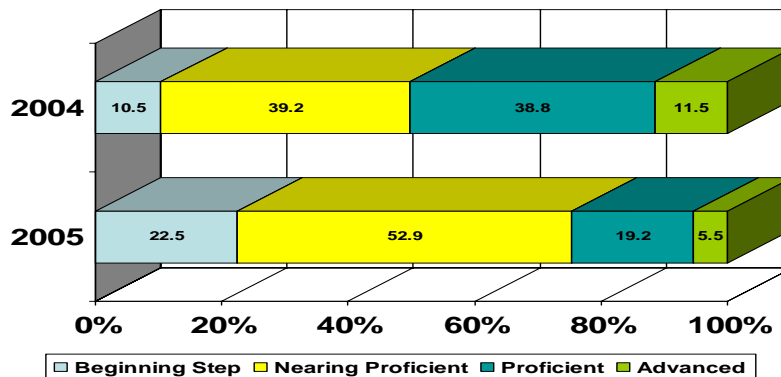
	Ethnicity									
	Asian/Pacific Islander		Black		White		Hispanic		American Indian/Alaskan	
	Count	%	Count	%	Count	%	Count	%	Count	%
Beginning Step	24	8.4%	150	30.9%	884	12.2%	2639	29.0%	762	27.8%
Nearing Proficient	111	38.9%	270	55.7%	3323	45.9%	5102	56.1%	1686	61.5%
Proficient	85	29.8%	53	10.9%	2228	30.8%	1173	12.9%	262	9.6%
Advanced	65	22.8%	12	2.5%	802	11.1%	182	2.0%	30	1.1%

Table 5.25 Mathematics Performance Distribution by County (Selection)

County Name		Performance Category				Total
		Beginning Step	Nearing Proficient	Proficient	Advanced	
BERNALILLO	Count	1081	2315	1182	484	5062
	% within County Name	21.4%	45.7%	23.4%	9.6%	100.0%
CATRON	Count	8	17	10	0	35
	% within County Name	22.9%	48.6%	28.6%	.0%	100.0%
CHAVES	Count	112	314	122	35	583
	% within County Name	19.2%	53.9%	20.9%	6.0%	100.0%
CIBOLA	Count	67	189	36	8	300
	% within County Name	22.3%	63.0%	12.0%	2.7%	100.0%
COLFAX	Count	23	78	39	9	149
	% within County Name	15.4%	52.3%	26.2%	6.0%	100.0%

Figure 5.23 Mathematics Performance Distribution for 2004 and 2005

Current year distribution for Math is significantly different from last year's distribution.



The item-level impact data included (Table 5.26):

- Basic item information: item position, key, type, and weights
- Item descriptive statistics: item mean score, percentage correct²
- Omit rate: percentage of students did not attempt the item

Table 5.26 Item Information and Data (Selection)

Item Position	Operational or Vertical Linking Study Item	Key	Item Type	Points	Item Average Score	Percent Correct	Omit Rate
9	Operational	B	MC ³	1	0.39	39	4.4
10	Operational	D	MC	1	0.55	55	4.5
11	Operational	A	MC	1	0.39	39	5
12	Operational	D	MC	1	0.37	37	4.5
13	Operational	C	MC	1	0.72	72	4.5
14	Operational	NA	OE ⁴	4	1.06	27	22.3
15	Operational	NA	OE	3	0.77	26	14.6
16	Operational	NA	OE	4	0.69	17	23.9
17	Operational	NA	OE	4	2.03	51	11.5
18	Operational	NA	OE	2	0.75	38	11.6

² Defined as percentage of students received full or partial score on the item.

³ Multiple-choice item

⁴ Open-ended item

Impact Data for the Reading Assessment

During the meeting of the reading group, both test-level and item-level impact data were presented to the panelists.

The test-level impact data included:

- Raw score mean based on the entire 11th grade student population (Figure 5.24)
- Overall performance distribution: proportion of students in each performance level based on previously set cut scores (Figure 5.25, and Figure 5.26)
- Performance distribution by gender, ethnicity, and geographical location (Figure 5.27, Figure 5.28, and Table 5.27)
- Performance distribution from last year (Figure 5.29)

Figure 5.24 Expected and Actual Performance for 2005 Reading

Reading Test Level Data: Comparing Expected and Actual Performance

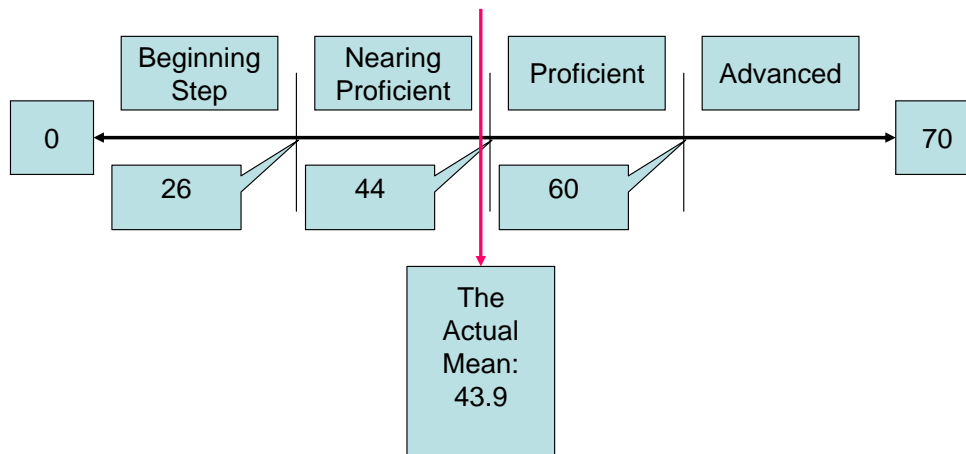


Figure 5.25 2005 NMHSSA Reading Raw Score Distribution

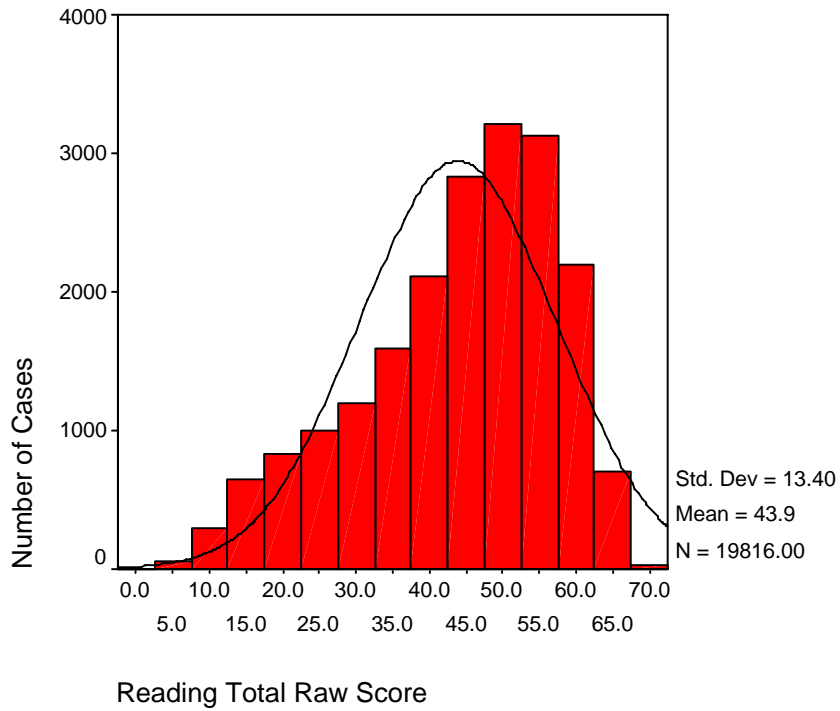
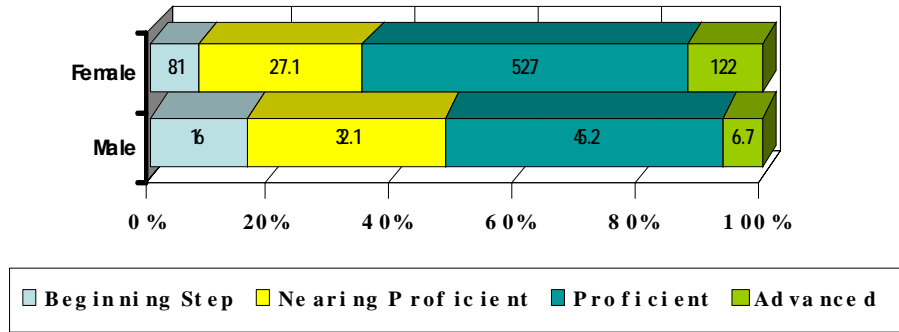


Table 5.27 Performance Level Distribution for Reading

	Frequency	Percent	Cumulative Percent
Beginning Step	2403	12.1	12.1
Nearing Proficient	5869	29.6	41.7
Proficient	9686	48.9	90.6
Advanced	1858	9.4	100.0
Total	19816	100.0	

11544 out of 19816 (58.2%) students reached the level of "Proficient" or above.

Figure 5.26 Reading Performance Distribution by Gender



Female students performed better than male students in general.

Figure 5.27 Reading Performance Distribution by Ethnicity

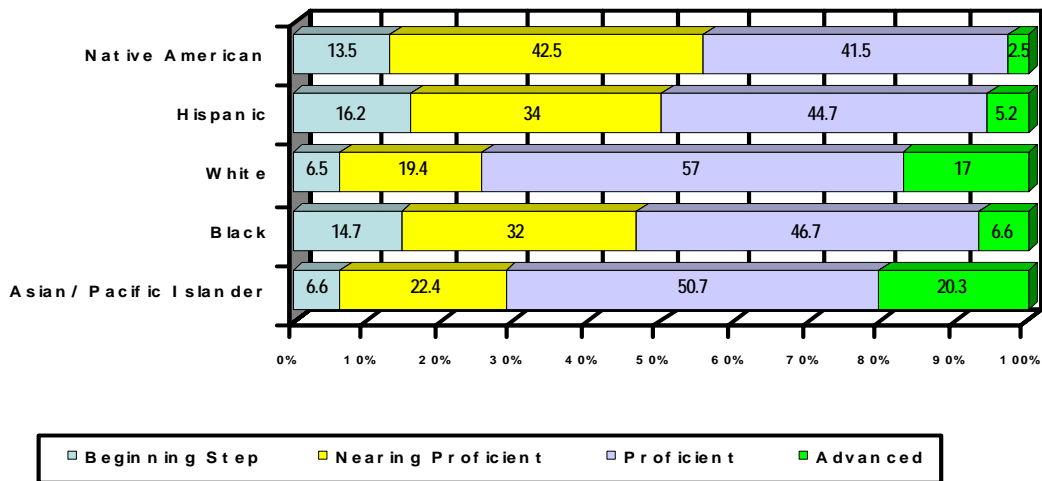
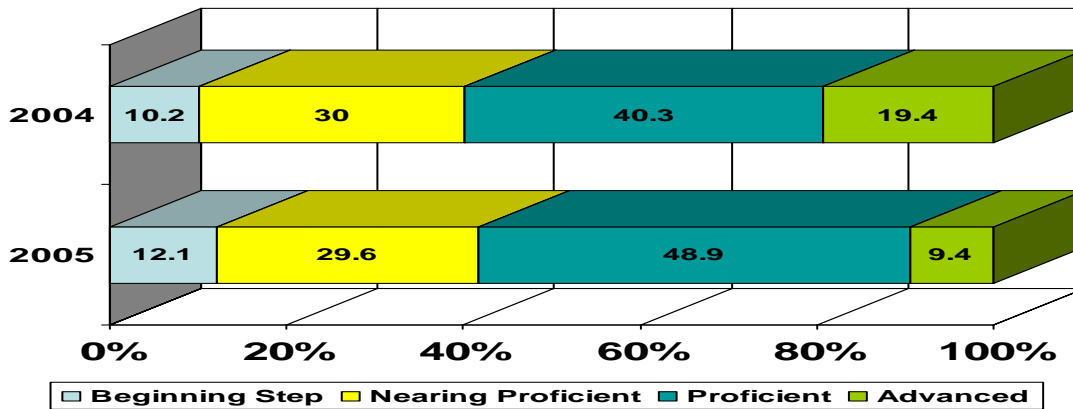


Table 5.28 Reading Performance Distribution by County (Selection)

			Performance Category				Total
			Beginning Step	Nearing Proficient	Proficient	Advanced	
County Name	BERNALILLO	Count	645	1251	2507	652	5055
		% within County Name	12.8%	24.7%	49.6%	12.9%	100.0%
	CATRON	Count	6	11	16	2	35
		% within County Name	17.1%	31.4%	45.7%	5.7%	100.0%
	CHAVES	Count	50	178	287	62	577
		% within County Name	8.7%	30.8%	49.7%	10.7%	100.0%
	CIBOLA	Count	25	74	147	19	265
		% within County Name	9.4%	27.9%	55.5%	7.2%	100.0%
	COLFAX	Count	7	46	81	15	149
		% within County Name	4.7%	30.9%	54.4%	10.1%	100.0%

Figure 5.28 Reading Performance Distribution for 2004 and 2005

Current year distribution for Reading is somewhat deviated from last year's distribution.



The item-level impact data included (Table 5.28):

- Basic item information: item position, key, type, and weights
- Item descriptive statistics: item mean score, percentage correct
- Omit rate: percentage of students did not attempt the item

Table 5.29 Item Information and Data (Selection)

Item Position	Operational or Vertical Linking Study Item	Key	Item Type	Points	Item Average Score	Percent Correct	Omit Rate
13	Operational	D	MC ⁵	1	0.68	68	4.3
14	Operational	C	MC	1	0.69	69	4.3
15	Operational	C	MC	1	0.52	52	4.4
16	Operational	NA	OE ⁶	2	0.56	28	12.7
17	Operational	NA	OE	2	1.52	76	10.4
18	Operational	D	MC	1	0.46	46	4.7
19	Operational	A	MC	1	0.8	80	4.7
20	Operational	C	MC	1	0.79	79	4.8
21	Operational	NA	OE	2	0.92	46	8.6
22	Operational	NA	OE	2	0.97	49	10.5

5.3 EFFECTS OF STANDARD VERIFICATION ON PERFORMANCE DISTRIBUTION

In addition to the test and item level impact data, information regarding the effect of standards verification was also provided to the panelists. With such information, panelists were able to examine the change of performance distribution influenced by standards (or cut scores) verification.

Figure 5.31, 5.32, and 5.33 show how changing the cut scores of the **Mathematics** test can affect the performance distribution. Figure 5.34, 5.35, and 5.36 show how changing the cut scores of the **Reading** test can affect the performance distribution.

⁵ Multiple-choice item

⁶ Open-ended item

Figure 5.31 Effect of Cut Score Change on the Distribution (Math: Beginning Step/Nearing Proficient)

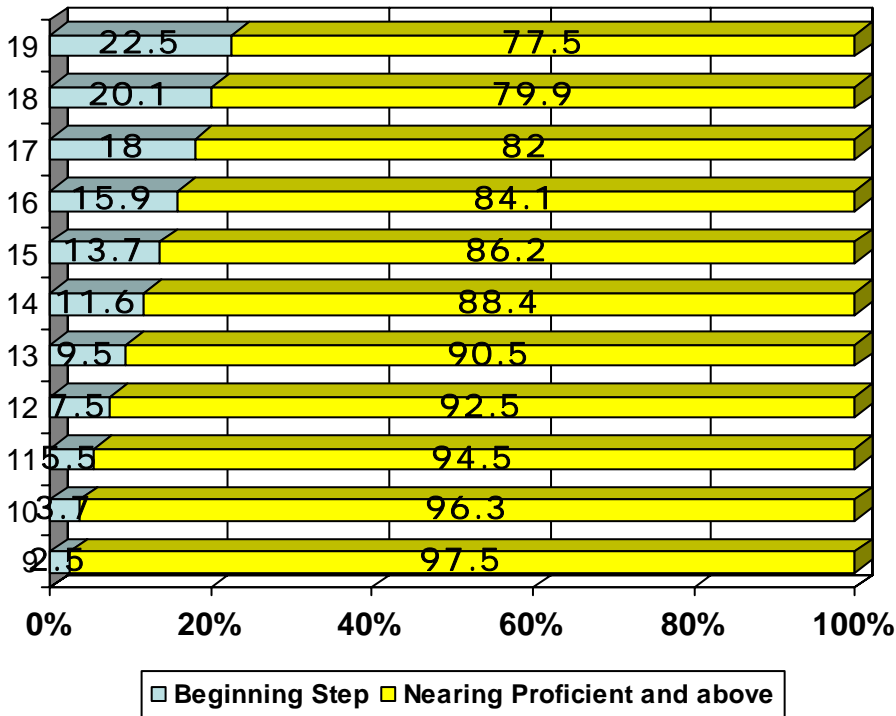


Figure 5.32 Effect of Cut Score Change on the Distribution (Math: Nearing Proficient/Proficient)

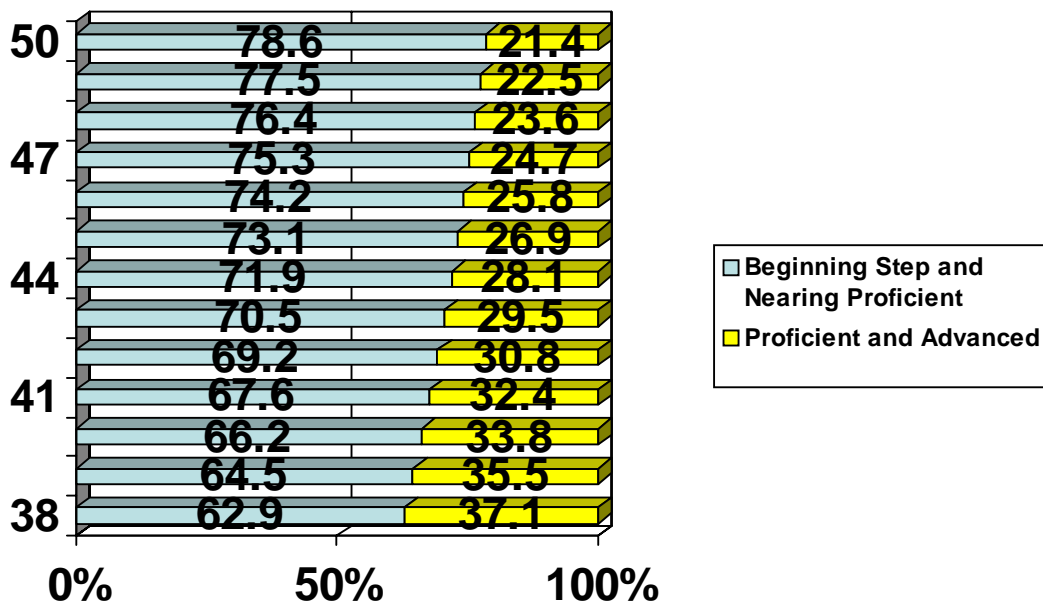


Figure 5.33 Effect of Cut Score Change on the Distribution (Math: Proficient/Advanced)

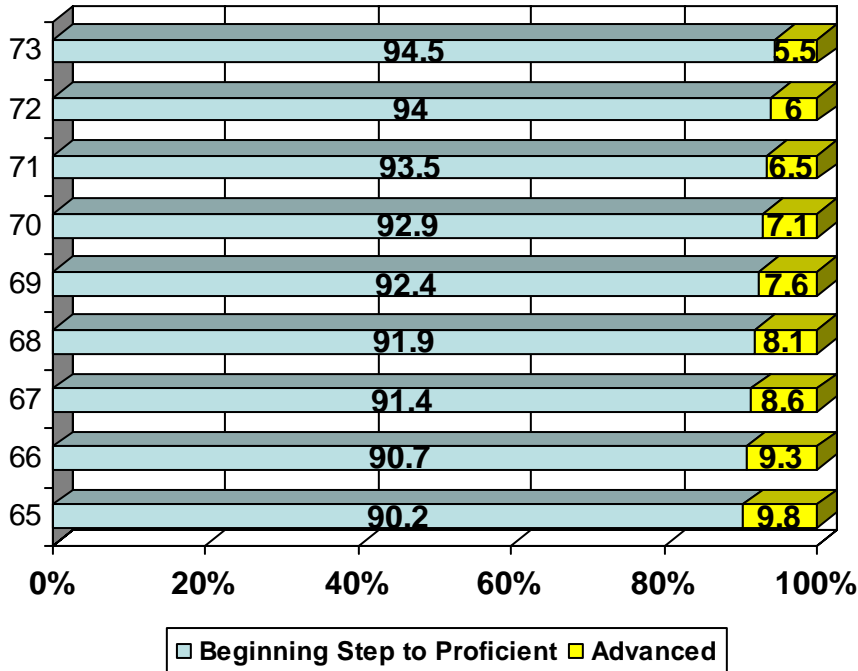


Figure 5.34 Effect of Cut Score Change on the Distribution (Reading: Beginning Step/Nearing Proficient)

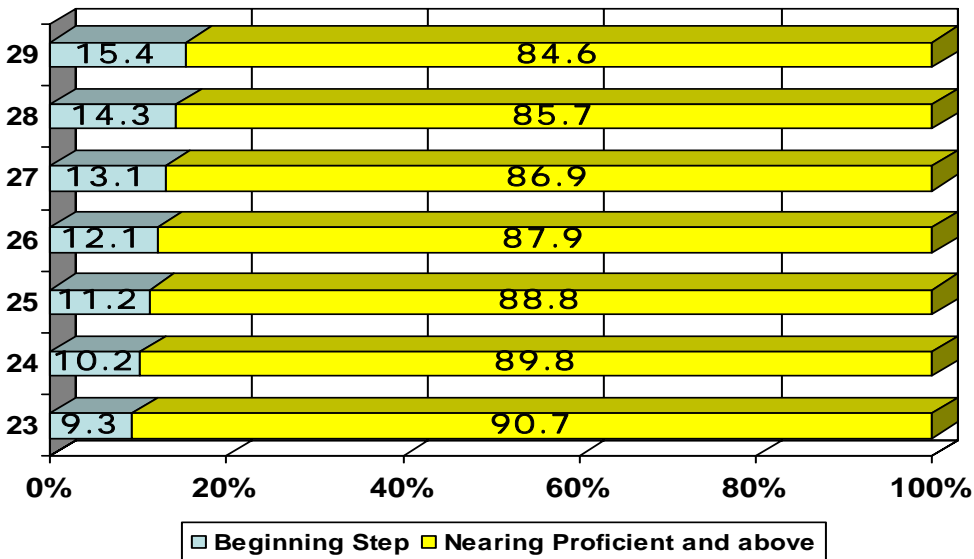


Figure 5.35 Effect of Cut Score Change on the Distribution (Reading: Nearing Proficient/Proficient)

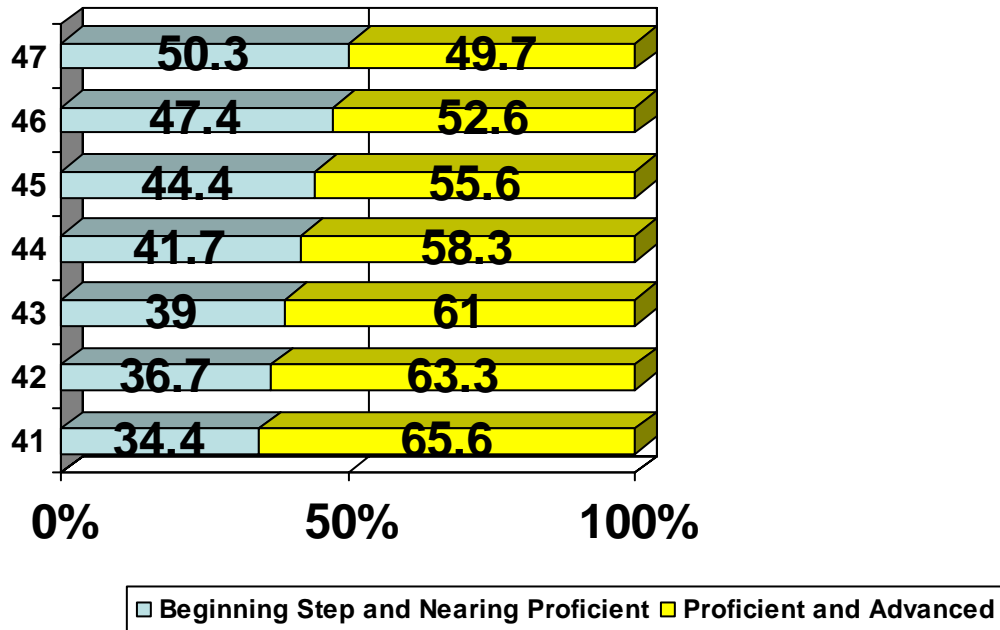
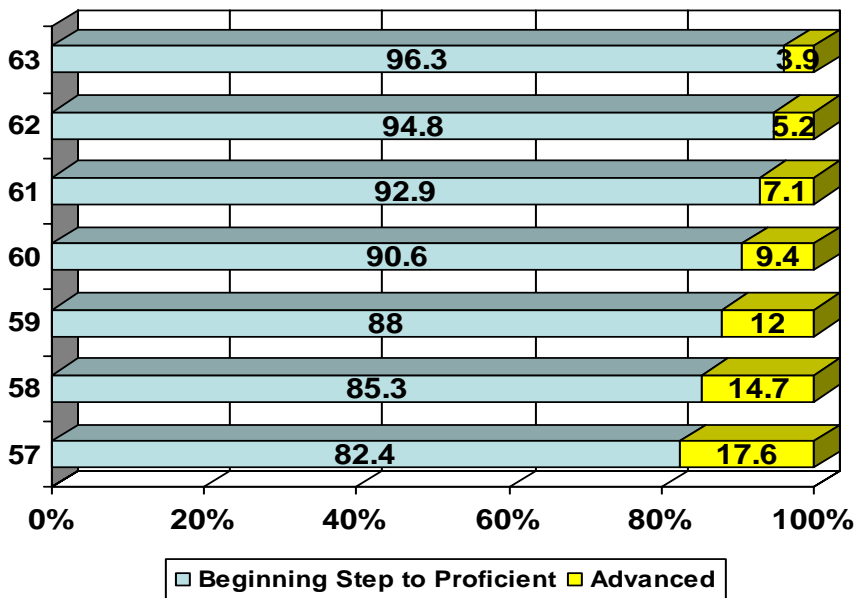


Figure 5.36 Effect of Cut Score Change on the Distribution (Reading: Proficient/Advanced)



Standards Verification Procedure

After reviewing the test- and item-level impact data, as well as the performance distributions based on different cut scores, panelists were encouraged to evaluate and discuss the reasonableness of the current cut scores and whether it was necessary to modify the standards and the necessity of modifying the standards.

Performance standards were then reexamined and verified level-by-level as in the initial standard setting, starting from the cut score between “Nearing Proficient” and “Proficient,” to the cut score between “Beginning Step” and “Nearing Proficient,” and to the cut score between “Proficient” and “Advanced.”

At each level, each panelist was asked to write down a cut score on a recommendation sheet by confidential ballot. If a panelist believed the initial cut score was reasonable and decided to keep the original cut score, he/she would write the original cut score on the sheet. Cut scores from all panelists were collected and the median of the cut scores was used as the new cut score at the specific performance level.

Performance distributions based on the new cut scores were presented to the Panel. Panelists were again asked to evaluate and discuss the reasonableness of the new cut scores. If there was disagreement about a cut score, the verification process was repeated, where panelist wrote down a cut score and data was tallied for all panelists to define a new median cut score.

Results

After two rounds of verification, the Mathematics Standards Verification Panel had decided to keep the cut score between Basic Step and Nearing Proficient (19), and lower the cut score between Nearing Proficient and Proficient by 5 raw score points (from 47 to 42), and lower the cut score between Proficient and Advanced by 3 raw score points (from 73 to 70).

For the Reading standards, the Panel made the decision after two rounds to keep the first two cut scores (Basic Step/ Nearing Proficient; Nearing Proficient/ Proficient) as is, and lower the cut score between Proficient and Advanced by 1 raw score point (from 60 to 59). The modified cut scores for Mathematics and Reading are shown in Table 5.31 below.

The change of performance distribution at different performance levels can be observed from Figure 5.31 to Figure 5.36 provided on pages 53-55. The overall performance

distribution based on the modified cut scores are shown in Table 5.32 for Mathematics and in Table 5.33 for Reading. New distributions by gender are shown in Table 5.34 for Mathematics and in Table 5.35 for Reading. New distributions by ethnic groups are shown in Table 5.36 for Mathematics and in Table 5.37 for Reading.

These cut scores represent the new standards established by the standards verification committee. These updated performance standards will serve as a reference to the final decision to be made by the NMPED.

Table 5.31 Modified Performance Level Standards

Subject	Cut Scores		
	Beginning Step/ Nearing Proficient	Nearing Proficient/ Proficient	Proficient/Advanced
Mathematics	19 (695)	42 (735)	70 (772)
Reading	26 (644)	44 (688)	59 (735)

Note: Numbers in the brackets indicate the relevant scale scores.

Table 5.32 Overall Mathematics Performance Level Distribution Based on the Modified Cut Scores

	Frequency	Percent	Cumulative Percent
Beginning Step	4459	22.5	22.5
Nearing Proficient	9271	46.7	69.2
Proficient	4707	23.7	92.9
Advanced	1406	7.1	100.0
Total	19843	100.0	

Table 5.33 Overall Reading Performance Distribution Based on the Modified Cut Scores

	Frequency	Percent	Cumulative Percent
Beginning Step	2403	12.1	12.1
Nearing Proficient	5869	29.6	41.7
Proficient	9160	46.2	88.0
Advanced	2384	12.0	100.0
Total	19816	100.0	

Table 5.34 Mathematics Performance Distribution by Gender Based on the Modified Cut Scores

	Gender			
	Female		Male	
	Count	%	Count	%
Beginning Step	2006	20.5%	2453	24.4%
Nearing Proficient	4858	49.7%	4413	43.8%
Proficient	2301	23.5%	2406	23.9%
Advanced	612	6.3%	794	7.9%

Figure 5.37 Reading Performance Distribution by Gender Based on the Modified Cut Scores

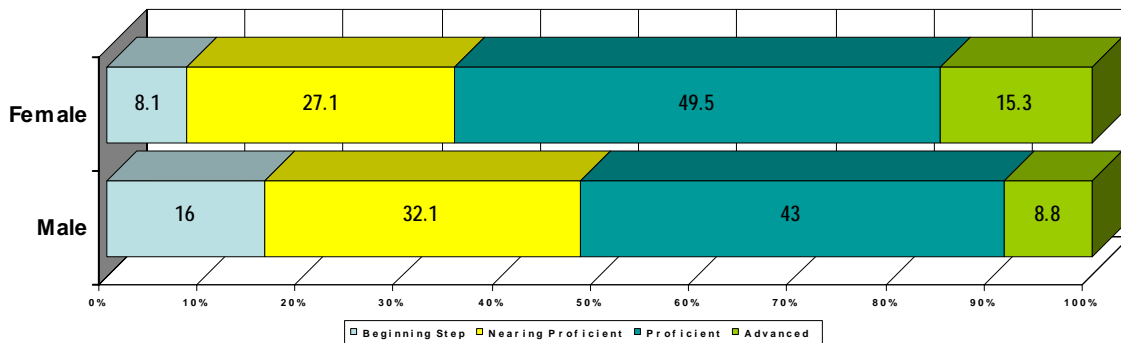
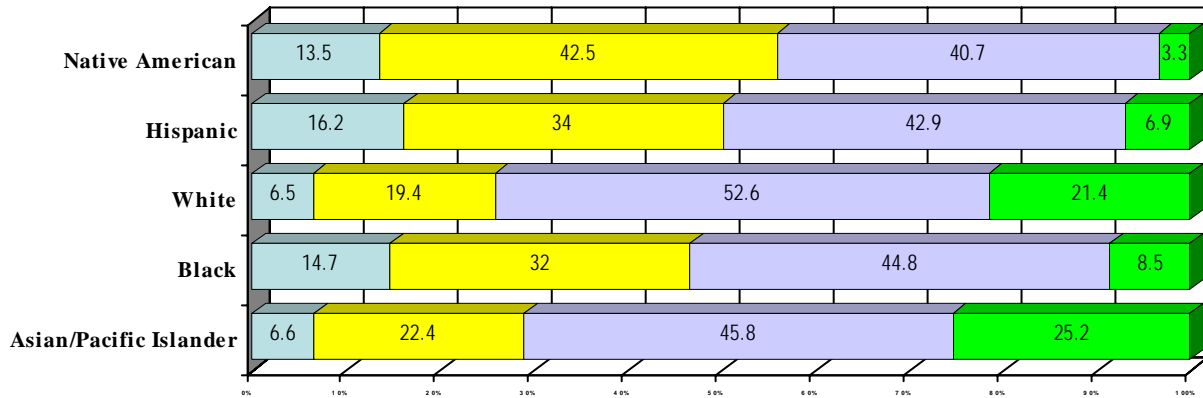


Table 5.35 Mathematics Performance Distribution by Ethnicity Based on the Modified Cut Scores

	Ethnicity									
	Asian/Pacific Islander		Black		White		Hispanic		American Indian/Alaskan	
	Count	%	Count	%	Count	%	Count	%	Count	%
Beginning Step	24	8.4%	150	30.9%	884	12.2%	2639	29.0%	762	27.8%
Nearing Proficient	92	32.3%	243	50.1%	2766	38.2%	4606	50.6%	1564	57.1%
Proficient	96	33.7%	77	15.9%	2557	35.3%	1605	17.6%	372	13.6%
Advanced	73	25.6%	15	3.1%	1030	14.2%	246	2.7%	42	1.5%

Figure 5.38 Reading Performance Distribution by Ethnicity Based on the Modified Cut Scores



CHAPTER 6

RELIABILITY

This chapter provides several reliability indices for the 2005 NMHSSA overall and content strand scores as well as standard and conditional standard errors of the measurement (SEM and CSEM). Special reliability considerations, such as the effects of test content specifications and different item types on reliability. Decision consistency is also addressed. Unless otherwise noted, the results in this section are based on the same data set used in the test and item analyses.

Test reliability refers to the expected consistency of the test scores and is commonly quantified with two indices. One index, the reliability coefficient, expresses the consistency of test scores as the ratio of true score variance to total score variance (true score variance plus error variance). If all test scores were true, the index would equal 1.0. Conversely, the index will be 0.0 if none of the test score variance were true. Clearly, a larger coefficient is better as it indicates the test scores are influenced less by random sources of error. The reliability coefficient is an “unitless” index, which can be compared from test to test. A second statistical index used to describe test reliability is the SEM. The SEM is an index of the random variability in test scores in actual score units, and thus, likely has greater utility for test score users. The CSEM also indicates the degree of measurement error in score units; however, it does so as a function of one’s actual test score. Therefore, the CSEM may be especially useful in characterizing measurement precision in the neighborhood of a score level used for decision-making - such as cut scores for identifying students who meet a performance standard. The CSEM for Reading and Mathematics is indicated on the NMHSSA individual student score points.

6.1 TOTAL TEST RELIABILITY AND OVERALL SEM

The reliability index for overall test scores is based upon Winsteps’ Pearson Separation Index. Because the total score is based on MC, SR, and ER items, this index also combines dichotomous and polytomous item formats because all items were scaled together. This index is similar to the computational formula most often referred to as Coefficient Alpha (Cronbach, 1951) but on the Rasch ability metric instead of the raw score scale. Full test reliabilities were 0.90 for Reading and 0.88 for Mathematic subtests respectively (see Table 6.11). Table 6.11 also represents an IRT analog to the classical test theory overall SEM, which is defined as $s_x \sqrt{1 - r_{xx}}$, where s_x is the standard deviation of the scale score and r_{xx} is the reliability coefficient for the test based on the population.

Coefficient Alpha indicates the internal consistency over the responses to a set of items measuring an underlying unidimensional trait, in this case Reading and Mathematics achievement. While sensitive to random errors associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability or variability in performance that might occur across testing occasions. It is also not sensitive to rater error. Consequently, this index might be positively biased by these factors. On the other hand, there are also factors that might negatively bias the estimate. These issues are addressed in more detail in Sections 6.3 and 6.4.

Table 6.11 Reliability Coefficients and Standard Errors of Measurement

Subject	Reliability	SD	Classical SEM
Reading	.90	75	23.7
Mathematics	.88	60	20.8

Note: Results are based on the scale score metric.

6.2 CSEM

The overall SEM presented in Table 6.11 is the standard deviation of projected replications of the testing procedure averaged over all students. In contrast, the CSEM is conditioned on the ability of the student. The Rasch-based CSEM at the scale score cut points is defined as the reciprocal of the square root of the test information function at the point on the ability continuum that corresponds to the final scale score cut points (Hambleton & Swaminathan, 1985). The CSEM at the three cut scores is presented in Table 6.21. The CSEM at other score points is reported in Table 4.21 and 4.22, in the “S.E. of Scale Score” column. Note that CSEM is smaller in the middle of the score distribution than at the extremes. This pattern is expected for Rasch-based CSEMs.

Table 6.21 Conditional Standard Errors of Measurement at Cut Scores

	Rasch-Based CSEMs at Cut Scores		
	Beginning Step/ Nearing Proficient	Nearing Proficient/ Proficient	Proficient/ Advanced
Reading	9	9	12
Mathematics	9	7	7

6.3 STRATIFICATION BY CONTENT

Because Coefficient Alpha is influenced by content homogeneity, the fact that the tests have items testing different content strands based on test content specifications might reduce the value of the index. Stratified Alpha may be employed in these circumstances to get a more appropriate indication of reliability (Feldt & Brennan, 1989). Stratified Alpha will result in a greater reliability coefficient than Coefficient Alpha when the covariance within content strands is greater than the covariance between content strands. These results are presented in Table 6.31 and 6.32. Stratification by content resulted in

little or no improvement over Coefficient Alpha for the NMHSSA Reading and Mathematics Tests.

Table 6.31 Coefficient Alpha Stratified by Content Strand: Reading

Content Strand	Items	Mean	SD	Alpha	SEM
Total	55	43.86	13.40	.92	3.79
1	37	30.20	9.86	.89	3.27
3	18	13.66	4.11	.84	1.64
Stratified	55	43.86	13.40	.92	3.79

Note: Results are in raw score metric.

Table 6.32 Coefficient Alpha Stratified by Content Strand: Mathematics

Content Strand	Items	Mean	SD	Alpha	SEM
Total	54	34.59	19.19	.93	5.08
2	15	11.50	6.89	.83	2.84
3	19	9.11	6.97	.82	2.96
5	20	13.98	6.91	.79	3.17
Stratified	54	34.59	19.17	.93	5.08

Note: Results are in raw score metric.

6.4 STRATIFICATION BY ITEM TYPE

It should be remembered that the true length of the exam is the number of score points. Because polytomous SR and ER items have more score points than the dichotomous (right or wrong; 0, 1 scored) items, the number of score points is more than the number of items. Strictly speaking, Coefficient Alpha may not be the most appropriate index of reliability under these circumstances (Feldt & Brennan, 1989). Stratified Alpha may also be employed in these cases to get a more appropriate indication of reliability (Qualls, 1995). These results are presented in Table 6.41 and 6.42. Stratification over item type resulted in slight improvements over Coefficient Alpha for both the Reading and Mathematics subtests.

Table 6.41 Coefficient Alpha Stratified by Item Type: Reading

Item Type	Items	Mean	SD	Alpha	SEM
Total	55	43.86	13.40	.92	3.79
MC	46	31.91	8.85	.91	2.66
SR	6	6.42	2.84	.71	1.53
ER	3	5.53	2.91	.53	1.99
Stratified	55	43.86	13.40	.94	3.28

Note: Results are in raw score metric.

Table 6.42 Coefficient Alpha Stratified by Item Type: Mathematics

Item Type	Items	Mean	SD	Alpha	SEM
Total	54	34.59	19.19	.93	5.08
MC	39	18.77	7.57	.86	2.83
SR	8	7.66	5.27	.82	2.24
ER	7	8.15	7.92	.85	3.07
Stratified	54	34.59	19.19	.95	4.29

Note: Results are in raw score metric.

6.5 DECISION CONSISTENCY

In a standards-based testing program there is also interest in knowing how accurately students are classified into the various performance categories. As there was only one administration of the NMHSSA in 2005, it was not feasible to utilize repeat testing in order to estimate the proportion of students who would be reclassified in the same performance levels. However, there are statistical models that can be used to estimate the consistency of classifications when data from only one administration is available. A method based on the beta-binomial model (Huynh, 1976 and 1979) was used in this study for this purpose.

Using a FORTRAN computer program (Huynh, 1979), consistency indices for performance levels were calculated. Using the maximum possible score (MPS), mean and standard deviation, the program first estimates the KR-21 reliability index and the two parameters of the beta-binomial distribution for the test data. Then using the formulae referenced in Huynh 1976, the program estimates the univariate and bivariate frequency distributions. The program then computes the agreement and kappa indices and associated standard errors for the classification of students to performance levels. The agreement index depicts the proportion of students who are consistently classified in the same achievement levels on two equivalent administrations of the test. The kappa index, on the other hand, reflects the level of improvement beyond the chance level in the consistency of classifications. Kappa is a “threshold loss” index that treats all misclassifications as having the same relative importance.

Both raw consistency and kappa indices were computed for each test for the case four performance levels: *Beginning Step*, *Nearing Proficient*, *Proficient* and *Advanced*. With the NMHSSA, test results are also used for AYP decisions that only involve two outcomes: below *Proficient* classification and *Proficient* and above classification. Consequently, computations were also carried out for the two categories *Below Proficient* and *Proficient or above*.

It may be noted that consistency indices for the four performance levels are lower than those based on two categories. This is not surprising since classification using four levels would allow more opportunity to change the achievement levels. Hence there would be

more classification errors in the four achievement levels, resulting in lower consistency indices.

Table 6.5 Consistency Indices for Performance Levels for the 2005 NMHSSA

Content	Two Achievement Levels		Four Achievement Levels	
	Proportion of Agreement	Kappa	Proportion of Agreement	Kappa
Reading	.87	.75	.75	.66
Mathematics	.89	.77	.78	.69

CHAPTER 7

VALIDITY

As noted in the *Standard for Educational and Psychological Testing*, “validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by the proposed uses of the tests” (AERA, APA, & NCME, 1999, p.9). Thus, the validity of the NMHSSA must be judged in relation to its primary purpose: to evaluate school Annual Yearly Progress (AYP) for No Child Left Behind (NCLB).

The following sections outline evidence of valid score use with respect to the NMHSSA. The sections are structured based on two of the five sources of validity evidence identified in the *Standards for Educational Psychological Testing*, that is, evidence based on test content and internal structure.

7.1 EVIDENCE RELATED TO TEST CONTENT

An important source of information regarding the validity of the NMHSSA scores is the link of each price of the assessment to New Mexico’s Content Standards and Benchmarks. Such content-related evidence of validity is presented in terms of how the 2005 NMHSSA assessments were assembled. Included is detailed information regarding both the item development procedures and the content coverage of the Reading and Mathematics subtests.

2004-2005 Item Development

Pearson Educational Measurement (PEM) content specialists performed the following tasks in developing the test items for the NMHSSA Reading and Mathematics.

- Reviewed the *New Mexico Content Standards and Benchmarks* and the *New Mexico Career Readiness Standards*
- Developed the *Item Specifications* document based on the *New Mexico Career Standards* and the RFP
- Developed the *Item Specifications* document based on *Test Specifications* document and New Mexico’s Reading and Mathematics standards
- Held a conference call to review the *Test Specifications and Item Specifications* documents with a committee of New Mexico educators and NMPED staff
- Made changes to the *Test and Items Specifications* documents as requested by the NMPED
- Recruited and selected experienced item writers

- Trained item writers on the NMPED approved *NMHSSA Test and Item Specifications*
- Conducted internal reviews of items, rubrics, passages and stimulus materials submitted
- Obtained permissions for passages and stimulus materials as needed
- Revised items, rubrics, passages and stimulus materials as necessary
- Prepared materials for external content and bias review
- Revised items per New Mexico educator committees as necessary

PEM developed and selected all items in strict accordance with the *NMHSSA Test and Item Specifications* document that was developed for, and approved by, the NMPED. Following each stage of development, PEM content specialists made requested revisions to items. At all phases, PEM reviewed the items to ensure that, in addition to matching the item specifications, the items were:

- Aligned to the *New Mexico Content Standards*
- Fair for students in all gender and ethnic groups
- Grade-level appropriate for the students who would be taking the NMHSSA

Educator Review

PEM worked closely with the NMPED to train New Mexico educators who have content expertise to review items. The Bias and Content Review Committees were representative of the diverse geographic and ethnic groups of New Mexico.

Bias Review Meeting: PEM’s content specialists conducted the bias review sessions. PEM took great care throughout the item-development process to monitor items for potential bias and to ensure appropriate representation of the various segments and groups of the New Mexico population. An important part of the item development process was the training item writers received to help them identify and eliminate potential bias from the materials they created.

New Mexico educators reviewed the items to judge the extent to which they may reflect gender and racial/ethnic stereotypes, favor one group over another, or offend a particular subgroup. PEM prepared materials for use in the review session. All meeting materials were given to committee participants on the day of the meeting, and security procedures were established.

Content Review: Items were reviewed for content and alignment to the New Mexico standards. PEM began the sessions with a training period led by the NMPED staff and PEM content specialists. PEM prepared copies of the agenda, all materials needed for the content review sessions and the items to be reviewed.

After training, New Mexico educators reviewed the items and identified those that needed modification or revision. The committee reviewed the items one by one, focusing on each item’s fit with the *Item Specifications*, the appropriateness of the item’s content,

and the item's match to the *New Mexico Content Standards*. The committee either approved the test item, suggested edits to the item, or recommended that the item be dropped.

Item Distribution Across Strands

Section 3.1 presents the general construct of the 2005 NMHSSA Reading and Mathematics subtests. More detailed information regarding the content coverage of the items on the operational NMHSSA tests are provided in Table 7.11 and 7.12 below.

Table 7.11 Spring 2005 NMHSSA Reading Item Distribution by Content Standards and Benchmarks

Content Strands		Standards	Benchmark	Number of Item			Score Points
				MC	SR	ER	
1	Reading and Listening for Comprehension	Students will apply strategies and skills to comprehend information that is read.	A. Read, react to, and analyze information	5	1	0	7
			C. Critical thinking to evaluate information and solve problems	18	1	3	31
			D. Evaluate print, non-print, and technology-based information	6	3	0	12
3	Literature and Media		B. Understand literary elements, concepts, and genres	17	1	0	19
Total				46	12	12	70

Table 7.12 Spring 2005 NMHSSA Mathematics Item Distribution by Standards and Benchmarks

Content Strands	Standards	Benchmark	Number of Item			Score Points	
			MC	SR	ER		
2	Algebra	Students will understand algebraic concepts and applications.	1. Understand patterns, relations, functions and graphs	4	1	0	7
			2. Represent and analyze mathematical situations and structures using algebraic symbols	1	3	0	7
			3. Use mathematical models to represent and understand quantitative relationships	0	1	1	7
			4. Analyze changes in various contexts	3	0	1	7
3	Geometry	Students will understand geometric concepts and applications.	1. Analyze characteristics and properties two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships	5	0	1	11
			2. Specify locations and describe spatial relationships using coordinate geometry and other representational systems	2	1	1	9
			3. Apply transformations and use symmetry to analyze mathematical situations	3	0	0	3
			4. Use visualization, spatial reasoning, and geometric modeling to solve problems	5	0	1	9
5	Data Analysis and Probability	Students will understand how to formulate questions, analyze data, and determine probabilities.	1. Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them	4	0	0	4
			2. Select and use appropriate statistical methods to analyze data	4	0	1	8
			3. Develop and evaluate inferences and predictions that are based on data	6	1	0	8
			4. Understand and apply basic concepts of probability	2	1	1	10
Total			39	8	7	90	

7.2 EVIDENCE BASED ON INTERNAL STRUCTURE

Inter-correlations among NMHSSA Strands within a Content Strand

Table 7.21 gives the correlations among the strands within the NMHSSA Reading and Mathematics. As can be seen, scores for Reading strands (Reading and Listening for Comprehension and Literature and Media) are fairly well correlated (.86). Correlations among the Mathematics strands (Algebra, Geometry; Data Analysis and Probability) are slightly lower than the Reading strands but are still quite significant (.78 to .82). The lowest correlation within a content strand is the correlation between Geometry and Data Analysis and Probability. However, it only differs with the highest correlation (between the two Reading Strands) by .08.

Inter-correlations among NMHSSA Strands between Two Content Strands

Even though correlations between Reading and Mathematics strand scores are all lower than those within the content strand, the values are still moderately significant, which range from .52 to .75. This result suggests 2005 NMHSSA Reading and Mathematics are of different constructs but also correlated to a certain degree.

Table 7.21 Content Strand Correlation Matrix for the 2005 NMHSSA

	Reading Strand 1	Reading Strand 3	Mathematics Strand 2	Mathematics Strand 3	Mathematics Strand 5
Reading Strand 1	-	-	-	-	-
Reading Strand 3	.86	-	-	-	-
Mathematics Strand 2	.69	.60	-	-	-
Mathematics Strand 3	.62	.52	.80	-	-
Mathematics Strand 5	.75	.65	.82	.78	-

Factor Analysis

In order to follow up on the strand inter-correlations across Reading and Mathematics, an exploratory factor analysis was conducted with the Mathematics and Reading strand scores. A principal components analysis was conducted using SAS. The number of factors was determined using three criteria: eigen values greater than 1, a scree test for the eigen values, and finding the solution in which close to 60 percent of the variance was explained. The result was a one-factor solution in which 77 percent of the variance was explained. Table 7.22 shows the factor loadings from the component matrix for the one-factor solution. Table 7.23 gives the total variance explained by the five components. Figure 7.23 is the scree plot that illustrates the variance explained by each factor (or component). This result is consistent with the analysis of the inter-correlations.

While both analyses suggest that 2005 NMHSSA Reading and Mathematics scores may be correlated, clearly Reading and Mathematics are separate dimensions of performance on the NMHSSA as a whole. Based on the a priori belief that there should be two distinct factors, a two-factor solution was further pursued. As shown in Table 7.24, the factor loadings of the two-factor solution also have “simple structure” and show that the Reading content strands clearly load on the second factor while Mathematics content strands do not. Finally, the correlation between the two latent factors was .85. Such a strong, positive correlation between latent factors is fairly typical for academic achievement constructs.

Table 7.22 Factor Loadings of the Five Strand Scores Based on One-Factor Solution

	Extracted Factor 1
Reading Strand 1	.90
Reading Strand 3	.82
Mathematics Strand 2	.89
Mathematics Strand 3	.85
Mathematics Strand 5	.92

Table 7.23 Total Variance Explained by the Five Components

Factor	Eigen Values	Percent of Variance	Cumulative Percent of Variance
1	3.84	76.81	76.81
2	.66	13.16	89.96
3	.20	3.95	93.92
4	.18	3.62	97.54
5	.12	2.46	100.00

Figure 7.23 Scree Plot from the Factor Analysis of the 2005 NMHSSA

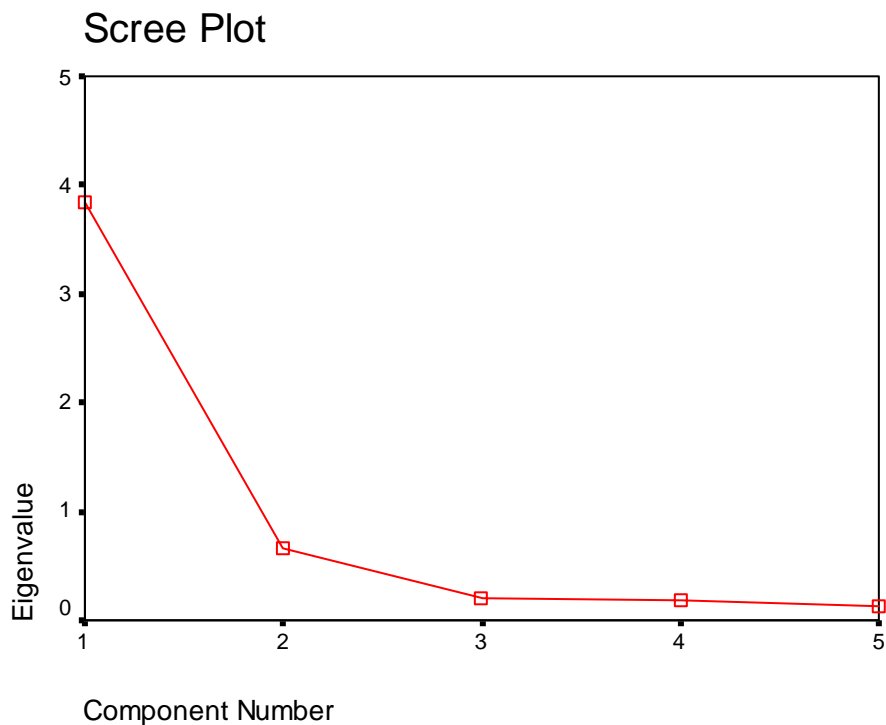


Table 7.24 Factor Loadings of the Five Strand Scores Based on Two-Factor Solution

	Extracted Factor 1	Extracted Factor 2
Reading Strand 1	.90	.35
Reading Strand 3	.82	.51
Mathematics Strand 2	.89	-.27
Mathematics Strand 3	.85	-.41
Mathematics Strand 5	.92	-.16

Differential Item Functioning

Differential Item Functioning (DIF) with respect to gender and ethnicity was used to address construct-irrelevant variance, which represents an important threat to the validity of achievement tests. DIF analyses were conducted for all groups of interest to the NMPED: Caucasian and Hispanic, Caucasian and Native American ethnic groups, and male and female. For this analysis, subgroups were assigned to either the Reference or Focal group. Males were used as the reference group for gender comparisons while Caucasian served as the reference group in the ethnicity comparisons. There was an insufficient population of African-American and Asian students to perform reliable DIF analyses for these groups.

Though the statistics in DIF analysis can reveal differential item functioning between groups, these procedures cannot attribute cause. Like all statistical procedures, DIF methods will identify some items as biased, when in fact, no bias exists. For this reason, DIF analysis is best viewed as a procedure for flagging items that warrant closer inspection.

The Mantel-Haenszel (MH) procedure was used to flag items showing DIF. Examinees were grouped according to their ability (based on their total test score) in each content area. This was done because DIF analyses are primarily designed to detect differential item performance across subgroups of populations while controlling for ability. For example, a group of high-ability students are expected to have better performance in an item than a group of low-ability students. Such performance is the desired result in a “fair” testing environment. In contrast, DIF seeks to identify those items where subgroups of students with similar ability perform at measurably different levels.

For both dichotomous and polytomous items, DIF classification was based on the statistical significance of the group differences. The differences of the difficulty measures (b parameters) between the two contrasting groups were calculated. Items with MH Delta (MHD) significantly different from 0 (based on $\alpha = 0.05$) were flagged. The polytomous DIF classification, using Generalized Mantel-Haenszel procedure, was based on chi-square significance ($\chi^2 > 3.841, p < .05$).

Table 7.25 and 7.26 summarize the items flagged for DIF based on the statistical analysis.

Table 7.25 Differential Item Functioning Summary for the Reading Test

Item Number	Item Type	Ethnic DIF Caucasian/ Native American		Ethnic DIF Caucasian/ Hispanic		Gender DIF Male/Female	
		Difference in Difficulty ⁷	Flag	Difference in Difficulty	Flag	Difference in Difficulty	Flag
1	MC	.4168	D	.1545		.1884	
2	MC	.0570		-.0011		.0498	
3	MC	.2621		.2695	D	.1787	
4	MC	-.0843		-.1712		.3669	D
5	MC	.5579	D	.3607	D	.1807	
6	MC	.4153	D	.1559		.2150	
7	MC	.5161	D	.3791	D	-.2701	D
8	MC	.1114		.1176		-.0739	
9	MC	-.2011		-.1420		-.4433	D
10	MC	.2960	D	-.0770		.0819	
11	MC	-.0308		-.0754		.5760	D
12	ER	-.1800	D	-.1535		.1043	
13	MC	.8280	D	.2953	D	.3373	D
14	MC	-.2338		-.2519	D	.0358	
15	MC	-.3328	D	-.2153		.0873	
16	SR	.0680		.1664		-.0014	
17	SR	-.3782	D	-.3204	D	-.1805	
18	MC	.9286	D	.4246	D	.0493	
19	MC	.2506		.1371		.0432	
20	MC	.0385		.1181		-.1175	
21	SR	-.1926	D	-.1196		.0185	
22	SR	.1384		.0383		-.1062	
31	MC	-.1599		-.4357	D	.2687	D
32	MC	-.1344		-.2118		.2464	D
33	MC	-.1164		-.0737		.1047	
34	MC	-.1339		-.0661		.1346	
35	MC	.9547	D	.4513	D	.2559	D
36	MC	.7433	D	.4347	D	.2090	
37	MC	.2815	D	.1845		.1375	
38	MC	.2655		.1863		.3722	D
39	MC	.2119		.0006		.0511	
40	MC	.4800	D	.2005		-.0140	
41	MC	.4747	D	.2142		-.0443	
42	ER	-.0663		-.0893		-.1366	
55	MC	.1105		.1165		.1854	
56	MC	.3121		.1250		-.3206	D
57	MC	.2550		.1312		-.0887	
58	MC	.1114		.2127		-.1774	
59	MC	-.1846		.1727		-.3941	D
60	MC	.0143		.2796	D	.0260	
61	MC	-.4766	D	-.0452		.0806	
62	MC	.1499		.3751	D	.0783	
63	MC	-.4996	D	-.0986		-.0922	
64	MC	-.2053		-.0821		.1048	
65	MC	.3571	D	.2263		.0059	
66	MC	-.3416		.2213		-.2494	D
67	MC	-.1146		.2041		-.1076	
68	MC	-.8497	D	-.4096	D	-.2844	D
69	MC	.1869		.2286	D	-.1895	
70	MC	-.0200		.0850		.2159	
71	MC	.5014	D	.3130	D	-.1947	
72	MC	.0374		.3140	D	-.0116	
73	SR	-.5702	D	-.2063		-.4204	D
74	SR	-.3402	D	-.0670		-.1626	
75	ER	-.3649	D	-.1360		-.2842	D

⁷ Positive difference indicates that the item is in favor of the reference group.

Table 7.26 Differential Item Functioning Summary for the Mathematics Test

Item Number	Item Type	Ethnic DIF Caucasian/ Native American		Ethnic DIF Caucasian/ Hispanic		Gender DIF Male/Female	
		Difference in Difficulty	Flag	Difference in Difficulty	Flag	Difference in Difficulty	Flag
1	MC	.2379		.0420		.1246	
2	MC	-.0129		.1472		.1266	
3	MC	.7279	D	.3903		.1992	
4	MC	.3185	D	.2602	D	.4662	D
5	MC	.3862	D	.2556	D	.1449	
6	MC	-.1002		-.1490		.1497	
7	MC	.3054	D	.0424		.1465	
8	MC	-.2041		.1571		.2486	D
9	MC	-.1805		-.2500	D	-.0685	
10	MC	-.3685	D	-.3815	D	-.2996	D
11	MC	.2773	D	.0727		.1125	
12	MC	.1981		.0800		.1126	
13	MC	-.3824	D	-.2060		-.2421	D
14	ER	-.0496		-.0258		-.1698	
15	SR	-.1630		-.0958		-.1442	
16	ER	-.1707		-.1058		.0271	
17	ER	-.0770		.0321		-.1196	
18	SR	.2894	D	.1029		.1207	
19	MC	-.0894		-.1415		-.1756	
20	MC	-.3552	D	-.3550	D	.2177	
22	MC	.0644		.0650		-.0298	
23	MC	.2994	D	.2130		.2212	
24	MC	-.0559		.0273		-.1370	
25	MC	-.2245		-.2805		.1376	
26	MC	.3073	D	.2095		.1095	
27	MC	.2493	D	.1998		-.0537	
28	MC	.4275	D	.1852		.5026	D
30	MC	.2111		.0304		.2109	
32	MC	.4321	D	.1852		.4669	D
34	MC	-.2676	D	-.2925	D	.1877	
35	MC	.6521	D	.4947	D	.1875	
38	MC	.6437	D	.2404	D	.6840	D
39	SR	-.3002		-.1136		-.1465	
40	SR	-.4354	D	-.2156		-.1773	
41	SR	.2209		.1519		.0466	
42	SR	-.3174	D	-.1333		-.1257	
43	SR	-.1796		-.0641		-.1374	
45	ER	-.0060		.0054		.0567	
46	MC	.3504	D	.2642	D	.1527	
47	MC	.1687		.1127		.0330	
49	MC	-.4974	D	-.3765	D	.1722	
50	MC	.4540	D	.1597		.1217	
51	MC	-.0148		-.0596		.4465	D
52	MC	.4935	D	.3257	D	.1515	
54	MC	-.2096		-.1761		.3040	D
59	MC	.2823	D	.2130		-.2715	D
62	MC	-.1107		-.0811		.2158	
63	MC	-.8967	D	-.7698	D	-.1454	
64	MC	.0566		.0801		-.0150	
65	MC	.4159	D	.2385	D	.1285	
69	ER	.1745	D	.0348		-.0811	
70	SR	.2468	D	.0853		-.0263	
72	ER	-.1622		.0132		-.0537	
74	ER	.0892		.1545		-.1246	

REFERENCES, RESOURCES, AND RELATED DOCUMENTS

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 105-146). New York, NY: ACE/Macmillan.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Pub.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253-64.

Huynh, H. (1979). FORTRAN program to calculate beta-binomial decision consistency reliability. Columbia, SC: College of Education, University of South Carolina.

Maters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

New Mexico Public Education Department (2003a). *New Mexico high school standards assessment (NMHSSA) test and items specifications mathematics draft*.

New Mexico Public Education Department (2003b). *New Mexico high school standards assessment (NMHSSA) test and item specifications part one: Reading test Specifications*.

New Mexico Public Education Department. (2003c). *Test administration manual: Reading and mathematics*.

New Mexico Public Education Department. (2003d). *New Mexico high school assessment test coordinator manual: Reading and mathematics*.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*(2), 11-120.

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

Smith, R. M., & Miao C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (vol. 2, chap. 18). Norwood, NJ: Ablex.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Wright, B. D., & Linacre, J. M. (1999). *A user's guide to WINSTEPS, MINISTEP, BIGSTEPS, Rasch-Model Computer Programs*. Chicago: Mesa Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Educational Measurement: Issues and Practices*, 7(4), pp. 16-17.