



Spring 2011

Nebraska State Accountability (NeSA)

Grades 3-8 and 11

Reading Operational and Field Test

Mathematics Operational and Field Test

Science Field Test

Technical Report

October 2011

Prepared by Data Recognition Corporation





2010 NEBRASKA STATE ACCOUNTABILITY (NeSA) TECHNICAL REPORT

TABLE OF CONTENTS

1. BACKGROUND

1.1. Purpose and Organization of This Report	1
1.2. Background of the NeSA Assessment	1
• Previous Nebraska Assessment (STARS)	
• Purpose of the NeSA	
• Phase-In Schedule for NeSA	
• Advisory Committees	

2. ITEM AND TEST DEVELOPMENT

2.1. Content Standards.....	3
2.2. Test Blueprints.....	4
2.3. Multiple-Choice Items.....	4
2.4. Passage Selection.....	4
2.5. Item Development and Review	4
• Item Writer Training	
• Item Writing	
• Item Review	
• Editorial Review of Items	
• Universal Designed Assessments	
• Depth of Knowledge	
• Item Content Review	
• Sensitivity and Bias Review	
2.6. Item Banking	16
2.7. The Operational Forms Construction Process	16
• DIF in Operational Forms Construction	
• Review of the Items and Test Forms	
2.8. Reading Assessment.....	18
• Test Design	
• Psychometric Targets	
• Equating Design	
2.9. Math Assessment.....	19
• Test Design	
• Psychometric Targets	
• Equating Design	

2.10 Science Assessment	20
• Initial Standalone Field Test	
• Forms Assembly	
• Forms Approval Meeting	
• Equating Design	
3. READING AND MATHEMATICS OPERATIONAL ASSESSMENT	
3.1. Rasch Calibrating and Equating	22
3.2. Validity and Reliability	22
• Items Analyses	
• Item Difficulty	
• Forms Performance Summary	
• State Performance Summary	
• Decision Consistency	
3.3. Setting Performance Standards	36
3.4. Scale Score Metric	38
3.5. Reading Pre- and Post-equated Comparison	40
4. READING AND MATH EMBEDDED FIELD TEST	
4.1. Psychometric Summary	42
• Traditional Item Statistics	
• Differential Item Functioning (DIF)	
5. SCIENCE STANDALONE FIELD TEST	
5.1. Sampling Design	48
5.2. Psychometric Summary	48
• Traditional Item Statistics	
• Differential Item Functioning (DIF)	
6. ONLINE TESTING TIMES	51
7. REFERENCES.....	54
8. APPENDICES	
A. Legislative Bill 1157.....	56
B. NeSA-R Test Blueprint.....	68
C. NeSA-M Test Blueprint.....	86
D. Confidentiality Agreement	123
E. Fairness in Testing Manual.....	124
F. Overview of Rasch Measurement Models.....	141

G. Reading Key Verification and Foil Analysis.....	145
H. Mathematics Key Verification and Foil Analysis	167
I. Science Key Verification and Foil Analysis.....	191
J. Reading Field Test Differential Item Functioning.....	208
K. Mathematics Differential Item Functioning	223
L. Science Differential Item Functioning.....	253
M. Reading and Mathematics Operational Form Calibration Summary	274
N. Mathematics Performance Level Descriptors.....	281
O. Reading and Mathematics Raw-to-Scale Conversion Tables and Distributions of Ability	289
P. Reading Item Bank Difficulties	321
Q. Mathematics Item Bank Difficulties.....	333
R. Science Field Test Estimated Item Bank Difficulties.....	345
S. Reading Pre- and Post-Equating Summary	353
T. Reading and Mathematics Demographic Summary Sheets.....	358



1. BACKGROUND

1.1 PURPOSE AND ORGANIZATION OF THIS REPORT

This report documents the technical aspects of the 2011 Nebraska Reading (NeSA-R) and Mathematics (NeSA-M) operational tests, NeSA-R and NeSA-M embedded field tests, and the Nebraska Science (NeSA-S) standalone field test, covering details of item and test development process, administration procedures, and psychometric methods and summaries.

1.2 BACKGROUND OF THE NEBRASKA STUDENT ASSESSMENT (NESEA)

Previous Nebraska Assessments: In previous years, Nebraska administered a blend of local and state-generated assessments to meet NCLB requirements called STARS (School-based Teacher-led Assessment and Reporting System). STARS was a decentralized local assessment system that measured academic content standards in reading, mathematics, and science. The state reviewed every local assessment system for compliance and technical quality. The Nebraska Department of Education (NDE) provided guidance and support for Nebraska educators by training them to develop and use classroom-based assessments. For accreditation, districts were also required to administer national norm-referenced tests (NRT).

As a component of STARS, NDE administered one writing assessment annually in grades 4, 8, and 11. In addition, NDE provided an alternate assessment for students severely challenged by cognitive disabilities.

Purpose of the NeSA: Legislative Bill 1157 passed by the 2008 Nebraska Legislature (<http://uniweb.legislature.ne.gov/FloorDocs/Current/PDF/Slip/LB1157.pdf>) required a single statewide assessment of the Nebraska academic content standards for writing, reading, mathematics, and science in Nebraska's K-12 public schools. The new assessment system was named NeSA (Nebraska State Accountability), with NeSA-R for reading assessments, NeSA-M for mathematics, and NeSA-S for science. The assessments in reading and math were administered in grades 3-8 and 11; science will be administered in grades 5, 8, and 11 in 2012.

NeSA replaced previous school-based assessments for purposes of local, state, and federal accountability. NeSA consists entirely of multiple choice items and will be administered, to the extent practicable, online. In January 2009, the Nebraska Department of Education (NDE) contracted with Data Recognition Corporation (DRC) to support the Department of Education with the administration, record keeping, and reporting of statewide student assessment and accountability.

Phase-In Schedule for NeSA: The NDE prescribed such assessments starting in the 2009-2010 school year to be phased in as shown in Table 1.2.1. The state intends to use the expertise and experience of in-state educators to participate, to the maximum extent possible, in the design and development of the new statewide assessment system.

Table 1.2.1: NeSA Administration Schedule

Subject	Administration Year		Grades
	Field Test	Operational	
Reading	2009	2010	3 through 8 plus high school
Mathematics	2010	2011	3 through 8 plus high school
Science	2011	2012	Elementary, middle, and high school

Advisory Committees: LB 1157 added a governor-appointed Technical Advisory Committee (TAC) with three nationally recognized experts in educational assessment, one Nebraska administrator, and one Nebraska teacher. The TAC reviewed the development plan for the NeSA, and provided technical advice, guidance, and research to help the NDE make informed decisions regarding standards, assessment, and accountability.

2. ITEM AND TEST DEVELOPMENT

2.1 CONTENT STANDARDS

In April of 2008, the Nebraska Legislature passed into state law Legislative Bill 1157 (Appendix A). This action changed previous provisions related to standards, assessment, and reporting. Specific to standards, the legislation stated:

- The State Board of Education shall adopt measurable academic content standards for at least the grade levels required for statewide assessment. The standards shall cover the subject areas of reading, writing, mathematics, science, and social studies. The standards adopted shall be sufficiently clear and measurable to be used for testing student performance with respect to mastery of the content described in the state standards.
- The State Board of Education shall develop a plan to review and update standards for each subject area every five years.
- The State Board of Education shall review and update the standards in reading by July 1, 2009, the standards in mathematics by July 1, 2010, and these standards in all other subject areas by July 1, 2013.

The Nebraska Language Arts Standards are the foundation for Nebraska State Accountability – Reading (NeSA-R). This assessment instrument is comprised of items that address standards for grades 3–8 and 12. The standards are assessed at grade-level with the exception of grade 12. The grade 12 standards are assessed on the NeSA tests at grade 11. The reading standards for each grade are represented in items that are distributed between two reporting categories: Vocabulary and Comprehension. The Vocabulary standards include word structure, context clues, and semantic relationships. The Comprehension standards include author’s purpose, elements of narrative text, literary devices, main idea, relevant details, text features, genre, and generating questions while reading.

The mathematics component of Nebraska State Accountability is composed of items that address indicators in grades 3–8 and high school. The standards are assessed at grade level with the exception of high school. The high school standards are assessed on the NeSA-M at grade 11. The assessable standards for each grade level are distributed among the four reporting categories: Number Sense Concepts, Geometric/Measurement Concepts, Algebraic Concepts, and Data Analysis/Probability Concepts. The National Council of Teachers of Mathematics (NCTM) and the National Assessment of Educational Progress (NAEP) standards are the foundation of the Nebraska Mathematics standards.

The science component of the Nebraska State Accountability is composed of items that address indicators in grade-band strands 3–5, 6–8, and 9–12. The NeSA-S assesses the standards for each grade-band strand at a specific grade: 3-5 strand at grade 5, 6–8 strand at grade 8, and 9–12 strand at grade 11. The assessable standards for each grade level are distributed among the four reporting categories: Inquiry, The Nature of Science, and Technology; Physical Science; Life Science; and Earth and Space Sciences.

2.2 TEST BLUEPRINTS

The test blueprints for each assessment include lists of all the standards, organized by reporting categories. The test blueprints also contain the Depth of Knowledge level assigned to each standard and the range of test items to be part of the assessment by indicator. The NeSA-R test blueprint was developed and approved in fall 2009 (Appendix B). The NeSA-M test blueprint was developed and approved in fall 2010 (Appendix C).

2.3 MULTIPLE-CHOICE ITEMS

Each assessment incorporates multiple-choice items to assess the content standards. Students are required to select a correct answer from four response choices with a single correct answer. Each multiple-choice item is scored as right or wrong and has a value of one raw score point. Multiple-choice items are used to assess a variety of skill levels in relation to the tested standards.

2.4 PASSAGE SELECTION

All items in the reading assessment were derived from a selection of narrative and informational passages. Passages acquired were “authentic” in that they were purchased from the test vendor that commissioned experienced passage writers to provide quality pieces of text. Passages were approved by a group of reading content specialists that have teaching experience at specific grade levels. These experts were given formal training on the specific requirements of the Nebraska assessment of reading. The group, under the facilitation of the NDE test development team, screened and edited passages for:

- interest and accuracy of information in a passage to a particular grade level;
- grade-level appropriateness of passage topic and vocabulary;
- rich passage content to support the development of high-quality test questions;
- bias, sensitivity, and fairness issues; and
- readability considerations and concerns.

Passages that were approved moved forward for the development of test items.

The readability of a passage was an evaluative process made by Nebraska educators, NDE’s test development team, DRC’s reading content specialists, and other individuals who understand each particular grade level and children of a particular age group. In addition, formal readability programs were also used by DRC to provide a “snapshot” of a passage’s reading difficulty based on sentence structure, length of words, etc. All of this information, along with the classroom context and content appropriateness of a passage, was taken into consideration when placing a passage at a particular grade.

2.5 ITEM DEVELOPMENT AND REVIEW

The most significant considerations in the item and test development process are: aligning the items to the grade level indicators; determining the grade-level appropriateness; depth of knowledge; estimated difficulty level; and determining style, accuracy, and correct terminology. In addition, the *Standards*

for Educational and Psychological Testing (AERA, APA, & NCME, 1999) and *Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the following steps in the item development process.

- Analyze the grade-level indicators and test blueprints.
- Analyze item specifications and style guides.
- Select qualified item writers.
- Develop item-writing workshop training materials.
- Train Nebraska educators to write items.
- Write items that match the standards, are free of bias, and address fairness and sensitivity concerns.
- Conduct and monitor internal item reviews and quality processes.
- Prepare passages and items for review by a committee of Nebraska educators (content and bias/sensitivity).
- Select and assemble items for field testing.
- Field test items, score the items, and analyze the data.
- Review items and associated statistics after field testing, including bias statistics.
- Update item bank.

Item Writer Training: The test items were written by Nebraska educators who were recommended for the process by an administrator. Three criteria were considered in selecting the item writers: educational role, geographic location, and experience with item writing.

Prior to developing items for NeSA, a cadre of item writers was trained with regard to:

- Nebraska content standards and test blueprints;
- cognitive levels, including depth of knowledge;
- principles of Universal Design;
- skill-specific and balanced test items for the grade level;
- developmentally appropriate structure and content;
- item-writing technical quality issues;
- bias, fairness, and sensitivity issues; and
- style considerations and item specifications.

Item Writing: To ensure that all test items met the requirements of the approved target content test blueprint and were adequately distributed across subcategories and levels of difficulty, item writers were asked to document the following specific information as each item was written.

- **Alignment to the Nebraska Standards:** There must be a high degree of match between a particular question and the standard it is intended to measure. Item writers were asked to clearly indicate which standard each item was measuring.
- **Estimated Difficulty Level:** Prior to field testing items, the item difficulties were not known, and writers could only make approximations as to how difficult an item might be. The estimated difficulty level was based upon the writer's own judgment as directly related to his or her classroom teaching and knowledge of the curriculum for a given subject area and grade level. The purpose for indicating estimated difficulty levels as items were written was to help

ensure that the pool of items would include a range of difficulty (easy, medium, and challenging).

- **Appropriate Grade Level, Item Context, and Assumed Student Knowledge:** Item writers were asked to consider the conceptual and cognitive level of each item. They were asked to review each item to determine whether or not the item was measuring something that was important and could be successfully taught and learned in the classroom.
- **Multiple-Choice (MC) Item Options and Distractor Rationale:** Writers were instructed to make sure that each item had only one clearly correct answer. Item writers submitted the answer key with the item. All distractors were plausible choices that represented common errors and misconceptions in student reasoning.
- **Face Validity and Distribution of Items Based Upon Depth of Knowledge:** Writers were asked to classify the depth of knowledge of each item, using a model based on Norman Webb's work on depth of knowledge (Webb, 2002). Items were classified as one of four depth of knowledge categories: recall, skill/concept, strategic thinking, and extended thinking.
- **Readability:** Writers were instructed to pay careful attention to the readability of each item to ensure that the focus was on the concepts; not on reading comprehension of the item. Resources writers used to verify the vocabulary level were the *EDL Core Vocabularies* (Taylor, Frackenpohl, White, Nieroroda, Browning, & Brisner, 1989) and the *Children's Writer's Word Book* (Mogilner, 1992). In addition, every test item was reviewed by grade-level experts. They reviewed each item from the perspective of the students they teach, and they determined the validity of the vocabulary used.
- **Grammar and Structure for Item Stems and Item Options:** All items were written to meet technical quality, including correct grammar, syntax, and usage in all items, as well as parallel construction and structure of text associated with each multiple-choice item.

Item Review: Throughout the item development process, independent panels of reading content experts reviewed the items. The following guidelines for reviewing assessment items were used during each review process.

A quality item should:

- have only one clear correct answer and contain answer choices that are reasonably parallel in length and structure;
- have a correctly assigned content code (item map);
- measure one main idea or problem;
- measure the objective or curriculum content standard it is designed to measure;
- be at the appropriate level of difficulty;
- be simple, direct, and free of ambiguity;
- make use of vocabulary and sentence structure that is appropriate to the grade level of the student being tested;
- be based on content that is accurate and current;

- when appropriate, contain stimulus material that are clear and concise and provide all information that is needed;
- when appropriate, contain graphics that are clearly labeled;
- contain answer choices that are plausible and reasonable in terms of the requirements of the question, as well as the students' level of knowledge;
- contain distractors that relate to the question and can be supported by a rationale;
- reflect current teaching and learning practices in the subject area; and
- be free of gender, ethnic, cultural, socioeconomic, and regional stereotyping bias.

Following each review process, the item writer group and the item review panel discussed suggestions for revisions related to each item. Items were revised only when both groups agreed on the proposed change.

Editorial Review of Items: After items were written and reviewed, Nebraska Department of Education test development specialists reviewed each item for item quality, making sure that the test items were in compliance with guidelines for clarity, style, accuracy, and appropriateness for Nebraska students. Additionally, DRC test development content experts worked collaboratively with NDE to review and revise the items prior to field testing to ensure highest level of quality possible.

Review of the Online Items: All items for online assessment were reviewed by the Nebraska Department of Education, Computerized Assessments and Learning (CAL), DRC's online partner, and DRC. In addition to DRC's standard review process to which all items are subjected, and to ensure comparability with paper and pencil versions, all items were reviewed for formatting and scrolling concerns.

Universally Designed Assessments: Universally designed assessments allow participation of the widest possible range of students and result in valid inferences about performance of all students who participate and are based on the premise that each child in school is a part of the population to be tested, and that testing results should not be affected by disability, gender, race, or English language ability (Thompson, Johnstone, & Thurlow, 2002). The Nebraska Department of Education and Data Recognition Corporation (DRC) are committed to the development of items and tests that are fair and valid for all students. At every stage of the item and test development process, procedures ensure that items and tests are designed and developed using the elements of universally designed assessments that were developed by the National Center on Educational Outcomes (NCEO).

Federal legislation addresses the need for universally designed assessments. The *No Child Left Behind Act* (Elementary and Secondary Education Act) requires that each state must "provide for the participation in [statewide] assessments of all students" [Section 1111(b)(3)(C)(ix)(I)]. Both Title 1 and IDEA regulations call for universally designed assessments that are accessible and valid for all students including students with disabilities and students with limited English proficiency. NDE and DRC recognize that the benefits of universally designed assessments not only apply to these groups of students, but to all individuals with wide ranging characteristics.

The NDE test development team and Nebraska item writers have been fully trained in the elements of Universal Design as it relates to developing large scale statewide assessments. Additionally, NDE and DRC partner to ensure that all items meet the Universal Design requirements during the item review process.

After a review of research relevant to the assessment development process and the principles of Universal Design (Center for Universal Design, 1997), NCEO has produced seven elements of Universal Design as they apply to assessments (Thompson, Johnstone, & Thurlow, 2002).

Inclusive Assessment Population

When tests are first conceptualized, they need to be thought of in the context of who will be tested. If the test is designed for state, district, or school accountability purposes, the target population must include every student except those who will participate in accountability through an alternate assessment. NDE and DRC are fully aware of increased demands that statewide assessment systems must include and be accountable for ALL students.

Precisely Defined Constructs

An important function of well-designed assessments is that they actually measure what they are intended to measure. NDE item writers and DRC carefully examine what is to be tested and design items that offer the greatest opportunity for success within those constructs. Just as universally designed architecture removes physical, sensory, and cognitive barriers to all types of people in public and private structures, universally designed assessments must remove all non-construct-oriented cognitive, sensory, emotional, and physical barriers.

Accessible, Non-biased Items

NDE conducts both internal and external review of items and test specifications to ensure that they do not create barriers because of lack of sensitivity to disability, cultural, or other subgroups. Items and test specifications are developed by a team of individuals who understand the varied characteristics of items that might create difficulties for any group of students. Accessibility is incorporated as a primary dimension of test specifications, so that accessibility is woven into the fabric of the test rather than being added after the fact.

Amenable to Accommodations

Even though items on niversally designed assessments will be accessible for most students, there will still be some students who continue to need accommodations. Thus, another essential element of any universally designed assessment is that it is compatible with accommodations and a variety of widely used adaptive equipment and assistive technology. NDE, DRC, and Computerized Assessment and Learning (CAL), DRC's online testing partner, work to ensure that state guidelines on the use of accommodations are compatible with the assessment being developed.

Simple, Clear, and Intuitive Instructions and Procedures

Assessment instructions should be easy to understand, regardless of a student's experience, knowledge, language skills, or current concentration level. Directions and questions need to be in simple, clear, and understandable language. Knowledge questions that are posed within complex language certainly invalidate the test if students cannot understand how they are expected to respond to a question.

Maximum Readability and Comprehensibility

A variety of guidelines exist to ensure that text is maximally readable and comprehensible. These features go beyond what is measured by readability formulas. Readability and comprehensibility are affected by many characteristics, including student background, sentence difficulty, organization of text, and others. All of these features are considered as NDE develops the text of assessments.

Plain language is a concept now being highlighted in research on assessments. Plain language has been defined as language that is straightforward and concise. The following strategies for editing text to produce plain language are used during NDE's editing process.

- Reduce excessive length.
- Use common words.
- Avoid ambiguous words.
- Avoid irregularly spelled words.
- Avoid proper names.
- Avoid inconsistent naming and graphic conventions.
- Avoid unclear signals about how to direct attention.
- Mark all questions.
- Maximum Legibility.

Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. Bias results when tests contain physical features that interfere with a student's focus on or understanding of the constructs that test items are intended to assess. DRC works closely with NDE to develop a style guide that includes dimensions of style that are consistent with universal design.

Depth of Knowledge: Interpreting and assigning depth of knowledge levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. Four levels of depth of knowledge are used for this analysis. The Nebraska State Accountability assessments include items written at levels 1, 2, and 3. Level 4 items are not included due to the test being comprised of only multiple-choice items.

Reading Level 1

Level 1 requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text as well as basic comprehension of a text is included. Items

require only a shallow understanding of text presented and often consist of verbatim recall from text or simple understanding of a single word or phrase. Some examples that represent but do not constitute all of Level 1 performance are:

- Support ideas by reference to details in the text.
- Use a dictionary to find the meaning of words.
- Identify figurative language in a reading passage.

Reading Level 2

Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Intersentence analysis of inference is required. Some important concepts are covered but not in a complex way. Standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 assessment item may require students to apply some of the skills and concepts that are covered in Level 1. Some examples that represent but do not constitute all of Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Reading Level 3

Deep knowledge becomes more of a focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent but do not constitute all of Level 3 performance are:

- Determine the author's purpose and describe how it affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Reading Level 4

Higher-order thinking is central and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with extended time provided. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. Students take information from at least one passage and are asked to apply this information to a new task. They may also be

asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent but do not constitute all of Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Mathematics Level 1

Level 1 includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include “identify,” “recall,” “recognize,” “use,” and “measure.” Verbs such as “describe” and “explain” could be classified at different levels depending on what is to be described and explained.

Mathematics Level 2

Level 2 includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as “explain,” “describe,” or “interpret” could be classified at different levels depending on the object of the action. For example, if an item required students to explain how light affects mass by indicating there is a relationship between light and heat, this is considered a Level 2. Interpreting information from a simple graph, requiring reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills, and such interpretation excludes from this level other skills such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Mathematics Level 3

Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include drawing conclusions from observations, citing evidence and developing a logical argument for concepts, explaining phenomena in terms of concepts, and using concepts to solve problems.

Mathematics Level 4

Level 4 requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student were to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments, making connections between a finding and related concepts and phenomena, combining and synthesizing ideas into new concepts, and critiquing experimental designs.

Science Level 1

Level 1 (Recall and Reproduction) requires the recall of information, such as a fact, definition, term, or a simple procedure, as well as performance of a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A “simple” procedure is well defined and typically involves only one step. Verbs such as “identify,” “recall,” “recognize,” “use,” “calculate,” and “measure” generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as “describe” and “explain” could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the item does not need to be “figured out” or “solved.” In other words, if the knowledge necessary to answer

an item automatically provides the answer to it, then the item is at Level 1. If the knowledge needed to answer the item is not automatically provided in the stem, the item is at least at Level 2. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Recall or recognize a fact, term, or property.
- Represent in words or diagrams a scientific concept or relationship.
- Provide or recognize a standard scientific representation for simple phenomenon.
- Perform a routine procedure, such as measuring length.

Science Level 2

Level 2 (Skills and Concepts) includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is **more complex** than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply **more than one step**. For example, to compare data requires first identifying characteristics of the objects or phenomena and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts. Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3. Some examples that represent, but do not constitute all of, Level 2 performance, are:

- Specify and explain the relationship between facts, terms, properties, or variables.
- Describe and explain examples and non-examples of science concepts.
- Select a procedure according to specified criteria and perform it.
- Formulate a routine problem, given data and conditions.
- Organize, represent, and interpret data.

Science Level 3

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation or a word or two should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a

Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent, but do not constitute all of Level 3 performance, are:

- Identify research questions and design investigations for a scientific problem.
- Solve non-routine problems.
- Develop a scientific model for a complex situation.
- Form conclusions from experimental data.

Science Level 4

Level 4 (Extended Thinking) involves high cognitive demands and complexity. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select or devise one approach among many alternatives to solve the problem. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. “Develop generalizations of the results obtained and the strategies used and apply them to new problem situations,” is an example of a grade 8 objective that is a Level 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought will be Level 4.

Level 4 requires complex reasoning, experimental design and planning, and probably will require an extended period of time either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4. Some examples that represent, but do not constitute all of, a Level 4 performance are:

- Based on data provided from a complex experiment that is novel to the student, deduct the fundamental relationship between several controlled variables.
- Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.

Source of Challenge Criterion

Source of Challenge criterion is only used to identify items where the major cognitive demand is inadvertently placed and is other than the targeted language arts skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source of challenge problem.

Such items' characteristics may cause some students to not answer an assessment item or answer an assessment item incorrectly or at a lower level even though they have the understanding and skills being assessed.

Item Content Review: Prior to field testing, all newly developed test passages/items were submitted to grade-level content committees for review. The content committees consisted of Nebraska educators from school districts throughout the state. The primary responsibility of the content committees was to evaluate items with regard to quality and content classification, including grade-level appropriateness, estimated difficulty, depth of knowledge, and source of challenge. They also suggested revisions, if appropriate. The committees also reviewed the items for adherence to the principles of universal design, including language demand and issues of bias, fairness, and sensitivity.

Item review committee members were selected by the Nebraska Department of Education. NDE test development team members facilitated the process. Training was provided by NDE and included how to review items for technical quality and content quality, including depth of knowledge and adherence to principles of universal design. In addition, training included providing committee members with the procedures for item review.

Committee members reviewed the items for quality and content, as well as for the following categories.

- Indicator (standard) Alignment
- Difficulty Level (classified as Low, Medium, or High)
- Depth of Knowledge (classified as Recall, Application, or Strategic Thinking)
- Correct Answer
- Quality of Graphics
- Appropriate Language Demand
- Freedom from Bias (classified as Yes or No)

Committee members were asked to flag items that needed revision and to denote suggested revisions on the flagged item cards.

Security was addressed by adhering to a strict set of procedures. Items in binders did not leave the meeting rooms and were accounted for at the end of each day before attendees were dismissed. All attendees, with the exception of NDE staff, were required to sign a Confidentiality Agreement (Appendix D).

Sensitivity and Bias Review: Prior to field testing items, all newly developed test items were submitted to a Bias and Sensitivity Committee for review. The committee's primary responsibility was to evaluate passages and items as to acceptability with regard to bias and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the area of concern. The bias/sensitivity committee was composed of Nebraska educators who represented the diversity of students. All committee members were trained by a Nebraska Department of Education test development lead to review items for bias and sensitivity issues using a Fairness in Testing training manual developed by Data Recognition Corporation (Appendix E).

All passages/items were read by all of the respective committee members. Each member noted bias and/or sensitivity comments on a review form. All comments were then compiled and the actions taken on these items were recorded by NDE. Committee members were required to sign a Confidentiality Agreement and strict security measures were in place to ensure that secure materials remained guarded (Appendix D).

2.6 Item Banking

DRC maintains an item bank (IDEAS) that provides a repository of item image, history, statistics, and usage. IDEAS includes a record of all newly created items together with item data from each item field test. It also includes all data from the operational administration of the items. Within IDEAS, DRC

- updates the Nebraska item bank after each administration;
- updates the Nebraska item bank with newly developed items;
- monitors the Nebraska item bank to ensure an appropriate balance of items aligned with content standards, goals, and objectives;
- monitors item history statistics; and
- monitors the Nebraska item bank for an appropriate balance of Depth of Knowledge (DOK) levels.

2.7 The Operational Form Construction Process

The Spring 2011 operational forms were constructed in Lincoln, Nebraska in August 2010 (Reading) and in September 2010 (Mathematics). The forms were constructed by NDE representatives and DRC content specialists. Training was provided by DRC for the forms construction process.

Prior to the construction of the operational forms, DRC Test Development content specialists reviewed the test blueprints to ensure that there was alignment between the items and the indicators, including the number of items per standard for each content-area test.

DRC Psychometricians provided Test Development specialists with an overview of the psychometric guidelines and targets for operational forms construction. The foremost guideline was for item content to match the test blueprint (Table of Specifications) for the given content. The point-biserial correlation guideline was to be greater than 0.3 (with a requirement for no point-biserial correlation less than zero). In addition, the average target p-value for each test was to be about 0.65. A Differential Item Functioning (DIF) code of C was to be avoided (unless no other items were available to fulfill a blueprint requirement). The overall summary of the actual approved p-value and biserial of the forms is provided in the summary table later in this document.

DRC Test Development specialists printed a copy of each item card, with accompanying item characteristics, image, and psychometric data. Test Development specialists verified the accuracy of each item card, making sure that the item image has its correct item characteristics. Test Development specialists carefully reviewed each item card's psychometric data to ensure it is complete and

reasonable. For Reading, the item cards (items and passages) were compiled in binders and sorted by p-values from highest to lowest by passage with associated items. For Mathematics, the item cards were compiled in binders and sorted by p-values from highest to lowest by standard and indicator.

NDE and DRC also checked to see that each item met technical quality for well-crafted items, including:

- only one correct answer,
- wording that is clear and concise,
- grammatical correctness,
- appropriate item complexity and cognitive demand,
 - appropriate range of difficulty,
 - appropriate depth-of-knowledge alignment,
- aligned with principles of Universal Design, and
- free of any content that might be offensive, inappropriate, or biased (content bias).

NDE representatives and DRC Test Development specialists made initial grade-level selections of the items (passages and items for Reading), known as the “pull list,” to be included on the 2011 operational forms. The goal was for the first pull of the items to meet the Table of Specification (TOS) guidelines and psychometric guidelines specific to each content. As items were selected, the unique item codes were entered into a form building template which contained the item pool with statistics and item characteristics. The template automatically calculated the P-value, biserial, number of items per indicator and standard, number of items per DOK level (1, 2, or 3), and distribution of answer key as items were selected for each grade. As items were selected, the item characteristics (key, DOK, and alignment to indicator) were verified.

Differential Item Functioning in Operational Form Construction: Differential Item Functioning (DIF) is present when the likelihood of success on an item is influenced by group membership. A pattern of such results may suggest the presence of, but does not prove, *item bias*. Actual item bias may present negative group stereotypes, may use language that is more familiar to one subpopulation than to another, or may present information in a format that disadvantages certain learning styles. While the source of item bias is often clear to trained judges, many instances of DIF may have no identifiable cause (resulting in false positives). As such, DIF is not used as a substitute for rigorous, hands-on reviews by content and bias specialists. Instead, DIF helps to organize the review of the instances in which bias is suggested. No items are automatically rejected simply because a statistical method flagged them or automatically accepted because they were not flagged.

During the operational form-pull process, the DIF code for every item proposed for use in the operational (core) is examined. To the greatest extent possible, the blueprint is met through the use of items with statistical DIF codes of A. Although DIF codes of B and C are not desirable and are deliberately avoided, the combination of the require blueprint and the depth of the available operational-ready item pool occasionally requires that items with B and C DIF are considered for operational use. In addition, for passage-based tests like reading (in which each item available in the

item pool is linked to a set of passage-based items), the ability to use a minimum number of items associated with a passage may require the use of an item with a B or C DIF code. In any case, prior to allowing exceptions of this nature, every attempt is made to re-craft the core to avoid the use of the item with B or C DIF. Before allowing any exception to be made, the item in question is examined to determine whether the suggested bias is identifiable. If the suggested bias is determined to be valid, the item is not used.

Review of the Items and Test Forms: At every stage of the test development process the match of the item to the content standard was reviewed and verified since establishing content validity is one of the most important aspects in the legal defensibility of a test. As a result, it is essential that an item selected for a form link directly to the content curriculum standard and performance standard to which it is measuring. Test development specialists verified all items against their classification codes and item maps, both to evaluate the correctness of the classification and to ensure that the given task measures what it purports to measure.

2.8 READING ASSESSMENT

Test Design: The NeSA-Reading operational test includes operational passages with associated items and one field test passage with associated items. This test was administered online via the test engine developed and managed by CAL, DRC’s online testing partner. One form of the test was also published in a printed test booklet for schools that did not have students participating in the online system. Depending on grade, the forms contained 45 to 50 operational items.

Table 2.8.1 Reading 2011 Operational Test

Grade	Total No. of MC Core Items	No. of Embedded FT Items per Form (1 passage)	Total Items per Form	Total No. of Equivalent FT Forms	Total Core Points	Total No. of MC Items Added to the Bank
3	45	10	55	5	45	50
4	45	10	55	5	45	50
5	48	10	58	5	48	50
6	48	10	58	5	48	50
7	48	10	58	5	48	50
8	50	10	60	5	50	50
11	50	10	60	5	50	50

Psychometric Targets: The goal for the operational forms was to meet a mean p-values of approximately 0.65 with values restricted to the range of 0.30 to 0.90 and point-biserial correlations greater than 0.25, based on previous field test results. However, these targets are secondary to

constructing the best test possible. Some compromises were allowed when necessary to best meet the objective of the assessment, to conform to the test specifications, and to operate within the limitations of the item bank.

Equating Design: Spring 2011 was the second operational administration of NeSA-R. Approximately 70% of the assessment was constructed from passages and related items field tested in Spring 2010. The approximate remaining 30% of the assessment was constructed from an overlap of items and passages from the 2010 operational (core) item positions from the Spring 2010 operational forms.

In addition to the operational passage sets, each student received one randomly selected field test passage with items. The passages and items taken by each student were administered in two testing sessions each intended to be administered in a single class period. The operational passages were administered to the student in a random order, but the field test passage was maintained in a fixed position. Items within a passage were administered in a fixed order for the passage. Equating was accomplished by anchoring on the operational passage items and calibrating the field test items concurrently.

2.9 MATHEMATICS ASSESSMENT

Test Design: The NeSA-Mathematics operational test includes operational and field test items. This test was administered online via the test engine developed and managed by CAL. One form of the test was also published in a printed test booklet for schools that did not have students participating in the online system. Depending on grade, the forms contained 50 to 60 operational items.

Table 2.9.1 Mathematics 2011 Operational Test

Grade	Total No. of MC Core Items	No. of Embedded FT Items per Form	Total Items per Form	Total No. of Equivalent FT Forms	Total Core Points	Total No. of MC Items Added to the Bank
3	50	10	60	5	50	50
4	55	10	65	5	55	50
5	55	10	65	5	55	50
6	58	10	68	5	58	50
7	58	10	68	5	58	50
8	60	10	70	5	60	50
11	60	10	70	5	60	50

Psychometric Targets: The goal for the operational forms was to meet a mean p-values of approximately 0.65 with values restricted to the range of 0.3 to 0.9 and point-biserial correlations

greater than 0.25, based on previous field test results. However, these targets are secondary to constructing the best test possible. Some compromises were allowed when necessary to best meet the objective of the assessment, to conform to the test specifications, and to operate within the limitations of the item bank.

Equating Design: Spring 2011 was the first operational administration of NeSA-M. The assessment was constructed from items field tested in Spring 2010. While preliminary item parameter estimates were available from the field test, the operational data were used for the final estimates; no equating was necessary.

In addition to the operational items, each student received 10 randomly selected field test items. The items taken by each student were administered in two testing sessions each intended to be administered in a single class period. The operational items were administered to the student in a random order, but the field test items were maintained in fixed positions. Equating was accomplished by anchoring on the operational items and calibrating the field test items concurrently.

2.10 SCIENCE ASSESSMENT

Initial Standalone Field Test: The main purpose of the 2011 NeSA-Science Field Test was to collect data for item screening and parameter calibration. This is critical to ensuring a large item pool from which operational forms can be constructed. Errors in the field test form-construction phase can result in a depleted item pool or a mis-estimation of item parameters that perpetuates throughout the form-construction process. The standalone Spring 2011 Science Field Test forms were constructed in Lincoln, Nebraska in September 2010.

Forms Assembly: The field test forms were constructed from the items in the field test item pool. Items from this pool were selected to meet the requirements described in the test specifications. Subject to the constraints of the pool, the forms were constructed according to the accepted standards of content balance and difficulty with the intent that the forms be as parallel as practical.

Table 2.10.1 Science Standalone Field Test (2011)

Grade	Total Items per Form	Total No. of Equivalent FT Forms	Total No. of FT Items
5	50	3	141
8	60	3	163
11	60	4	213

Form Approval Meeting: The items and forms for the standalone Spring 2011 Science Field Test were reviewed and approved by the NDE staff in collaboration with DRC science content specialists and the project lead in September 2010 in Lincoln, Nebraska. The items were reviewed for technical quality, alignment to indicator, and adherence to style guide formats.

[Equating Design](#): The field tests were administered online with each student receiving a random selection of items administered in a random order. This process ensures a randomly equivalent sample receiving each item and permits the concurrent calibration of all items. The result is a common calibration and equated item difficulties for all field test items.

3. Reading and Mathematics Operational Assessment

3.1 RASCH CALIBRATION AND EQUATING

Calibration of NeSA was accomplished with *Winsteps version 3.71.00* (Linacre, 2011). This provided the final estimates of the item logit difficulties for the reading and mathematics operational items. These estimates were the basis for the standard setting and scale definition to be used throughout the duration of the program. The first calibration run established the parameter estimates for the operational items without interference from the newly written field test items. Once the difficulties for the operational items were obtained, they were used as *anchors* to evaluate and equate the field test items to the operational metric. This was accomplished by using the anchor difficulties to define the metric and obtain estimates for the unanchored (field test) items in that metric. The results are estimated difficulties relative to the anchors and are, hence, scaled to the existing metric. The final reading values can be viewed in Appendix P and the final mathematics values can be viewed in Appendix Q. For summary demographic breakdowns for reading and mathematics please see Appendix T.

An overview of Rasch Measurement Models is provided in Appendix F as well as in several of the references (see, for example, Wright & Stone, 1979).

3.2 Validity and Reliability

Items: For criterion-referenced, standards-based assessment, the strongest validity evidence is derived directly from the test construction process and the item scaling. The item development and test construction process, described above, ensures that every item aligns directly to one of the content standards. This alignment is foremost in the minds of the item writers and editors. As a routine part of item selection prior to an item appearing on a test form, the review committees check the alignment of the items with the standards and make any adjustments necessary. The result is consensus among the content specialists and teachers that the assessment does in fact assess what was intended.

The empirical item scaling, which indicates where each item falls on the logit ability-difficulty continuum, should be consistent with what theory suggests about the items. Items that require more knowledge, more advanced skills, and more complex behaviors should be empirically more difficult than those requiring less. Evidence of this agreement is contained in the item summary tables in Appendix G and H, as well as the success of the Bookmark standard setting process (in the separate *2010 NeSA-R Standard Setting Technical Report and 2011 NeSA-M Standard Setting Technical Report*). Panelists participating in the Bookmark process work from an item booklet in which items are ordered by their empirical difficulties. Discussions about placement of the bookmarks almost invariably focus on the knowledge, skills, and behaviors required of each item, and, overall, panelists were comfortable with the item ordering and spacing.

Items Analyses: Traditional item analysis is a straightforward approach to examining the quality of the items that is rooted in true score theory. Although these are sample-specific statistics, they are entirely adequate for assessing the effectiveness of items in this context. The statistics provide information about the quality of the items based on student responses in an operational setting. The following sections provide descriptions of the item summary statistics found in Appendices G and H.

Item Difficulty: (p -value) is the percent of examinees in the sample who answered the item correctly. Typically, test developers target p -values in the range of 0.30 to 0.90. Mathematically, information is maximized and standard errors minimized when the p -value equals 0.50. Experience suggests that multiple choice items are effective when the student is more likely to succeed than fail and it is important to include a range of difficulties matching the distribution of student abilities (Wright & Stone, 1979). Occasionally, items that fall outside the desired range can be justified for inclusion when the educational importance of the item content or the desire to measure students with very high or low achievement override the statistical considerations.

Table 3.2.1: Summary of Traditional Item Percent Correct for NeSA-R Operational Items

Grade	Item Percent Correct										Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	
3	0	0	0	1	3	6	12	14	9	0	45
4	0	0	1	1	2	9	9	16	7	0	45
5	0	0	0	0	6	9	9	15	8	1	48
6	0	0	0	3	2	6	11	14	10	2	48
7	0	0	0	0	4	10	9	14	11	0	48
8	0	0	0	0	4	12	15	12	6	1	50
11	0	0	0	2	2	16	7	16	7	0	50

Table 3.2.2: Summary of Traditional Item Percent Correct for NeSA-M Operational Items

Grade	Item Percent Correct										Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	
3	0	0	0	0	2	5	14	11	17	1	50
4	0	0	0	0	3	7	7	19	14	5	55
5	0	0	0	0	1	8	10	16	19	1	55
6	0	0	0	0	2	5	13	15	21	2	58
7	0	0	0	1	5	10	7	20	13	2	58
8	0	0	0	0	0	7	21	22	9	1	60
11	0	0	0	0	4	15	26	11	3	1	60

Percent selecting each response option indicates the effectiveness of each distractor. In general, one expects the correct response to be the most attractive, although this need not hold for unusually

challenging items. This statistic for the correct response option is identical to the p -value when considering multiple-choice items with a single correct response.

Item-total correlation describes the relationship between performance on the specific item and performance on the entire form. Total test score is the best available indicator of proficiency; success on individual items should correlate with success on the total test. For multiple-choice items, the statistic is the *point-biserial correlation*, which is a special case of the Pearson product moment correlation for the keyed correct response with total test score. Items with negative correlations are flagged and referred to Test Development as possible mis-keys. Mis-keyed items will be corrected and rescored prior to computing the final item statistics. Negative correlations can also indicate problems with the item content, structure, or students' opportunity to learn. Items with point-biserial values of less than 0.2 were flagged and referred to content specialists for review before being considered for use on future forms. As seen below, no items had negative point-biserial correlations.

Table 3.2.3 Summary of Point-biserial Correlations for NeSA-R

Grade	Item Point-biserial Correlation							Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	
3	0	0	11	15	18	1	0	45
4	0	2	12	21	10	0	0	45
5	0	3	2	27	13	3	0	48
6	0	3	8	24	12	1	0	48
7	0	0	6	22	17	3	0	48
8	0	0	7	22	21	0	0	50
11	0	0	12	12	22	4	0	50

Table 3.2.4 Summary of Point-biserial Correlations for NeSA-M

Grade	Item Point-biserial Correlation							Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	
3	0	0	5	21	23	1	0	50
4	0	0	5	28	21	1	0	55
5	0	0	4	18	29	4	0	55
6	0	0	4	16	33	5	0	58
7	0	0	5	17	29	7	0	58
8	0	0	2	13	38	7	0	60
11	0	0	3	9	38	10	0	60

Point-biserial correlations of response options describe the relationship between selecting a response option for a specific item and performance on the entire test. They can be interpreted as the standardized mean score of examinees selecting the response. The correlation between an incorrect answer and total test performance should be negative. The desired pattern is strong positive values for

the correct option and strong negative values for the incorrect options. Any other pattern indicates a problem with the item or with the key. These patterns would imply a high ability way to answer incorrectly or a low ability way to answer correctly. Examples of these situations could be an item with an ambiguous or misleading distractor that was attractive to high-performing examinees or an item that depended on experience outside of instruction that was unrelated to ability.

This statistic for the correct option is identical to the item-total correlation for multiple-choice items.

Percent of students omitting an item is useful for identifying problems with testing time and test layout. When the pattern of omits increases at the end of a timed section, there may not have been sufficient time for students to complete all items. Alternatively, if the omit percentage is large for a single item, it could indicate a problem with the layout or content of an item. For example, students tend to skip items with wordy stems or that otherwise appear difficult or time consuming. While there is no hard and fast rule for what *large* means, and it varies with groups and ages of students, five percent omits is often used as a preliminary screening value.

Detailed results of the item analyses for the NeSA-R operational items are presented in Appendices G and M. Detailed results of the item analyses for the NeSA-M operational items are presented in Appendices H and M. Based on these analyses, items were selected for review if the *p*-value was less than 0.25 and the *item-total correlation* was less than 0.2. Items were identified as probable mis-keys if the *p*-value for the correct response was less than one of the incorrect responses and the *item-total correlation* was negative. No items on the NeSA-R were miskeyed.

Differential item functioning (DIF) is defined as the situation in which the likelihood of success on an item is partially predicted by group membership. Operationally, this is computed as the difference in the likelihood of success for examinees with the same level of proficiency but who were members of different sub-groups. DIF can occur if the item involves factors that differentially advantage or disadvantage specific groups of students. Items exhibiting DIF were referred to content specialists to determine possible bias.

Within the context of the Rasch measurement models, DIF is a direct violation of the model requirements that the probability of success depends only on item difficulty and person ability. Hence, DIF analysis is a natural consequence of Rasch analysis. The Winsteps software was used to compute DIF statistics that directly compare group performance on the items after adjusting for any differences in the ability distributions of the examinees. Items with DIF codes of C (significance level less than 0.01) or B (significance level less than 0.1) were flagged for review by content and bias specialists, with emphasis on items that disadvantage the focal group. The level depends on the magnitude of the discrepancy between the groups of interest and the likelihood it could arise by chance. Large group sizes and equal numbers result in a very sensitive test. Table 3.2.5 shows a summary of the DIF statistics. The plus and minus codes on the B and C indicates which group is favored. Plus means the

focal group was favored; minus means the focal group was disadvantaged. Detailed analyses are included in Appendix K. The first column indicates the focal group.

Table 3.2.5: Summary of NeSA-M Differential Item Functioning by Code

Grade 3	A	B+	B-	C+	C-	Total Items
Female	50	0	0	0	0	50
Black	48	0	2	0	0	50
Hispanic	49	1	0	0	0	50
Native American	43	0	7	0	0	50
Asian	41	2	6	0	1	50
Multiple	50	0	0	0	0	50

Grade 4	A	B+	B-	C+	C-	Total Items
Female	55	0	0	0	0	55
Black	53	0	2	0	0	55
Hispanic	52	1	2	0	0	55
Native American	50	0	5	0	0	55
Asian	47	0	6	0	2	55
Multiple	55	0	0	0	0	55

Grade 5	A	B+	B-	C+	C-	Total Items
Female	53	1	1	0	0	55
Black	52	0	3	0	0	55
Hispanic	55	0	0	0	0	55
Native American	51	0	4	0	0	55
Asian	43	4	6	1	1	55
Multiple	54	1	0	0	0	55

Grade 6	A	B+	B-	C+	C-	Total Items
Female	56	0	2	0	0	58
Black	54	1	3	0	0	58
Hispanic	58	0	0	0	0	58
Native American	50	0	8	0	0	58
Asian	55	2	1	0	0	58
Multiple	58	0	0	0	0	58

Grade 7	A	B+	B-	C+	C-	Total Items
Female	55	2	1	0	0	58
Black	57	0	1	0	0	58
Hispanic	58	0	0	0	0	58
Native American	57	0	1	0	0	58
Asian	57	0	1	0	0	58
Multiple	56	0	2	0	0	58

Grade 8	A	B+	B-	C+	C-	Total Items
Female	59	0	1	0	0	60
Black	59	0	1	0	0	60
Hispanic	60	0	0	0	0	60
Native American	57	0	3	0	0	60
Asian	49	3	5	1	2	60
Multiple	60	0	0	0	0	60

Grade 11	A	B+	B-	C+	C-	Total Items
Female	57	2	0	0	1	60
Black	58	0	2	0	0	60
Hispanic	60	0	0	0	0	60
Native American	59	0	1	0	0	60
Asian	52	2	4	1	1	60
Multiple	59	1	0	0	0	60

Forms Performance Summary: The NeSA-R operational forms contained five passages for all grades and a total of 45 to 50 items, depending on the grade, as shown in Table 2.8.1. The passages were administered online in a random order with the items in a fixed order within each passage. The percent correct means and traditional form reliabilities for NeSA-R are shown in Table 3.2.6 and 3.2.7 for NeSA-M. More detail on the performance of the forms is given in Appendix M.

Table 3.2.6: 2011 NeSA-R Form Summary

Grade	Mean Percent Correct	Form Reliability
3	68.6	0.885
4	67.9	0.862
5	67.8	0.889
6	69.3	0.880
7	68.5	0.897
8	66.8	0.897
11	65.9	0.900

Table 3.2.7: 2011 NeSA-M Form Summary

Grade	Mean Percent Correct	Form Reliability
3	73.2	0.913
4	74.3	0.914
5	73.5	0.925
6	74.4	0.929
7	70.2	0.929
8	71.9	0.937
11	63.9	0.941

Tables 3.2.8 and 3.2.9 provide more detail on the performance of the content area assessments by subgroup. Mean percent correct are typical of the group historical performances. The form reliabilities were on the order of 0.90 for all groups, with none below 0.85, which is often cited as the acceptable level for this type of data.

Nebraska State Accountability Technical Report 2011

Table 3.2.8: 2011 NeSA-R Reliability Subgroup Form Summary

Reading	Grade	3			4			5			6			7		
		Reliability	Mean	Std Dev												
Ethnicity*	AM	0.88	23.6	8.7	0.85	24.5	7.9	0.87	25.4	8.8	0.89	27.0	9.1	0.91	26.5	9.9
	AS	0.93	32.4	9.7	0.91	31.2	9.0	0.91	33.7	9.5	0.91	35.1	9.0	0.93	33.9	10.1
	BL	0.87	26.5	8.2	0.86	25.3	8.0	0.88	26.7	9.1	0.89	28.5	9.0	0.90	25.9	9.7
	PI	0.91	30.6	9.1	0.79	28.4	6.5	0.88	26.9	8.7	0.86	30.4	8.1	0.91	29.0	9.8
	WH	0.88	32.3	7.7	0.85	31.9	7.2	0.88	33.9	8.1	0.87	34.5	7.8	0.89	34.3	8.3
	HI	0.87	26.9	8.1	0.84	27.3	7.4	0.88	29.2	8.7	0.87	26.6	8.3	0.89	28.7	8.9
	MU	0.88	30.7	8.1	0.86	30.2	7.6	0.89	32.2	8.9	0.88	33.0	8.2	0.90	31.1	9.4
Gender	Male	0.89	30.2	8.5	0.87	30.1	7.8	0.89	31.9	8.8	0.89	32.5	8.5	0.90	31.9	9.3
	Female	0.88	31.4	8.0	0.86	31.0	7.5	0.89	33.1	8.6	0.88	33.9	8.0	0.90	33.6	8.7
Free/Reduced	Yes	0.87	28.0	8.2	0.85	27.8	7.6	0.88	29.5	8.8	0.88	30.3	8.5	0.89	29.2	9.2
	No	0.87	33.2	7.5	0.85	32.8	6.9	0.88	34.9	7.8	0.87	35.5	7.4	0.88	35.4	7.9
ELL	Yes	0.84	24.7	7.7	0.81	25.2	7.0	0.84	26.0	7.9	0.83	25.5	7.6	0.83	23.5	7.8
	No	0.88	31.4	8.1	0.86	31.0	7.6	0.89	32.9	8.6	0.88	33.6	8.1	0.90	33.1	8.9
SPED	Yes	0.89	26.4	8.8	0.87	25.8	8.3	0.89	26.2	9.3	0.88	26.2	8.9	0.89	24.3	9.2
	No	0.88	31.6	7.9	0.85	31.3	7.3	0.88	33.6	8.1	0.87	34.3	7.6	0.89	34.1	8.3

*AM=American Indian, AS=Asian, BL=African American/Black, HI= Hispanic, MU=Multiple Ethnicities, PI=Pacific Islander, WH=White

Nebraska State Accountability Technical Report 2011

Reading	Grade	8			11		
		Reliability	Mean	Std Dev	Reliability	Mean	Std Dev
Ethnicity	AM	0.88	27.0	9.4	0.91	26.9	10.3
	AS	0.93	33.5	10.7	0.93	31.6	11.0
	BL	0.89	26.4	9.7	0.90	25.5	10.0
	PI	0.88	35.8	8.4	0.91	30.5	10.2
	WH	0.89	35.2	8.6	0.89	34.4	8.8
	HI	0.89	28.2	9.5	0.90	27.7	9.8
	MU	0.89	31.3	9.4	0.90	31.1	9.7
Gender	Male	0.90	32.3	9.6	0.91	31.9	9.9
	Female	0.90	34.3	9.2	0.90	33.7	9.2
Free/ Reduced	Yes	0.89	29.3	9.5	0.90	28.6	9.8
	No	0.88	36.1	8.3	0.89	34.9	8.8
ELL	Yes	0.84	21.8	8.2	0.83	20.0	7.8
	No	0.90	33.7	9.2	0.90	33.1	9.5
SPED	Yes	0.87	24.3	9.0	0.87	23.2	9.0
	No	0.89	34.7	8.8	0.90	34.0	9.0

*AM=American Indian, AS=Asian, BL=African American/Black, HI= Hispanic, MU=Multiple Ethnicities, PI=Pacific Islander, WH=White

Nebraska State Accountability Technical Report 2011

Table 3.2.9: 2011 NeSA-M Reliability Subgroup Form Summary

Math	Grade	3			4			5			6			7		
		Reliability	Mean	Std Dev												
Ethnicity	AM	0.92	27.9	10.6	0.91	31.8	11.0	0.92	31.7	11.6	0.94	32.2	13.1	0.92	32.3	12.0
	AS	0.94	38.7	10.2	0.94	42.5	10.5	0.94	42.8	10.5	0.94	46.6	10.7	0.95	42.3	12.3
	BL	0.90	29.8	9.9	0.90	33.1	10.2	0.92	31.2	11.2	0.92	34.2	11.5	0.90	30.7	11.0
	PI	0.94	35.2	11.1	0.85	39.5	7.8	0.86	38.3	8.3	0.93	42.7	11.4	0.92	35.1	11.7
	WH	0.90	38.4	8.4	0.90	42.7	8.7	0.91	42.3	9.4	0.92	45.1	9.9	0.92	43.0	10.2
	HI	0.90	32.4	9.4	0.90	37.2	9.7	0.91	36.6	10.5	0.92	38.9	11.0	0.92	35.2	11.3
	MU	0.90	36.3	9.0	0.91	40.2	9.7	0.92	40.0	10.2	0.93	42.4	10.9	0.93	38.3	11.8
Gender	Male	0.92	36.9	9.5	0.92	40.9	9.8	0.93	40.4	11.5	0.93	43.1	11.2	0.93	40.8	11.6
	Female	0.91	36.3	9.1	0.91	40.8	9.4	0.92	40.4	10.3	0.92	43.3	10.6	0.92	40.6	10.9
Free/ Reduced	Yes	0.91	33.4	9.5	0.91	37.6	9.9	0.92	36.9	10.9	0.93	39.4	11.4	0.92	36.2	11.5
	No	0.90	39.4	8.1	0.90	43.6	8.4	0.91	43.3	9.1	0.92	46.2	9.4	0.92	44.1	9.8
ELL	Yes	0.89	30.2	9.3	0.89	35.1	9.6	0.90	33.8	10.1	0.90	34.5	10.5	0.88	30.6	10.1
	No	0.91	37.3	9.0	0.91	41.4	9.4	0.92	40.9	10.3	0.93	43.7	10.7	0.93	41.2	11.1
SPED	Yes	0.92	31.9	10.6	0.91	35.3	10.6	0.92	33.2	11.5	0.93	33.9	12.1	0.91	30.6	11.3
	No	0.91	37.4	8.8	0.91	41.9	9.1	0.91	41.7	9.6	0.92	44.7	9.9	0.92	42.3	10.4

*AM=American Indian, AS=Asian, BL=African American/Black, HI= Hispanic, MU=Multiple Ethnicities, PI=Pacific Islander, WH=White

Nebraska State Accountability Technical Report 2011

Math	Grade	8			11		
		Reliability	Mean	Std Dev	Reliability	Mean	Std Dev
Ethnicity	AM	0.94	34.3	13.2	0.93	29.4	13.25
	AS	0.95	45.5	13.0	0.95	41.5	14.02
	BL	0.92	32.5	12.1	0.91	26.6	11.44
	PI	0.89	48.1	8.5	0.93	37.7	12.36
	WH	0.93	45.6	10.8	0.94	40.8	12.50
	HI	0.93	37.0	12.2	0.92	30.9	12.07
	MU	0.93	39.5	12.4	0.94	34.5	13.31
Gender	Male	0.94	42.8	12.4	0.94	38.4	13.59
	Female	0.93	43.5	11.7	0.94	38.3	12.98
Free/ Reduced	Yes	0.93	38.1	12.4	0.93	30.8	12.74
	No	0.93	46.7	10.5	0.93	32.2	12.63
ELL	Yes	0.92	32.0	11.8	0.86	24.9	9.52
	No	0.94	43.5	11.9	0.94	38.7	13.21
SPED	Yes	0.92	31.9	12.0	0.89	26.0	10.59
	No	0.93	44.8	11.2	0.94	39.9	12.79

*AM=American Indian, AS=Asian, BL=African American/Black, HI= Hispanic, MU=Multiple Ethnicities, PI=Pacific Islander, WH=White

State Performance Summary: Complete frequency distributions for the NeSA-R and NeSA-M are provided in Appendix O as part of the raw-to-scale score conversion tables. A simple summary of the reading and mathematics distributions can be found in Table 3.2.10 and 3.2.11. While the distribution appears consistent across grades, there was no attempt at longitudinal equating beyond the articulation of the performance level definitions done in conjunction with standard setting. This is described briefly in Section 3.3 and in detail in the separate *2010 NeSA-R Standard Setting Technical Report and 2011 NeSA-M Standard Setting Technical Report*.

Table 3.2.10: 2011 NeSA-R State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	21852	104.3	31.5	81	103	123
4	21545	109.0	35.2	86	111	130
5	21328	107.7	41.3	80	108	134
6	20805	108.9	38.5	83	111	138
7	20652	110.5	41.3	82	109	140
8	20516	106.2	38.5	79	107	133
11	20896	102.6	41.5	75	106	129

Table 3.2.11: 2011 NeSA-M State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	21921	103.5	37.1	79	104	127
4	21598	102.6	35.3	79	102	127
5	21384	102.7	38.2	76	102	128
6	20857	100.5	40.4	71	100	127
7	20690	98.8	38.6	72	95	123
8	20544	98.0	40.0	70	97	123
11	20822	95.5	46.3	58	90	127

For NeSA-R, between 16% and 21% of students took the assessment in the paper-based version with the lower percentages occurring in middle schools. Table 3.2.12 provides counts of the numbers tested in each mode and the percent testing with paper.

Table 3.2.12: 2011 NeSA-R Number of Students Tested

Grade	Total	Online	Paper	Percent Paper
3	21852	17537	4315	20
4	21545	17430	4115	19
5	21328	17516	3812	18
6	20805	16512	4293	21
7	20652	16572	4080	20
8	20516	16577	3939	19
11	20896	17572	3324	16

For NeSA-M between 39% and 47% of students took the assessment in the paper-based version. Table 3.2.13 provides counts of the numbers tested in each mode and the percent testing with paper.

Table 3.2.13: 2011 NeSA-M Number of Students Tested

Grade	Total	Online	Paper	Percent Paper
3	21921	13199	8722	40
4	21598	13079	8519	39
5	21384	12919	8465	40
6	20857	11868	8989	43
7	20690	12104	8586	41
8	20544	12041	8503	41
11	20822	11129	9693	47

Decision Consistency: In a standards-based testing program, there is great interest in how accurately students are classified into achievement categories. Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test (Huynh, 1976). If two equivalent forms were given to the same students, the consistency of the measure would be reflected by the extent that the classification decisions made from the first set of test scores matched the decisions based on the second set of test scores. In contrast to Coefficient Alpha, which describes the relative ordering of students, it is the actual student scores that are important in decision consistency.

Table 3.2.14. Pseudo-Decision Table for Two Hypothetical Categories

		TEST ONE		
		LEVEL I	LEVEL II	MARGINAL
TEST TWO	LEVEL I	Φ_{11}	Φ_{12}	$\Phi_{1\bullet}$
	LEVEL II	Φ_{21}	Φ_{22}	$\Phi_{2\bullet}$
	MARGINAL	$\Phi_{\bullet 1}$	$\Phi_{\bullet 2}$	1

Table 3.2.15. Pseudo-Decision Table for Four Hypothetical Categories

		TEST ONE				
		LEVEL I	LEVEL II	LEVEL III	LEVEL IV	MARGINAL
TEST TWO	LEVEL I	Φ_{11}	Φ_{12}	Φ_{13}	Φ_{14}	$\Phi_{1\bullet}$
	LEVEL II	Φ_{21}	Φ_{22}	Φ_{23}	Φ_{24}	$\Phi_{2\bullet}$
	LEVEL III	Φ_{31}	Φ_{32}	Φ_{33}	Φ_{34}	$\Phi_{3\bullet}$
	LEVEL IV	Φ_{41}	Φ_{42}	Φ_{43}	Φ_{44}	$\Phi_{4\bullet}$
	MARGINAL	$\Phi_{\bullet 1}$	$\Phi_{\bullet 2}$	$\Phi_{\bullet 3}$	$\Phi_{\bullet 4}$	1

If a student is classified as being in one category based on Test One’s score, how probable would it be that the student would be classified in the same category based on Test Two?

The proportions of correct decisions, ϕ for two and four categories are computed by the following two formulas, respectively:

$$\begin{aligned}\phi &= \phi_{11} + \phi_{22} \\ \phi &= \phi_{11} + \phi_{22} + \phi_{33} + \phi_{44}.\end{aligned}$$

It is the proportion of students classified by the two forms into exactly the same achievement level that represents the overall consistency.

Since it is not possible to retest in order to estimate the proportion of students who would be reclassified in the same performance levels, a statistical model needs to be imposed on the data in order to project the consistency of classifications solely using data from the available administration (Hambleton & Novick, 1973). Although a number of procedures are available, two well-known methods were developed by Hanson and Brennan (1990) and Livingston and Lewis (1995) utilizing specific True Score Models.

RESULTS AND OBSERVATIONS

Several factors might affect decision consistency. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications. Another factor is the location of the cutscore in the score distribution. More consistent classifications are observed when the cutscores are located away from the mass of the score distribution. The number of performance levels is also a consideration. Consistency indices for four performance levels should be lower than those based on two categories because classification using four levels would allow more opportunity to change achievement levels. Finally, some research has found that results from the Hanson and Brennan (1990) method on a dichotomized version of a complex assessment yields similar results to the Livingston and Lewis method (1995) and the method by Smith and Stearns (Stearns & Smith, 2007).

Across all grades, the overall decision consistencies were around 0.90, with only trivial differences between the algorithms. Consistency around the Exceeds the Standards cut score tended to be lower than around the Meets the Standards cutscore, reflecting the higher standard errors for the more extreme scores. The tables below provide the results for each grade and cutscore for both algorithms. The tables also distinguish between Decision Consistency and Decision Accuracy.

Decision Consistency: *the degree of agreement between two classifications based on non-overlapping, equally difficult forms of the test.* This is the agreement between two independent, observable but imperfect classification decisions. It is analogous to test-retest reliability. It is an index of how consistent the classification would be if the student could be tested again without contamination from the first testing. Both classifications would involve measurement error.

Decision Accuracy: *the degree of agreement between actual classification, based on the single-form score, with the classification that would be made on the basis of the true scores?* This is the agreement between the observed classification and the unobservable *true* classification. While the observed classification would involve measurement error, the *true* classification would not.

Table 3.2.16 NeSA-R Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Proficient	Advanced	Proficient	Advanced	Proficient	Advanced	Proficient	Advanced
Reading	3	0.92	0.91	0.89	0.87	0.92	0.91	0.89	0.88
	4	0.92	0.89	0.89	0.84	0.92	0.89	0.89	0.85
	5	0.92	0.90	0.89	0.86	0.92	0.90	0.89	0.86
	6	0.92	0.89	0.89	0.84	0.92	0.89	0.90	0.85
	7	0.93	0.90	0.90	0.86	0.93	0.90	0.90	0.86
	8	0.92	0.91	0.89	0.87	0.92	0.91	0.90	0.87
	11	0.92	0.90	0.89	0.86	0.92	0.90	0.89	0.87

Table 3.2.17 NeSA-M Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Proficient	Advanced	Proficient	Advanced	Proficient	Advanced	Proficient	Advanced
Math	3	0.93	0.91	0.90	0.87	0.93	0.91	0.90	0.88
	4	0.93	0.92	0.90	0.89	0.93	0.92	0.90	0.89
	5	0.93	0.92	0.90	0.88	0.93	0.92	0.90	0.89
	6	0.93	0.92	0.90	0.89	0.93	0.92	0.90	0.89
	7	0.93	0.93	0.90	0.91	0.93	0.93	0.90	0.91
	8	0.94	0.92	0.91	0.89	0.94	0.93	0.91	0.90
	11	0.94	0.94	0.91	0.92	0.94	0.94	0.91	0.92

3.3 Setting Performance Standards

In spring and summer 2011 standard setting and contrasting groups events took place for NeSA Mathematics. NeSA Reading was phased in a year earlier in 2010. Complete documentation of the 2011 mathematics standard setting and standards validation events are presented in a separate document called *2011 NeSA-Mathematics Standard Setting Technical Report*.

Academic Performance Levels for the mathematics component of the Nebraska State Accountability assessments (NeSA-Mathematics) were developed in spring 2011 by establishing cut scores that define operationally the three Performance Levels: *Below the Standards*, *Meets the Standards*, *Exceeds the Standards*. These Performance Level designations will be used by local, state, and federal accountability programs and are central to communicating to parents, teachers and the public. The *Meets the Standards* and *Exceeds the Standards* levels are used for the *No Child Left Behind (NCLB)* Adequate Yearly Progress (AYP) proficiency goal.

The larger process comprised four events. First, a meeting was held February 28, 2011, with the Nebraska State Board of Education and other stakeholders to introduce the process and obtain feedback to ensure an effective, defensible process. Second, a *Contrasting Groups* survey of mathematics specialists and teachers was conducted in spring 2011 to obtain the teachers' overall perception of the proficiency level of their own students, independent of the state assessment. Third, a *Bookmark Standard Setting* was conducted June 27–29, 2011 in Lincoln, Nebraska, after the operational data were available. Finally, recommendations of the *Contrasting Groups* and *Bookmark* processes were presented to the State Board of Education July 12–13, 2011. The purpose of this meeting was for the State Board of Education to formally establish the Performance Levels. This report specifically documents the *Bookmark* and *Contrasting Groups* portions of the process.

The *Bookmark* method (Lewis, Mitzel, & Green, 1996) is, perhaps, the most philosophically consistent with criterion-referenced, standards-based¹ assessments like the NeSA. *Bookmark* is an *item-based* method. It requires panelists to determine which items can be successfully answered 67% of the time by students at the Performance Level boundaries. The *Contrasting Groups* method (Cizek & Bunch, 2007, chapter 8) is *student-based* which asks teachers to place students into one of the three Performance Levels based on their knowledge of the students from their classrooms without considering the assessment. The success of either approach requires an in-depth understanding of the skills and knowledge required at each level. This shared understanding is expressed in *Performance Level Descriptors* (Appendix N).

To assist the State Board of Education in determining appropriate cut scores, DRC presented the results of both studies: the *Bookmark* and the *Contrasting Groups*. A composite of the two studies was also considered. An analytical smoothing of the results was done to provide a coherent representation of the data across grades that, overall, did not raise or lower the panel recommendations. Ultimately, the State Board of Education approved cut scores that were above the recommendations but within one standard error of measurement from the smoothed values.

Board-Approved Cut Scores

The final State Board of Education approved cut scores and the percentage of spring 2011 students in each Performance Level are shown in Table 3.3.1. These values in the scale score metric will not change from year to year. The *Raw Score Ranges* may vary from year to year, depending on the difficulty of the specific form, and the *Percent in Each Performance Level* will vary, depending on the proficiency of the students at that time.

Cut scores are defined in a logit metric, which, like scale scores, are fixed. Logits are related to percentage correct scores but are preferred because they are not tied to a specific test form and will not change from year to year. This ensures a consistent definition of the Performance Levels even if

¹ The term *standard* is used in two different senses in this area. *Content standards* are written descriptions of the goals and expectations for learning and instruction at each grade level. *Performance standards*, which are the focus of this section, define the levels of achievement necessary for each Performance Level. In some contexts, the term *performance standard* is interchangeable with *cut score*.

different test forms vary in difficulty. For reporting purposes, logits are converted into the scale scores, which is mathematically equivalent but more user-friendly.

Table 3.3.1: Logit and 2011 Raw Score Cut points for NeSA-M

Grade	Scale Score Ranges by Performance Level			2011 Raw Score Ranges by Performance Level			Logit Cut Points		2011 Percent in Each Performance Level		
	Below	Meets	Exceeds	Below	Meets	Exceeds	B/M	M/E	Below	Meets	Exceeds
3	1 to 84	85-134	135 to 200	1 to 33	34 to 45	46 to 50	-0.6000	1.1000	32.7	49.8	17.5
4	1 to 84	85-134	135 to 200	1 to 37	38 to 50	51 to 55	-0.6000	1.2000	32.4	51.7	15.9
5	1 to 84	85-134	135 to 200	1 to 37	38 to 50	51 to 55	-0.5700	1.1597	34.1	48.2	17.7
6	1 to 84	85-134	135 to 200	1 to 41	42 to 53	54 to 58	-0.4700	1.1816	37.3	44.3	18.4
7	1 to 84	85-134	135 to 200	1 to 38	39 to 52	53 to 58	-0.4500	1.2500	38.5	45.3	16.2
8	1 to 84	85-134	135 to 200	1 to 41	42 to 55	56 to 60	-0.4000	1.3000	39.5	44.5	16.0
11	1 to 84	85-134	135 to 200	1 to 37	38 to 51	52 to 60	-0.2900	1.1000	46.0	32.8	21.2

The meaning of the logit and scale score values will not change in the future, but the raw score ranges may shift slightly to reflect the variation in item and form difficulty; a more difficult form will require fewer correct responses and an easier form will require more. With a stable scale score cut point, changes in the percentage of students in each proficiency level will reflect changes in student proficiency and not changes in form difficulty.

3.4 Scale Score Metric

Defining the scale score metric is an important, albeit arbitrary, step. Mathematically, scale scores are a linear transformation of the logit scores and thus do not alter the relationships or the displays. Scale scores simply provide more attractive labels for the scales. This is not meant to minimize the practical importance of this step because these are the numbers that will be reported to describe the performance of the students, schools, and systems. They will define the ranges of the performance levels, appear on individual student reports and school accountability analyses, and be dissected in newspaper accounts.

Appendix O contains the detailed raw score to scale score conversion tables that were used to assign Scale Scores to students based on the total number correct scores from the NeSA-R for 2010. Because the relationship between raw and scale scores depends on the difficulties of the specific items on the form, these tables will change for every operational form.

There are two primary considerations when establishing the metric:

- Multiply the logit by a value large enough to make decimal points unnecessary for student scores, and
- Shift the scale enough to avoid negative values for low Scales Scores.

The scale chosen for all grades of the NeSA will range from 0 to 200. The value of 0 is reserved for students who were not tested or were otherwise invalidated. Any student who attempted the test will receive a Scale Score equal to 1 even if the student gave no correct responses. No student tested will receive a Scale Score higher than 200 or lower than 1 even if this requires constraining the Scale Score

Table 3.4.2: NeSA-M Conversion of Logits to Scale Scores

Grade	Logit Cut Points		Scale Score Ranges by Performance Level			Conversion	
	B/M	M/E	Below	Meets	Exceeds	Slope b	Intercept a
3	-0.6000	1.1000	1 to 84	85-134	135 to 200	29.41176	102.15706
4	-0.6000	1.2000	1 to 84	85-134	135 to 200	27.77778	101.17667
5	-0.5700	1.1597	1 to 84	85-134	135 to 200	28.90675	100.98685
6	-0.4700	1.1816	1 to 84	85-134	135 to 200	30.27367	98.73862
7	-0.4500	1.2500	1 to 84	85-134	135 to 200	29.41176	97.74529
8	-0.4000	1.3000	1 to 84	85-134	135 to 200	29.41176	96.27470
11	-0.2900	1.1000	1 to 84	85-134	135 to 200	35.97122	94.94165

3.5 Reading Pre- and Post-Equated Comparison

The intent of the NeSA exams is that the item parameter estimates are established from the initial field test data and considered fixed over the life of the items. Any changes in curriculum or instruction will be reflected in improved student performance and with no opportunity of being absorbed by revisions in the parameter estimates. The underlying assumption is that these changes will affect all items uniformly rather than uniquely by item, item type, or content standard. At the initial stages of the assessment, at least, this assumption should be verified.

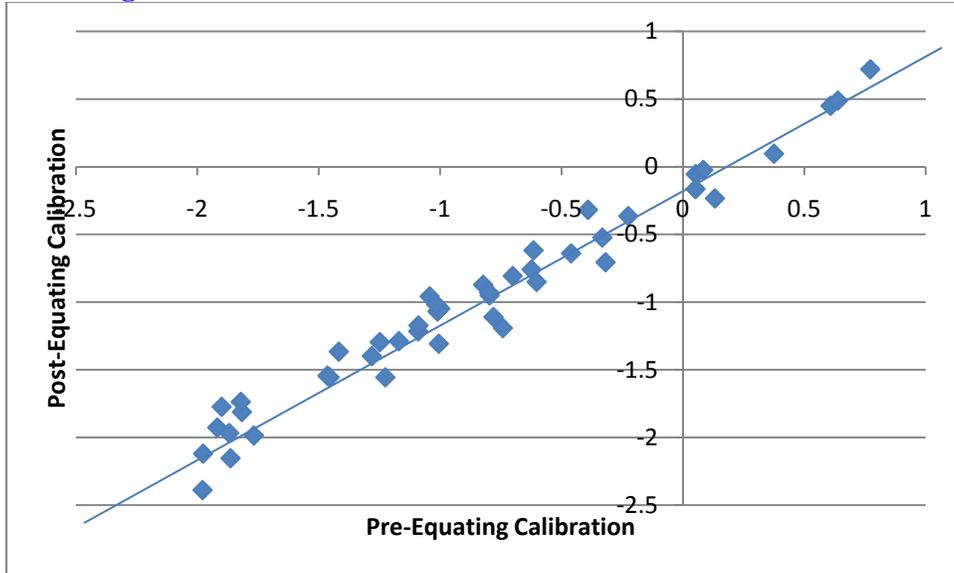
For the NeSA-R, which is the only assessment that has reached the stage where this assumption is an issue, the check was done by comparing the existing item calibrations from the field test with calibrations based on the current operational administration. One expects, through the measurement models invariance parameters, to obtain statistically equivalent estimates with a possible shift in the mean. These comparisons are present in Table 3.5.1 below.

Table 3.5.1: NeSA-R Pre- and Post Equating Comparison

	Grade						
	3	4	5	6	7	8	11
Correlation	0.98	0.97	0.98	0.98	0.97	0.95	0.96
SD pre	0.75	0.74	0.83	0.88	0.72	0.63	0.75
SD post	0.74	0.72	0.78	0.86	0.75	0.67	0.77
Ratio SD	1.02	1.03	1.06	1.03	0.95	0.94	0.98

Common criteria for comparing item calibrations across years are correlations of at least 0.95 and a ratio of standard deviations of between 0.90 and 1.10 (Huynh & Meyer, 2010). The high correlation ensures the items define the same construct and the ratio of SD's near one ensures a consistent unit. These data meet the criteria in all grades. The relationship for grade 3 is shown graphically below; detailed data are presented in Appendix S.

Figure 3.5.1: NeSA-R Grade 3 Pre- and Post-Calibrations



4. READING AND MATHEMATICS EMBEDDED FIELD TEST

4.1 Psychometric Summary

Traditional Item Statistics: The statistics computed are defined in detail in Section 3.1 above and traditional statistics for each NeSA-R field test item are in Appendix G and J and for NeSA-M Appendix H and K. The tables below provide summaries of the distributions of item percents correct, point-biserial correlations, and differential item functioning codes. Items with negative point-biserial correlations were never considered for operational use. Item with correlations less than 0.2 or percents correct less than 0.3 or greater 0.9 were avoided when possible.

Table 4.1.1: Summary of Traditional Item Statistics for NeSA-R 2011 Field Test Items

Grade	Item Percent Correct										Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	
3	0	0	1	2	2	9	8	12	12	4	50
4	0	0	0	4	3	9	11	9	12	2	50
5	0	0	6	6	6	6	8	8	6	4	50
6	0	0	1	2	6	10	6	11	13	1	50
7	1	1	0	6	6	8	10	8	10	0	50
8	0	1	1	1	7	4	11	11	13	1	50
11	0	1	2	1	1	7	8	6	17	7	50

Grade	Item Point-biserial Correlation							Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	
3	1	3	11	21	13	1	0	50
4	0	4	11	27	8	0	0	50
5	5	6	13	15	10	1	0	50
6	0	7	9	16	17	1	0	50
7	1	7	6	13	21	2	0	50
8	2	3	8	15	20	2	0	50
11	1	4	4	13	23	5	0	50

Table 4.1.2: Summary of Traditional Item Statistics for NeSA-M 2011 Field Test Items

Grade	Item Percent Correct										Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	
3	0	0	1	1	4	3	8	13	10	4	44
4	0	0	1	1	6	5	9	11	8	9	50
5	0	0	1	1	5	7	12	13	8	3	50
6	1	0	0	0	2	8	9	13	9	8	50
7	0	0	2	5	3	4	15	11	8	2	50
8	0	0	1	3	9	10	10	8	7	2	50
11	0	1	9	13	10	4	7	2	4	0	50

Grade	Item Point-biserial Correlation							Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	
3	1	2	9	12	18	2	0	44
4	1	3	3	22	15	6	0	50
5	2	1	11	16	15	5	0	50
6	0	4	6	12	22	6	0	50
7	2	2	4	14	19	9	0	50
8	1	3	4	15	20	7	0	50
11	7	5	9	12	14	3	0	50

Differential Item Functioning (DIF): The differential item function statistics are defined in detail in Section 3.1 above, and item statistics are included in Appendix J for NeSA-R and Appendix K for NeSA-M. Groups that were too small to provide meaningful results are labeled NA for Not Applicable. The first column defines the focal group; codes with a minus sign indicate items that disadvantaged this group in comparison to the reference group, which is male for gender or White for ethnicity.

Table 4.1.3: Summary of DIF by Code for NeSA-R 2011 Field Test

Grade 3	A	B+	B-	C+	C-	Total Items
Female	50	0	0	0	0	50
Black	47	0	1	1	1	50
Hispanic	48	0	1	0	1	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	10	0	0	0	0	10

Grade 4	A	B+	B-	C+	C-	Total Items
Female	49	0	1	0	0	50
Black	44	1	4	0	1	50
Hispanic	48	1	1	0	0	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	10	0	0	0	0	10

Grade 5	A	B+	B-	C+	C-	Total Items
Female	48	2	0	0	0	50
Black	38	0	2	0	0	40
Hispanic	48	0	1	0	1	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	10	0	0	0	0	10

Grade 6	A	B+	B-	C+	C-	Total Items
Female	46	1	2	1	0	50
Black	36	0	3	0	1	40
Hispanic	44	0	4	0	2	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	10	0	0	0	0	10

Nebraska State Accountability Technical Report 2011

Grade 7	A	B+	B-	C+	C-	Total Items
Female	46	2	1	0	1	50
Black	42	0	7	0	1	50
Hispanic	44	0	5	1	0	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	9	0	1	0	0	10

Grade 8	A	B+	B-	C+	C-	Total Items
Female	42	3	4	1	0	50
Black	24	0	3	0	3	30
Hispanic	47	0	2	0	1	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	9	0	1	0	0	10

Grade 11	A	B+	B-	C+	C-	Total Items
Female	42	7	1	0	0	50
Black	38	0	6	0	6	50
Hispanic	45	0	4	0	1	50
Native American	NA	NA	NA	NA	NA	0
Asian	NA	NA	NA	NA	NA	0
Multiple	NA	NA	NA	NA	NA	0

Table 4.1.4: Summary of DIF by Code for NeSA-M 2011 Field Test

Grade 3	A	B+	B-	C+	C-	Total Items
Female	43	0	0	0	1	44
Black	16	0	3	0	0	19
Hispanic	40	1	3	0	0	44
Native American	NA	NA	NA	NA	NA	NA
Asian	9	1	0	0	0	10
Multiple	10	0	0	0	0	10

Grade 4	A	B+	B-	C+	C-	Total Items
Female	49	0	1	0	0	50
Black	7	1	2	0	0	10
Hispanic	47	1	2	0	0	50
Native American	2	0	2	0	1	5
Asian	8	0	0	0	2	10
Multiple	10	0	0	0	0	10

Grade 5	A	B+	B-	C+	C-	Total Items
Female	46	2	2	0	0	50
Black	10	0	0	0	0	10
Hispanic	48	2	0	0	0	50
Native American	NA	NA	NA	NA	NA	NA
Asian	9	1	0	0	0	10
Multiple	10	0	0	0	0	10

Grade 6	A	B+	B-	C+	C-	Total Items
Female	45	1	3	0	1	50
Black	9	0	1	0	0	10
Hispanic	44	1	4	1	0	50
Native American	NA	NA	NA	NA	NA	NA
Asian	9	1	0	0	0	10
Multiple	10	0	0	0	0	10

Nebraska State Accountability Technical Report 2011

Grade 7	A	B+	B-	C+	C-	Total Items
Female	41	5	4	0	0	50
Black	9	0	1	0	0	10
Hispanic	47	0	3	0	0	50
Native American	NA	NA	NA	NA	NA	NA
Asian	10	0	0	0	0	10
Multiple	9	0	1	0	0	10

Grade 8	A	B+	B-	C+	C-	Total Items
Female	48	0	1	0	1	50
Black	9	0	1	0	0	10
Hispanic	50	0	0	0	0	50
Native American	NA	NA	NA	NA	NA	NA
Asian	8	1	1	0	0	10
Multiple	10	0	0	0	0	10

Grade 11	A	B+	B-	C+	C-	Total Items
Female	48	0	2	0	0	50
Black	9	0	1	0	0	10
Hispanic	42	0	3	0	0	45
Native American	NA	NA	NA	NA	NA	NA
Asian	9	1	0	0	0	10
Multiple	10	0	0	0	0	10

5. SCIENCE STANDALONE FIELD TEST

5.1 Sampling

Schools were recruited to participate in the online science field test. Every attempt was made to obtain a sample representative of the state covering all regions, ethnic-cultural groups, and district and school types. Psychometrically, any issues related to the students actually tested are mitigated by the use of the Rasch measurement model, which conditions out the influence of the ability distribution. However, it is still important to reflect the diversity of the state in the sample, both to foster acceptance of the assessment and to ensure the robustness of the measurement model.

For the NeSA-S, the sample is also representative of the state. Specifics of the NeSA-S sample are provided in Tables 5.1.1 and 5.1.2.

Table 5.1.1: Demographic Comparison of NeSA-S Field Test Schools

Grade	Online		Paper		Online		Paper		Online		Paper	
	White	Non	White	Non	FRL	Non	FRL	Non	SpEd	Non	SpEd	Non
5	92.8%	95.2%	7.2%	4.8%	93.7%	94.6%	6.3%	5.4%	98.3%	89.8%	1.7%	10.2%
8	92.7%	96.3%	7.3%	3.7%	94.8%	94.3%	5.2%	5.7%	98.7%	90.3%	1.3%	9.7%
11	89.2%	91.1%	10.8%	8.9%	91.1%	89.4%	8.9%	10.6%	98.0%	82.3%	2.0%	17.7%

*FRL=Free and reduced lunch status, SpEd=Special education status

Table 5.1.2: Summary Demographic Breakdown for NeSA-S Field Test

Group	Over-all	Gender		Ethnicity					Special Ed		ELL		FLS	
		Male	Female	BL	AM	HI	AS	WH	No	Yes	No	Yes	No	Yes
5	17343	8868	8474	1147	257	2988	283	12133	14702	2641	16097	1246	9422	7887
8	17360	8881	8478	1075	229	2657	331	12551	15271	2089	16815	545	10343	6995
11	14569	7380	7188	785	179	1730	257	11242	13135	1434	14297	272	9862	4681

*AM=American Indian, AS=Asian, BL=African American/Black, HI= Hispanic, MU=Multiple Ethnicities, WH=White

5.2 Psychometric Summary

Traditional Item Statistics: The statistics computed are defined in detail in Section 3.1 above, and traditional statistics for each NeSA-S field test item are in Appendix I, L and R. The tables below provide summaries of the distributions of item percents correct, point-biserial correlations, and differential item functioning codes. Items with negative point-biserial correlations were not considered for operational use; items with correlations less than 0.2 or percents correct less than 0.3 or greater than 0.9 were avoided when possible.

Table 5.2.1: Summary of Traditional Item Statistics for NeSA-S Field Test Items

Grade	Item Percent Correct										Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	<=0.7	<=0.8	<=0.9	>0.9	
5	0	3	5	10	22	23	25	30	16	7	141
8	0	4	13	27	37	23	32	19	8	0	163
11	1	7	16	34	33	41	40	25	12	4	213

Grade	Item Point-biserial Correlation							Total
	<=0.1	<=0.2	<=0.3	<=0.4	<=0.5	<=0.6	>0.6	
5	1	10	37	58	35	0	0	141
8	6	11	43	66	34	3	0	163
11	10	24	39	74	62	4	0	213

Differential Item Functioning (DIF): The differential item function statistics are defined in detail in Section 3.1 above, and item statistics are included in Appendix J. The NA indicates cells where the sample size was too small to be meaningful. The first column indicates the focal group; codes with a minus sign correspond to items the disadvantage the focal group in comparison to the Reference group. The Reference group was male for gender and white for Ethnicity.

Table 5.2.2: Summary of Differential Item Functioning by Code for NeSA-S Field Test

Grade 5	A	B+	B-	C+	C-	Total Items
Female	136	2	3	0	0	141
Black	126	5	8	0	2	141
Hispanic	137	2	1	0	1	141
Native American	NA	NA	NA	NA	NA	NA
Asian	NA	NA	NA	NA	NA	NA
Multiple	49	1	2	0	0	52

Grade 8	A	B+	B-	C+	C-	Total Items
Female	154	3	5	1	0	163
Black	152	1	5	0	5	163
Hispanic	153	2	7	0	1	163
Native American	NA	NA	NA	NA	NA	NA
Asian	16	1	0	0	0	17
Multiple	17	0	0	0	0	17

Grade 11	A	B+	B-	C+	C-	Total Items
Female	190	9	9	1	4	213
Black	64	4	3	1	1	73
Hispanic	207	2	3	1	0	213
Native American	NA	NA	NA	NA	NA	NA
Asian	NA	NA	NA	NA	NA	NA
Multiple	55	3	2	0	0	60

6. ONLINE TESTING TIMES

Figures 6.1.1 and 6.1.2 contain a breakout of testing times from the 2011 NeSA-R and NeSA-M assessments respectively. The data (see Table 6.1.1 and 6.1.2) were compiled based on students who had a *single login*, a *single logout*, and *responded to all the items*. In contrast to 2010, there was very little difference in the time spent in sessions 1 and 2, although still a slight tendency toward less time in the second session, particularly for mathematics.

Figure 6.1.1: Duration of Online Reading Testing Time by Grade and Session

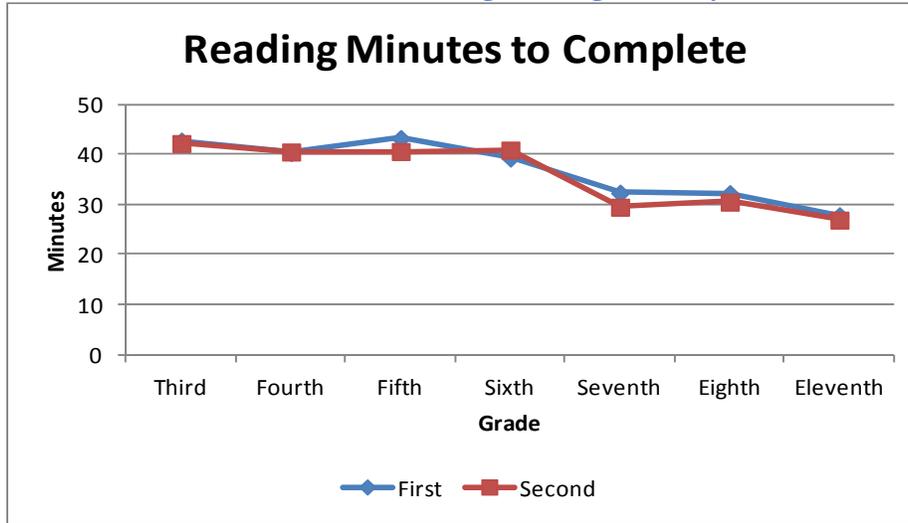
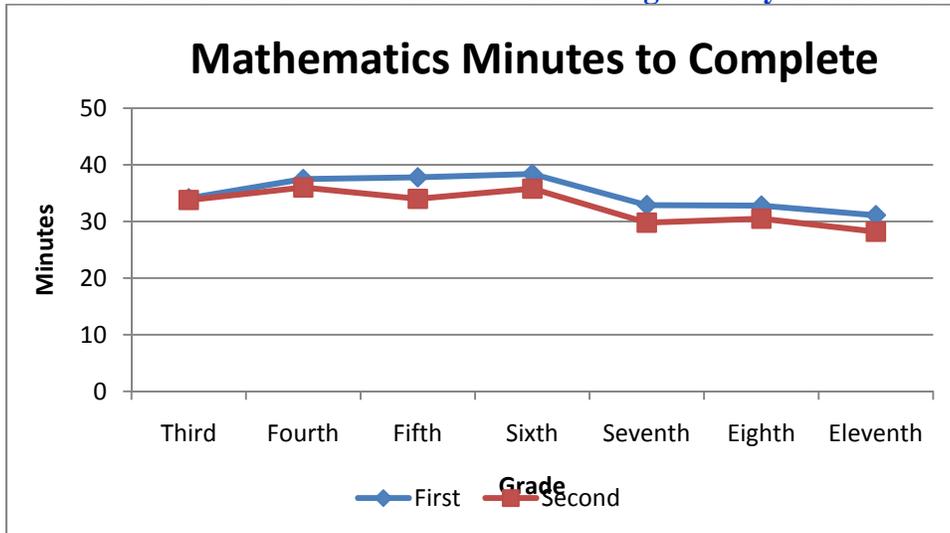


Figure 6.1.2: Duration of Online Mathematics Testing Time by Grade and Session



There were students who answered every item in less than five minutes. There was a remarkably constant number around 200 for all grades in session 1, which probably reflects something related to administration rather than to student behavior. The very short times in session 2 again increased with grade level. The outliers on the other end, greater than 90 minutes, are also interesting because these

data do not include students who *paused out*, had the test end due to inactivity, or were reactivated. It appears that they were actively involved with the test for the full time between the login and logout, but it raises the question of how fully engaged those students may have been for that amount of time.

Table 6.1.1: Duration of Online Reading Testing Sessions

Grade	3		4		5		6		7		8		11	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2
<5	206	9	227	3	222	7	182	4	217	11	225	33	220	113
5-10	53	14	48	13	51	10	45	9	67	69	85	86	132	316
10-15	56	128	77	142	23	84	55	58	229	453	190	477	495	1040
15-20	289	470	430	667	155	454	366	368	1156	2015	951	1671	1987	2677
20-25	974	1297	1371	1591	787	1395	1366	1189	2940	3660	2637	3284	4379	3912
25-30	1877	2170	2284	2316	1936	2566	2616	2296	3603	3794	3977	3698	4377	3656
30-35	2505	2506	2699	2493	2767	2916	3009	2903	3002	2667	3334	2905	2941	2539
35-40	2624	2489	2678	2515	2771	2586	2510	2567	1963	1400	2097	1704	1487	1360
40-45	2263	2091	2040	1949	2321	2016	1850	2021	1222	831	1242	1067	690	693
45-50	1915	1691	1695	1486	1758	1501	1378	1388	782	495	717	521	364	320
50-55	1370	1193	1186	1140	1233	1104	943	927	457	275	427	336	172	198
55-60	1002	922	806	830	992	709	658	733	277	196	249	204	110	106
60-65	726	632	645	534	679	495	475	515	232	133	185	130	49	54
65-70	449	447	439	398	502	372	318	340	119	91	104	79	40	52
70-75	303	332	273	268	343	266	256	211	88	53	56	59	31	28
75-80	209	198	171	188	244	194	163	189	43	38	40	43	15	15
80-85	144	132	104	159	196	145	91	129	27	16	24	28	16	11
85-90	110	91	73	101	136	83	68	83	39	19	26	31	3	4
>90	262	210	165	228	291	276	197	270	58	37	54	49	17	24
Total	17337	16803	17411	17021	17407	17179	16546	16200	16521	16253	16620	16405	17525	17118
Mean	42.7	42.2	40.4	40.5	43.2	40.6	39.3	40.9	32.3	29.5	32.2	30.5	27.8	26.9

Table 6.1.2: Duration of Mathematics Online Testing Sessions

rade	3		4		5		6		7		8		11	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2
<5	209	5	223	0	206	2	194	1	216	8	218	25	204	80
5-10	32	12	45	7	61	18	42	6	66	26	77	49	132	226
10-15	152	249	74	164	42	159	21	65	86	274	80	170	229	511
15-20	1029	1231	676	993	424	1174	375	745	674	1414	562	1077	686	1189
20-25	2140	2255	1758	2078	1462	2457	1316	1886	1945	2751	1848	2772	1751	2327
25-30	2455	2485	2191	2397	2342	2472	2055	2319	2651	2764	2850	2883	2516	2598
30-35	2083	2010	2073	1984	2258	2050	1958	1940	2302	1905	2383	2004	2312	1865
35-40	1651	1476	1572	1515	1836	1365	1631	1487	1602	1072	1648	1139	1514	1104
40-45	1082	1079	1263	1059	1222	919	1189	947	1011	697	1032	709	912	540
45-50	736	706	854	639	960	620	921	682	649	419	596	430	483	263
50-55	542	422	659	605	573	429	590	487	386	235	352	277	237	139
55-60	370	324	476	381	466	293	444	336	260	144	219	161	130	76
60-65	205	196	425	322	320	240	280	228	116	79	143	89	69	50
65-70	150	154	266	202	222	178	223	159	87	63	92	47	34	29
70-75	111	108	147	182	158	110	178	111	59	28	42	42	23	17
75-80	77	70	134	94	111	57	122	81	35	25	32	29	12	14
80-85	53	55	76	115	78	52	79	67	24	16	16	18	4	10
85-90	32	32	100	43	63	40	52	50	18	17	19	15	9	5
>90	78	94	171	146	204	145	183	136	42	19	29	28	11	10
Total	13187	12963	13183	12926	13008	12780	11853	11733	12229	11956	12238	11964	11268	11053
Mean	34.1	33.8	37.5	36.0	37.8	34.0	38.4	35.8	32.9	29.8	32.8	30.5	31.1	28.2

7. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D. (1988). *Rasch models for measurement*. Newberry Park, CA: Sage Publications.
- Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Fischer, G. & Molenaar, I. (1995). *Rasch models – Foundations, recent developments, and applications*. New York: Springer.
- Hambleton, R. & Novick, M. (1973). "Toward an Integration of Theory and Method for Criterion-Referenced Tests." *Journal of Educational Measurement*, 10, 159–170.
- Hanson, B. A. & Brennan, R. L. (1990). An Investigation of Classification Consistency Indexes Estimated Under Alternative Strong True Score Theory Models. *Journal of Educational Measurement*, 27(4), 345–359.
- Huynh, H. (1976). "On the Reliability of Decisions in Domain-Referenced Testing." *Journal of Educational Measurement*, 13, 253–264.
- Huynh, H. & Meyer, P. (2010). "Use of Robust z in Detecting Unstable Items in Item Response Theory Models." *Practical Research, Assessment, and Evaluation*, vol 15, no. 2.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-Based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- Linacre, J. (2004). Estimation methods for Rasch measures. In E. Smith & R. Smith (Eds.). *Introduction to Rasch measurement. Theory, models and applications*. Maple Grove, MN: JAM Press. 25-47.
- Lissitz, R. W. & H. H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Retrieved January 10, 2008 from <http://PAREonline.net/getvn.asp?v=8&n=10>
- Livingston, S. & Lewis, C. (1995). "Estimating the Consistency and Accuracy of Classifications Based on Test Scores." *Journal of Educational Measurement* 32, 179–197.
- Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10, 217–229.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174.

- Mead, R. J. (1976). *Assessing the fit of data to the Rasch model through the analysis of residuals*. Unpublished doctoral dissertation, Chicago: University of Chicago.
- Mead, R. J. (2008). *A Rasch primer: The measurement theory of Georg Rasch*. (Psychometrics services research memorandum 2008–001). Maple Grove, MN: Data Recognition Corporation.
- Mogilner, A. (1992). *Children's Writer's World Book*. Cincinnati, OH: Writer's Digest Books.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 14, 58-94.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1, 199–218.
- Smith, E. V. & Smith, R. M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.
- Stearns, M. & Smith R. M. (2007). *Estimation of classification consistency indices for complex assessments: Model based approaches*. Paper presented at the 2007 Annual Convention of the American Educational Research Association. Chicago, IL.
- Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Brisner, E. P. (1989). *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.
- Thompson, S., Johnston, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. National Center on Educational Outcomes Synthesis Report 44. Minneapolis, MN: University of Minnesota.
- Webb, N. L. (2002). *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessment for Four States*. Washington, D.C.: Council of Chief State School Officers.
- WINSTEPS. (2011). *WINSTEPS[®] Rasch Measurement*. [Computer Program]. Chicago: WINSTEPS.com.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.
- Wright, B. & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wright, B. & Mok, M. (2004). An overview of the family of Rasch measurement models. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 25-47) Maple Grove, MN: JAM Press.
- Wright, B. & Panchapakesan, N. (1969). A procedure of sample-free item analysis. *Educational and Psychological Measurement* 29, pp. 23-48.
- Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.