Spring 2016

Nebraska State Accountability (NeSA)

Reading, Mathematics, and Science

Technical Report

September 2016

Prepared by Data Recognition Corporation

DATA RECOGNITION

DRC

CORPORATION

2016 NEBRASKA STATE ACCOUNTABILITY (NeSA) TECHNICAL REPORT

TABLE OF CONTENTS

3. STUDENT DEMOGRAPHICS

4. CLASSICAL ITEM STATISTICS

5. RASCH ITEM CALIBRATION

6. EQUATING AND SCALING

7. FIELD TEST ITEM DATA SUMMARY

8. RELIABILITY

1. Background

1.1 Purpose and organization of this report

This report documents the technical aspects of the 2016 Nebraska State Accountability Reading (NeSA-R), Mathematics (NeSA-M), and Nebraska Science (NeSA-S) operational tests, along with the NeSA-R, NeSA-M and NeSA-S embedded field tests, covering details of item and test development processes, administration procedures, and psychometric methods and summaries.

1.2 Background of the Nebraska State Accountability (NeSA)

Previous Nebraska Assessments: In previous years, Nebraska administered a blend of local and state-generated assessments to meet No Child Left Behind (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, the NDE administered one writing assessment annually in grades 4, 8, and 11. In addition, the 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://www.legislature.ne.gov/laws/statutes.php?statute=79-760.03) required a single statewide assessment of the Nebraska academic content standards for reading, mathematics, science, and writing 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, NeSA-S for science, and NeSA-W for writing (Complete documentation of the technical details for NeSA-W are presented in a separate document labeled *NeSA 2016 Writing Test Technical Report*). The assessments in reading and mathematics were administered in grades 3-8 and 11; science was administered in grades 5, 8, and 11.

NeSA replaced previous school-based assessments for purposes of local, state, and federal accountability. The NeSA RMS consists entirely of multiple choice items and will be administered, to the extent practicable, online. In January 2009, the 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	5, 8 and 11

Advisory Committees: Legislative Bill 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.

New College and Career Ready Standards for English Language Arts:
New College and Career Ready Standards for English Language Arts were adopted by the State Board of Education in September of 2014. Districts had to adopt these standards within one year of their adoption.

The 2015-16 NeSA-Reading operational assessment was aligned to the 2009 Legacy Language Arts Standards and to the new College and Career Ready Standards for English Language Arts. The NeSA-R also included field test items for the new NeSA-English Language Arts (ELA) assessment. Student scores for the NeSA-R were calculated using only operational items aligned to 2009 Legacy Language Arts Standards. The NeSA-ELA field test items were aligned to the 2014 College and Career Ready Standards for English Language Arts and included new item types. This report includes technical information about the field test items. The NeSA-ELA will be implemented in the spring of 2017.

2. ITEM AND TEST DEVELOPMENT

2.1 CONTENT STANDARDS

In April of 2008, the Nebraska Legislature passed into state law Legislative Bill 1157. 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 content 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 content 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 content areas by July 1, 2013.
- New College and Career Ready Standards for English Language Arts were adopted by the State Board of Education in September of 2014. Spring 2016 was the final administration of the NeSA-R and spring 2017 will mark the first administration of the NeSA-English Language Arts (ELA) assessment.

The Nebraska Language Arts Standards are the foundation for 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 the NeSA 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 NeSA 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 (DOK) 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 A). The NeSA-M test blueprint was developed and approved in fall 2010 (Appendix B). The NeSA-S test blueprint was developed and approved in fall 2011 (Appendix C).

2.3 Multiple-Choice Items (MC)

Each assessment incorporates multiple-choice (MC) items to assess the content standards. Students are required to select a correct answer from four response choices with a single correct answer. Each MC item is scored as right or wrong and has a value of one raw score point. MC 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, the 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; DOK; estimated difficulty level; and determining style, accuracy, and correct terminology. In addition, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) 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 (reading only) 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 (DOK);
- 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 content 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.
- 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 DOK: Writers were asked to classify the DOK of each item, using a model based on Norman Webb's work on DOK (Webb, 2002). Items were classified as one of four DOK categories: recall (DOK Level 1), skill/concept (DOK Level 2), strategic thinking (DOK Level 3), and extended thinking (DOK Level 4).
- 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 MC 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 content 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, the NDE 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 the 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 NDE 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 NDE and 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) (l)]. 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. The 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, the 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. The 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. The 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

The 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 universally 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. The NDE and DRC 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 the 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 the 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 the NDE to develop a style guide that includes dimensions of style that are consistent with universal design.

DOK: Interpreting and assigning DOK levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. Four levels of DOK are used for this analysis. The NeSA assessments include items written at levels 1, 2, and 3. Level 4 items are not included due to the test being comprised of only MC 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. 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, Level 4 performance are:

- Based on data provided from a complex experiment that is novel to the student, deduce the fundamental relationship between a controlled variable and an experimental variable.
- 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 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, DOK, 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 NDE. The NDE test development team members facilitated the process. Training was provided by the NDE and included how to review items for technical quality and content quality, including DOK 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)
- DOK (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 the 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 an NDE test development lead to review items for bias and sensitivity issues using *Fairness in Testing* training manual developed by DRC (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 the 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 DOK levels.

2.7 The Operational Form Construction Process

The Spring 2016 operational forms were constructed in Lincoln, Nebraska in August 2015 (Mathematics) and early September 2015 (Reading, Science). 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 333the 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 and science, the item cards were compiled in binders and sorted by *p*-values from highest to lowest by standard and indicator.

The 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 2016 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 area. As items were selected, the unique item codes were entered into a form building template (Perform) which contained the item pool with statistics and item characteristics. Perform 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: 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 required blueprint and the depth of the available operational-ready item pool occasionally require 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-R 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 DRC (INSIGHT Online Learning System). One form of the test was also published in a printed test booklet for students needing accommodation provided by paper/pencil test. Depending on grade, the forms contained 45 to 50 operational items.

Table 2.8.1 Reading 2016 Operational Test

Grade	Total # of Operational Items per Form			Total # of Embedded Research (FT or EB) Items per Form*			Total # of FT Forms	Total # of Items Field Tested	Total # of Items per Form
	MC	ASCR/EBSR	TDA	MC	ASCR/EBSR	TDA †			
3	45	0	0	8	6	0	5	70	59
4	45	0	0	8	6	0	5	70	59
5	48	0	0	8	6	1	5	75	63
6	48	0	0	8	6	1	5	75	63
7	48	0	0	8	6	1	5	75	63
8	50	0	0	8	6	1	5	75	65
11	50	0	0	8	6	1	5	75	65

* Items are a part of a passage set. One FT or EB passage set is embedded per form.

† Grades 3 and 4 exempt from the 2016 TDA Field Test.

Psychometric Targets: The goal for the operational forms was to meet a mean *p*-value 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 2016 was the seventh operational administration of NeSA-R. Approximately 70% of the assessment was constructed from passages and related items field tested from Spring 2009–2015. The approximate remaining 30% of the assessment was constructed from an overlap of items and passages from the 2015 operational (core) item positions from the Spring 2015 operational forms.

In addition to the operational passage sets, each student received one randomly selected field test passage with 14-15 associated field test items. The passages and items taken by each student were administered in two (grades 3 and 4) or three (grades 5-8 and 11) testing sessions each intended to be administered in a single class period. Items within a passage set were administered in a fixed order after the passage. Equating was accomplished by anchoring on the operational passage items and calibrating the field test items concurrently.

With the adoption of revised NeSA-ELA standards in 2014, the NeSA-Reading tests will transition in order to meet the content and rigor of the revised English Language Arts standards. Starting with the 2016 NeSA-Reading administration, field tested items were aligned to the newly developed NeSA-ELA (English Language Arts) standards. The purpose of aligning to NeSA-ELA standards is to gather performance statistics and generate a pool of items for future use. Starting in 2017, the existing NeSA-Reading test will transition and report out as NeSA-ELA. Statistics for these field tested items can be found in Chapter 7 of this Technical Report.

The implementation of the transition will be phased in starting with the 2016 NeSA administration and concluding with a fully transition test with the 2018 NeSA administration. As part of the transition, three major changes will be implemented:

- The NeSA program will assess the newly updated ELA Standards and Indicators using a revised Table of Specifications (TOS) blueprint and a newly revised test design.
- In addition to traditional Multiple-Choice (MC) test questions, the program will incorporate both Technology-Enhanced (TE) test questions like Evidence-Based Selected Response (EBSR) and Auto-Scored Constructed Response (ASCR) test questions.
- In addition to auto-scored test questions (like the MC, EBSR, and ASCR), one hand-scored Text-Dependent Analysis (TDA) writing prompt will be added to the core (operational) of each grade. The inclusion of the TDA will be phased in at a different pace for grades 3 and 4 compared to the phase-in plan for grades 5 and higher.

The transition will be implemented in three waves as illustrated below. The subsections that follow describe the transition test plans in greater detail.

Grades	Test Attribute	Wave 1 Spring 2016 NeSA	Wave 2 Spring 2017 NeSA	Wave 3 Spring 2018 NeSA
3 and 4	**Alignment of Operational Items**	2009 Legacy Standards and 2014 Revised Standards	2014 Revised Standards	2014 Revised Standards
	Operational Test Items	MC items only	MC, EBSR, and ASCR	MC, EBSR, and ASCR
	Embedded Field Test Items	MC, EBSR, and ASCR	MC, EBSR, and ASCR	MC, EBSR, ASCR, and TDA*
5, 6, 7, 8, and 11	**Alignment of Operational Items**	2009 Legacy Standards and 2014 Revised Standards	2014 Revised Standards	2014 Revised Standards
	Operational Test Items	MC items only	MC, EBSR, ASCR, and TDA	MC, EBSR, ASCR, and TDA
	Embedded Field Test Items	MC, EBSR, ASCR, and TDA	MC, EBSR, ASCR, and TDA	MC, EBSR, ASCR, and TDA

* NDE is currently conducting research regarding future field testing and inclusion of TDA for grades 3-4

2.9 Mathematics Assessment

Test Design: The NeSA-M operational test includes operational and field test items. This test was administered online via the test engine developed and managed by DRC (INSIGHT Online Learning System). One form of the test was also published in a printed test booklet for students needing accommodation provided by paper/pencil test. Depending on grade, the forms contained 50 to 60 operational items.

Table 2.9.1 Mathematics 2016 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*-value 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 2016 was the sixth operational administration of NeSA-M. Approximately 70% of the assessment was constructed from items field tested from Spring 2010–2015. The approximate remaining 30% of the assessment was constructed from an overlap of items from the 2015 operational (core) item positions from the 2015 operational forms.

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

Test Design: The NeSA-S operational test includes operational and field test items. This test was administered online via the test engine developed and managed by DRC (INSIGHT Online Learning System). One form of the test was also published in a printed test booklet for students needing accommodation provided by paper/pencil test. Depending on grade, the forms contained 50 or 60 operational items.

Table 2.10.1 Science 2016 Operational Test

Grade	No. Operational Items	No. of Embedded FT Items per Form	Total Items	Total No. of FT Forms	Total No. of Items Field Tested
5	50	10	60	5	50
8	60	10	70	5	50
11	60	10	70	5	50

Psychometric Targets: The goal for the operational forms was to meet a mean *p*-value 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 2016 was the fifth operational administration of NeSA-S. Approximately 70% of the assessment was constructed from items field tested in Spring 2011–2015. The approximate remaining 30% of the assessment was constructed from an overlap of items from the 2015 operational (core) item positions from the 2015 operational forms.

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.

3. STUDENT DEMOGRAPHICS

Three areas of student demographics are discussed below, summary demographics and accommodations, summary information on the number of students tested with breakdowns by mode, and summary information on testing times.

3.1 DEMOGRAPGICS AND ACCOMMODATIONS

Gender, ethnicity, food program status (FRL), Limited English Proficiency/English Language Learners (LEP/ELL) status, Special Education status (SPED), and accommodation status data was collected for all students who participated and attempted the 2016 NeSA assessments. This summary of student demographics by grade and content area is provided in Tables 3.1.1– 3.1.7. These tables show around 22,000 students took the assessment in each grade. Of those students across grades, half are males, half are females, over two thirds white, and less than one fifth are Hispanic. Among the students across grades, about 37% to 47% are eligible for FRL, 2% to 9% are LEP/ELL, and 11% to 16% belong to at least one SPED category. For all three of these programs/categories, the participation rate is slightly lower for upper grade students. In terms of the test accommodations, there are about 6% to 16% of the students across grade and content area that report at least one type of accommodation (see row 'Total' for 'Accommodation' in the table). Similar to the rate for FRL, LEP/ELL, and SPED across grades, the rate for accommodation is lower for high school students (Grade 11). Across all grades, the 'Timing/Schedule/Setting' is the most utilized accommodation (about 6-10% for Grade 3-8, and 4% for Grade 11), followed by the 'Content Presentation' (about 6-9% for Grade 3-7, and 2-5% for Grade 8 and 11).

Table 3.1.1 Grade 3 NeSA Summary Data: Demographics and Accommodations

Grade 3		Reading		Mathematics	
		Count	%	Count	%
All Students		21887	100.00	22028	100.00
Gender	Female	10602	48.44	10663	48.41
	Male	11285	51.56	11365	51.59
Race/Ethnicity	American Indian/Alaska Native	293	1.34	295	1.34
	Asian	534	2.44	563	2.56
	Black	1496	6.84	1504	6.83
	Hispanic	4015	18.34	4093	18.58
	Native Hawaiian or other Pacific Islander	35	0.16	35	0.16
	White	14630	66.84	14652	66.52
	Two or More Races	884	4.04	886	4.02

Grade 3		Reading		Mathematics	
		Count	%	Count	%
Food Program	Yes	9842	44.97	9922	45.04
	No	11698	53.45	11709	53.16
LEP/ELL	Yes	1852	8.46	1965	8.92
	No	20035	91.54	20063	91.08
Special Education	Yes	3102	14.17	3116	14.15
	No	18785	85.83	18912	85.85
Accommo-dations	Content Presentation	1634	7.47	1657	7.52
	Response	900	4.11	1102	5.00
	Timing/Schedule/Setting	1897	8.67	1867	8.48
	Direct Linguistic Support with Test Directions	1135	5.19	1195	5.42
	Direct Linguistic Support with Content and Test items	1289	5.89	1393	6.32
	Indirect Linguistic Support	952	4.35	1035	4.70
	Spanish	18	0.08	79	0.36
	Braille*	1	0.00	1	0.00
	Large Print*	16	0.07	17	0.08
	Audio	506	2.31	521	2.37
	Total	3475	15.88	3586	16.28

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.2 Grade 4 NeSA Summary Data: Demographics and Accommodations

Grade 4		Reading		Mathematics	
		Count	%	Count	%
All Students		23039	100.00	23143	100.00
Gender	Female	11287	48.99	11333	48.97
	Male	11752	51.01	11810	51.03
Race/Ethnicity	American Indian/Alaska Native	322	1.40	320	1.38
	Asian	597	2.59	615	2.66

Grade 4		Reading		Mathematics	
		Count	%	Count	%
	Black	1600	6.94	1611	6.96
	Hispanic	4289	18.62	4356	18.82
	Native Hawaiian or other Pacific Islander	31	0.13	31	0.13
	White	15353	66.64	15365	66.39
	Two or More Races	847	3.68	845	3.65
Food Program	Yes	10606	46.03	10666	46.09
	No	12073	52.40	12076	52.18
LEP/ELL	Yes	1596	6.93	1687	7.29
	No	21443	93.07	21456	92.71
Special Education	Yes	3541	15.37	3545	15.32
	No	19498	84.63	19598	84.68
Accommo-dations	Content Presentation	2077	9.02	2086	9.01
	Response	1137	4.94	1438	6.21
	Timing/Schedule/Setting	2304	10.00	2294	9.91
	Direct Linguistic Support with Test Directions	1213	5.26	1276	5.51
	Direct Linguistic Support with Content and Test items	1383	6.00	1479	6.39
	Indirect Linguistic Support	1046	4.54	1101	4.76
	Spanish	26	0.11	88	0.38
	Braille*	2	0.01	2	0.01
	Large Print*	9	0.04	10	0.04
	Audio	620	2.69	629	2.72
	Total	4029	17.49	4115	17.78

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.3 Grade 5 NeSA Summary Data: Demographics and Accommodations

Grade 5		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		22689	100.00	22799	100.00	22798	100.00
Gender	Female	11018	48.56	11070	48.55	11067	48.54
	Male	11671	51.44	11729	51.45	11731	51.46
Race/Ethnicity	American Indian/Alaska Native	295	1.30	297	1.30	297	1.30
	Asian	567	2.50	590	2.59	589	2.58
	Black	1566	6.90	1581	6.93	1579	6.93
	Hispanic	4090	18.03	4158	18.24	4160	18.25
	Native Hawaiian or other Pacific Islander	32	0.14	32	0.14	32	0.14
	White	15291	67.39	15293	67.08	15293	67.08
	Two or More Races	848	3.74	848	3.72	848	3.72
Food Program	Yes	10129	44.64	10206	44.77	10202	44.75
	No	12237	53.93	12239	53.68	12241	53.69
LEP/ELL	Yes	1087	4.79	1197	5.25	1196	5.25
	No	21602	95.21	21602	94.75	21602	94.75
Special Education	Yes	3503	15.44	3503	15.36	3502	15.36
	No	19186	84.56	19296	84.64	19296	84.64
Accommo-dations	Content Presentation	2147	9.46	2211	9.70	2013	8.83
	Response	1179	5.20	1551	6.80	1115	4.89
	Timing/Schedule/Setting	2434	10.73	2454	10.76	2310	10.13
	Direct Linguistic Support with Test Directions	762	3.36	808	3.54	769	3.37
	Direct Linguistic Support with Content and Test items	956	4.21	1023	4.49	970	4.25
	Indirect Linguistic Support	761	3.35	828	3.63	808	3.54
	Spanish	31	0.14	89	0.39	89	0.39
	Braille*	1	0.00	0	0.00	1	0.00

Grade 5		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
	Large Print*	11	0.05	10	0.04	10	0.04
	Audio	610	2.69	646	2.83	653	2.86
	Total	3691	16.27	3805	16.69	3636	15.95

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.4 Grade 6 NeSA Summary Data: Demographics and Accommodations

Grade 6		Reading		Mathematics	
		Count	%	Count	%
All Students		22915	100.00	22997	100.00
Gender	Female	11231	49.01	11262	48.97
	Male	11684	50.99	11735	51.03
Race/Ethnicity	American Indian/Alaska Native	314	1.37	316	1.37
	Asian	583	2.54	601	2.61
	Black	1473	6.43	1476	6.42
	Hispanic	4098	17.88	4159	18.08
	Native Hawaiian or other Pacific Islander	29	0.13	29	0.13
	White	15631	68.21	15632	67.97
	Two or More Races	787	3.43	784	3.41
Food Program	Yes	10045	43.84	10106	43.94
	No	12548	54.76	12552	54.58
LEP/ELL	Yes	647	2.82	735	3.20
	No	22268	97.18	22262	96.80
Special Education	Yes	3397	14.82	3391	14.75
	No	19518	85.18	19606	85.25
Accommo-dations	Content Presentation	1934	8.44	1929	8.39
	Response	856	3.74	1534	6.67
	Timing/Schedule/Setting	2104	9.18	2122	9.23

Grade 6		Reading		Mathematics	
		Count	%	Count	%
	Direct Linguistic Support with Test Directions	426	1.86	486	2.11
	Direct Linguistic Support with Content and Test items	492	2.15	634	2.76
	Indirect Linguistic Support	444	1.94	475	2.07
	Spanish	40	0.17	96	0.42
	Braille*	1	0.00	1	0.00
	Large Print*	3	0.01	4	0.02
	Audio	601	2.62	603	2.62
	Total	3084	13.46	3211	13.96

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.5 Grade 7 NeSA Summary Data: Demographics and Accommodations

Grade 7		Reading		Mathematics	
		Count	%	Count	%
All Students		22598	100.00	22716	100.00
Gender	Female	10982	48.60	11028	48.55
	Male	11616	51.40	11688	51.45
Race/Ethnicity	American Indian/Alaska Native	299	1.32	301	1.33
	Asian	550	2.43	569	2.50
	Black	1502	6.65	1516	6.67
	Hispanic	3988	17.65	4086	17.99
	Native Hawaiian or other Pacific Islander	24	0.11	24	0.11
	White	15459	68.41	15443	67.98
	Two or More Races	776	3.43	777	3.42
Food Program	Yes	9776	43.26	9836	43.30
	No	12526	55.43	12528	55.15
LEP/ELL	Yes	579	2.56	699	3.08
	No	22019	97.44	22017	96.92

Grade 7		Reading		Mathematics	
		Count	%	Count	%
Special Education	Yes	3204	14.18	3188	14.03
	No	19394	85.82	19528	85.97
Accommo-dations	Content Presentation	1603	7.09	1646	7.25
	Response	667	2.95	1450	6.38
	Timing/Schedule/Setting	1754	7.76	1761	7.75
	Direct Linguistic Support with Test Directions	197	0.87	292	1.29
	Direct Linguistic Support with Content and Test items	180	0.80	312	1.37
	Indirect Linguistic Support	161	0.71	221	0.97
	Spanish	51	0.23	140	0.62
	Braille*	3	0.01	2	0.01
	Large Print*	9	0.04	7	0.03
	Audio	678	3.00	701	3.09
	Total	2363	10.46	2680	11.80

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.6 Grade 8 NeSA Summary Data: Demographics and Accommodations

Grade 8		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		22220	100.00	22320	100.00	22328	100.00
Gender	Female	10929	49.19	10973	49.16	10975	49.15
	Male	11291	50.81	11347	50.84	11353	50.85
Race/Ethnicity	American Indian/Alaska Native	283	1.27	282	1.26	282	1.26
	Asian	525	2.36	542	2.43	542	2.43
	Black	1370	6.17	1389	6.22	1387	6.21
	Hispanic	3968	17.86	4050	18.15	4053	18.15
	Native Hawaiian or other Pacific Islander	28	0.13	29	0.13	28	0.13
	White	15315	68.92	15299	68.54	15307	68.56

Grade 8		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
	Two or More Races	731	3.29	729	3.27	729	3.26
Food Program	Yes	9444	42.50	9507	42.59	9513	42.61
	No	12523	56.36	12515	56.07	12522	56.08
LEP/ELL	Yes	529	2.38	638	2.86	639	2.86
	No	21691	97.62	21682	97.14	21689	97.14
Special Education	Yes	2961	13.33	2949	13.21	2960	13.26
	No	19259	86.67	19371	86.79	19368	86.74
Accommo-dations	Content Presentation	1468	6.61	1492	6.68	1436	6.43
	Response	538	2.42	1400	6.27	653	2.92
	Timing/Schedule/Setting	1572	7.07	1598	7.16	1505	6.74
	Direct Linguistic Support with Test Directions	204	0.92	287	1.29	304	1.36
	Direct Linguistic Support with Content and Test items	184	0.83	327	1.47	288	1.29
	Indirect Linguistic Support	144	0.65	212	0.95	232	1.04
	Spanish	54	0.24	119	0.53	123	0.55
	Braille*	3	0.01	3	0.01	3	0.01
	Large Print*	10	0.05	10	0.04	10	0.04
	Audio	664	2.99	690	3.09	763	3.42
	Total	2233	10.05	2517	11.28	2354	10.54

*Count represents the number of booklets ordered. This is not tracked.

Table 3.1.7 Grade 11 NeSA Summary Data: Demographics and Accommodations

Grade 11		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
All Students		21400	100.00	21357	100.00	21366	100.00
Gender	Female	10405	48.62	10390	48.65	10390	48.63
	Male	10995	51.38	10967	51.35	10976	51.37
Race/Ethnicity	American Indian/Alaska Native	272	1.27	274	1.28	271	1.27

Grade 11		Reading		Mathematics		Science	
		Count	%	Count	%	Count	%
	Asian	541	2.53	540	2.53	538	2.52
	Black	1256	5.87	1254	5.87	1251	5.86
	Hispanic	3541	16.55	3547	16.61	3546	16.60
	Native Hawaiian or other Pacific Islander	33	0.15	34	0.16	34	0.16
	White	15081	70.47	15034	70.39	15053	70.45
	Two or More Races	676	3.16	674	3.16	673	3.15
Food Program	Yes	7858	36.72	7839	36.70	7845	36.72
	No	13241	61.87	13212	61.86	13220	61.87
LEP/ELL	Yes	481	2.25	500	2.34	502	2.35
	No	20919	97.75	20857	97.66	20864	97.65
Special Education	Yes	2218	10.36	2183	10.22	2205	10.32
	No	19182	89.64	19174	89.78	19161	89.68
Accommo-dations	Content Presentation	546	2.55	554	2.59	580	2.71
	Response	297	1.39	722	3.38	446	2.09
	Timing/Schedule/Setting	1021	4.77	1031	4.83	1016	4.76
	Direct Linguistic Support with Test Directions	149	0.70	163	0.76	161	0.75
	Direct Linguistic Support with Content and Test items	104	0.49	156	0.73	154	0.72
	Indirect Linguistic Support	127	0.59	142	0.66	142	0.66
	Spanish	76	0.36	100	0.47	116	0.54
	Braille*	3	0.01	2	0.01	3	0.01
	Large Print*	6	0.03	5	0.02	5	0.02
	Audio	200	0.93	181	0.85	232	1.09
	Total	1298	6.07	1450	6.79	1446	6.77

*Count represents the number of booklets ordered. This is not tracked.

3.2 STUDENTS TESTED AND MODE SUMMARY DATA

As noted in Chapters One and Two, the 2016 NeSA assessments were administered online to the extent practical. One form of the test was also published in a printed test booklet for students needing accommodation of a paper/pencil test. Tables 3.2.1 – 3.2.3 report the number of students in each test mode. For NeSA-R, between 2% and 7% of students took the assessment in the paper-based version with the lower percentages occurring in middle and high schools.

Table 3.2.1 NeSA-R Number of Students Tested

Grade	Total	Online	Paper	Percent Paper
3	21887	20409	1478	7
4	23039	21342	1697	7
5	22689	21216	1473	6
6	22915	21679	1236	5
7	22598	21704	894	4
8	22220	21439	781	4
11	21400	20896	504	2

For NeSA-M, between 2% and 8% of students took the assessment in the paper-based version.

Table 3.2.2 NeSA-M Number of Students Tested

Grade	Total	Online	Paper	Percent Paper
3	22028	20479	1549	7
4	23143	21390	1753	8
5	22799	21258	1541	7
6	22997	21716	1281	6
7	22716	21775	941	4
8	22320	21487	833	4
11	21357	20826	531	2

For NeSA-S, between 2% and 6% of students took the assessment in the paper version.

Table 3.2.3 NeSA-S Number of Students Tested

Grade	Total	Online	Paper	Percent Paper
5	22798	21325	1473	6
8	22328	21611	717	3
11	21366	20871	495	2

The number of students, across content area and grade level, who took the 2016 NeSA tests online instead of paper pencil is similar to that of the 2016 NeSA tests.

3.3 TESTING TIME

Online testing time for the 2016 NeSA assessments was examined for each grade and content area. The data in Tables 3.3.1, 3.3.2, and 3.3.3 were compiled based on students who had a *single login, a single logout,* and *responded to all the items.* Students from upper grade levels in mathematics and science generally indicated a tendency toward less time spent on the second session. Students who took greater than 90 minutes are interesting because this data does not include students who *paused out*, had the test ended 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 3.3.1 Duration of Reading Online Testing Sessions

Grade	3		4		5		6		7		8		11	
Session	1	2	1	2	1	2	1	2	1	2	1	2	1	2
<5 minutes	139	126	168	152	133	122	168	150	162	147	116	159	253	329
5-10	113	135	136	113	156	96	126	117	151	231	155	182	457	462
10-15	162	303	153	210	433	199	384	223	334	972	251	570	1524	1457
15-20	286	827	376	381	1481	732	1408	588	1050	2802	863	1594	3815	3144
20-25	578	1651	978	700	2764	1690	2866	1384	2330	4265	1795	2916	5234	4284
25-30	1204	2316	1894	1255	3387	2661	3656	2134	3374	3935	2888	3686	4013	4053
30-35	1818	2600	2466	1719	3204	3098	3279	2801	3592	3187	3417	3546	2358	2860
35-40	2236	2487	2549	2174	2536	2998	2447	2730	3058	2119	3276	2767	1237	1640
40-45	2282	2146	2521	2212	1955	2240	1858	2483	2358	1358	2635	2038	693	961
45-50	2118	1783	2116	2175	1379	1831	1446	2088	1691	879	1910	1303	435	555
50-55	1903	1383	1784	1943	940	1476	1019	1592	1093	585	1301	857	258	368
55-60	1633	1045	1381	1598	715	1046	685	1227	773	383	894	566	163	232
60-65	1239	855	1107	1398	518	758	530	948	486	209	572	340	113	150
65-70	1020	628	911	1118	357	548	433	681	333	149	403	239	67	105
70-75	861	463	644	872	293	369	298	571	236	101	237	190	56	68
75-80	656	385	465	670	205	252	230	373	164	69	175	102	34	38
80-85	454	254	348	543	176	230	206	333	113	50	134	58	31	26
85-90	330	211	262	417	121	185	136	241	98	39	76	45	16	22
>90	1267	716	983	1605	362	615	412	947	235	134	276	207	73	72
Total	20299	20314	21242	21255	21115	21146	21587	21611	21631	21614	21374	21365	20830	20826
Median	49	40	44	49	34	39	34	42	35	28	37	33	25	26

Table 3.3.1 Duration of Reading Online Testing Sessions (continued)

Grade	5	6	7	8	11
Session	3	3	3	3	3
<5 minutes	160	155	201	191	321
5-10	162	186	253	281	725
10-15	449	401	612	749	1900
15-20	1207	1023	1663	1717	3537
20-25	1924	2063	2597	2515	3752
25-30	2386	2615	2964	2858	2934
30-35	2516	2602	2854	2589	2163
35-40	2218	2370	2303	2059	1556
40-45	2020	2023	1737	1713	1102
45-50	1587	1643	1371	1391	772
50-55	1332	1355	1127	1029	491
55-60	1017	1114	768	843	354
60-65	880	858	642	706	297
65-70	669	733	532	536	189
70-75	551	550	422	382	158
75-80	454	426	305	319	118
80-85	350	314	229	253	88
85-90	242	258	207	195	58
>90	941	880	804	1013	223
Total	21065	21569	21591	21339	20738
Median	39.0	39.0	35.0	35.0	26.0

Table 3.3.2 Duration of Mathematics Online Testing Sessions

Grade	3		4		5		6		7		8		11	
Session	1	2	1	2	1	2	1	2	1	2	1	2	1	2
<5 minutes	132	171	128	130	120	119	146	144	126	146	119	117	236	334
5-10	128	184	124	135	112	150	135	142	133	147	123	133	313	530
10-15	254	714	278	463	151	222	152	161	173	255	185	217	687	1135
15-20	870	2003	916	1432	429	688	241	364	403	666	434	637	1716	2485
20-25	1763	3015	1863	2525	1074	1701	535	844	1158	1672	1176	1626	3484	3945
25-30	2545	3111	2494	2928	1869	2657	1311	1654	2261	2855	2287	2691	4213	4067
30-35	2622	2623	2745	2766	2466	2910	1978	2320	3028	3371	3077	3312	3529	3193
35-40	2425	1989	2519	2311	2608	2744	2483	2702	3073	3222	3114	3156	2517	1997
40-45	2077	1527	2104	1852	2366	2201	2530	2530	2703	2721	2752	2636	1548	1208
45-50	1648	1161	1749	1514	2147	1801	2462	2281	2315	1917	2213	1982	906	689
50-55	1318	890	1393	1206	1630	1362	2043	1857	1814	1481	1723	1423	543	389
55-60	1024	719	1094	909	1337	1045	1678	1501	1285	1027	1247	1038	353	261
60-65	741	511	805	758	1021	741	1399	1183	933	728	900	738	226	162
65-70	652	430	668	586	842	642	1047	913	691	475	595	512	148	129
70-75	519	323	549	401	663	435	796	652	451	276	419	318	92	74
75-80	396	239	419	318	507	354	613	524	300	192	281	213	57	54
80-85	289	178	301	225	405	282	436	399	205	151	205	180	40	22
85-90	242	131	244	184	317	224	346	300	174	111	143	126	40	23
>90	741	479	901	657	1111	914	1278	1154	490	314	439	381	124	72
Total	20386	20398	21294	21300	21175	21192	21609	21625	21716	21727	21432	21436	20772	20769
Median	39.0	32.0	40.0	36.0	44.0	39.0	48.0	45.0	41.0	38.0	41.0	39.0	30.0	28.0

Table 3.3.3 Duration of Science Online Testing Sessions

Grade	5		8		11	
Session	1	2	1	2	1	2
<5 minutes	131	153	129	150	315	486
5-10	238	432	384	942	1068	2539
10-15	1711	2910	2930	5222	5762	8387
15-20	4088	4923	5669	6237	6948	5590
20-25	4524	4158	4854	3999	3636	1967
25-30	3272	2920	3132	2149	1516	782
30-35	2333	1876	1778	1171	711	402
35-40	1458	1102	1002	646	326	205
40-45	929	728	607	384	180	115
45-50	665	529	381	200	115	94
50-55	497	396	235	139	78	60
55-60	333	268	132	99	48	46
60-65	279	213	103	66	40	37
65-70	215	138	65	40	27	42
70-75	137	117	48	17	17	17
75-80	119	98	21	14	7	11
80-85	65	61	15	18	7	7
85-90	53	53	15	13	6	10
>90	198	169	66	57	22	30
Total	21245	21244	21566	21563	20829	20827
Median	25.0	23.0	22.0	19.0	18.0	15.0

4. CLASSICAL ITEM STATISTICS

This chapter provides an overview of the most familiar item-level statistics obtained from classical (true-score model) item analysis: item difficulty, item discrimination, distractor distribution, and omits or blanks. The following results pertain only to operational NeSA items (i.e., those items that contributed to a student's total test score). Rasch item statistics are discussed in Chapter Five, and test-level statistics are found in Chapter Six. 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 F, G, and H.

4.1 ITEM DIFFICULTY

Item difficulty (*p*-value) is the proportion of examinees in the sample who answered the item correctly. For example, if an item has a *p*-value of 0.79, it means 79 percent of the students answered the item correctly. Relatively lower values correspond to more difficult items and those that have relatively higher values correspond to easier items. Items that are either very hard or very easy provide little information about student differences in achievement. On a standards-referenced test like the NeSA, a test development goal is to include a wide range of item difficulties. Typically, test developers target *p*-values in the range of 0.30 to 0.80. 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. Summary *p*-value information across all grades for each content area is shown in Tables 4.1.1 through 4.1.3. In general, most of the items fall into the *p*-value range of 0.4 to 0.8, which is appropriate for a criterion-referenced assessment. In reading the following tables, the heading ≤ 0.1 descrbes items between 0.0 and 0.1, and the heading ≤ 0.2 descirbes items between 0.1 and 0.2, etc.

Table 4.1.1 Summary of Proportion Correct for NeSA-R Operational Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	0	0	0	6	7	13	14	5	0	0.657	45
4	0	0	0	0	4	7	17	11	5	1	0.664	45
5	0	0	0	0	4	8	10	16	9	1	0.692	48
6	0	0	0	0	3	8	13	14	10	0	0.688	48
7	0	0	1	2	2	9	12	13	8	1	0.671	48
8	0	0	0	1	5	7	10	17	10	0	0.685	50
11	0	0	0	0	5	5	16	14	10	0	0.685	50

Table 4.1.2 Summary of Proportion Correct for NeSA-M Operational Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	0	0	2	2	10	17	18	1	0	0.654	50
4	0	0	0	1	4	12	23	10	5	0	0.651	55
5	0	0	0	0	4	20	15	13	3	0	0.630	55
6	0	0	0	0	5	11	26	11	5	0	0.640	58
7	0	0	0	0	4	17	20	17	0	0	0.637	58
8	0	0	0	0	5	17	20	15	3	0	0.639	60
11	0	0	0	0	7	13	23	16	1	0	0.635	60

Table 4.1.3 Summary of Proportion Correct for NeSA-S Operational Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
5	0	0	0	2	4	8	15	13	8	0	0.662	50
8	0	0	0	0	9	9	19	17	5	1	0.650	60
11	0	0	0	2	6	9	22	15	6	0	0.653	60

4.2 ITEM-TOTAL CORRELATION

Item-total correlation describes the relationship between performance on the specific item and performance on the entire form. For the NeSA tests, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate this relationship. For MC items, the statistic is typically referred to as point-biserial correlation. This index indicates an item's ability to differentiate between high and low achievers (i.e., item discrimination power). It is expected that students with high ability (i.e., those who perform well on the NeSA overall) would be more likely to answer any given NeSA item correctly, while students with low ability (i.e., those who perform poorly on the NeSA overall) would be more likely to answer the same item incorrectly. However, an interaction can exist between item discrimination and item difficulty. Items answered correctly (or incorrectly) by a large proportion of examinees (i.e., the items have extreme *p*-values) can have reduced power to discriminate and thus can have lower correlations.

The correlation coefficient can range from -1.0 to +1.0. If the aforementioned expectation is met (high-scoring students tend to get the item right while low-scoring students do not), the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero), meaning the item is a good discriminator between high- and low-ability students. 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 are flagged and referred to content specialists for review before being considered for use on future forms. As seen below in Tables 4.2.1 – 4.2.3, no items in the 2016 NeSA tests have negative point-biserial correlations and most are above 0.30, indicating good item discrimination.

Table 4.2.1 Summary of Point-biserial Correlations for NeSA-R

	Item Point-biserial Correlation							
Grade	≤=0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	0	1	7	17	18	2	0	45
4	0	0	10	27	7	1	0	45
5	0	1	10	26	11	0	0	48
6	0	1	11	22	14	0	0	48
7	0	0	14	23	10	1	0	48
8	1	3	13	19	14	0	0	50
11	0	2	5	25	17	1	0	50

Table 4.2.2 Summary of Point-biserial Correlations for NeSA-M

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	0	0	3	21	24	2	0	50
4	0	1	3	18	25	8	0	55
5	0	0	9	10	25	11	0	55
6	0	0	3	14	24	17	0	58
7	0	0	1	14	32	11	0	58
8	0	0	3	17	30	10	0	60
11	0	0	1	10	33	15	1	60

Table 4.2.3 Summary of Point-biserial Correlations for NeSA-S

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
5	0	3	9	25	13	0	0	50
8	0	2	11	27	17	3	0	60
11	0	2	9	24	25	0	0	60

4.3 PERCENT SELECTING EACH RESPONSE OPTION

This index 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 MC items with a single correct response. Please see the detailed summary statistics for each grade and content area in Appendices F, G, and H.

4.4 POINT-BISERIAL CORRELATIONS OF RESPONSE OPTIONS

This index describes the relationship between selecting a response option for a specific item and performance on the entire test. 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 MC items. Please see the detailed summary statistics for each grade and content area in Appendices F, G, and H.

4.5 PERCENT OF STUDENTS OMITTING AN ITEM

This statistic is useful for identifying problems with testing time and test layout. 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 Appendix F. Detailed results of the item analyses for the NeSA-M operational items are presented in Appendix G. Detailed results of the item analyses for the NeSA-S operational items are presented in Appendix H. 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.

5. RASCH ITEM CALIBRATION

The psychometric model used for the NeSA is based on the work of Georg Rasch (1960). Rasch models have had a long-standing presence in applied testing programs and have been the methodology used to calibrate NeSA items in recent history. Rasch models have several advantages over true-score test theory, so it has become the standard procedure for analyzing item response data in large-scale assessments. However, Rasch models have a number of strong requirements related to dimensionality, local independence, and model-data fit. Resulting inferences derived from any application of Rasch models rests strongly on the degree to which the underlying requirements are met.

Generally, item calibration is the process of estimating a difficulty-parameter estimate to each item on an assessment so that all items are placed onto a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch requirements, and summarizes Rasch item statistics for the 2016 NeSA Reading, Mathematics, and Science assessments.

5.1 DESCRIPTION OF THE RASCH MODEL

The Rasch dichotomous model was used to calibrate the NeSA items. All NeSA assessments contain only MC items. According to the Rasch model, the probability of answering an item correctly is based on the difference between the ability of the student and the difficulty of the item. The Rasch model places both student ability and item difficulty (estimated in terms of log-odds, or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of a person's ability that are independent of the items employed in the assessment and conversely, estimates item difficulty independently of the sample of examinees (Rasch, 1960; Wright & Panchapakesan, 1969). (As noted in Chapter Four, interpretation of item *p*-values confounds item difficulty and student ability.) Appendix I contains a more detailed overview of Rasch measurement.

5.2 CHECKING RASCH ASSUMPTIONS

Since the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the NeSA, the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. It should be noted that only operational items were analyzed since they are the basis of student scores.

Unidimensionality: Rasch models assume that one dominant dimension determines the difference among students' performances. Principal components analysis (PCA) of residuals can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant component(s) exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

Tables 5.2.1, 5.2.2, and 5.2.3 present the PCA of residuals results for the reading, mathematics, and science assessments, respectively. The results include the eigenvalues and the percentage of variance explained for the first five components. As can been seen in Table 5.2.1, the primary dimension for NeSA-R explained about 22 percent to 26 percent of the total variance across Grades 3–8 and 11. The eigenvalues of the second dimension ranged from 1.4 to 1.7. This indicates that the second dimension accounted for only 1.4 to 1.7 units out of 66 - 84 units of total variance. Similar patterns are observed for the Mathematics and the Science test. Overall, the PCA suggests that there is one clearly dominant dimension for each NeSA assessment.

Table 5.2.1 Results from PCA of Residuals – Reading

Grade	Component	Eigenvalue	Explained Variance
3	measures	13.5	23.1%
	1	1.6	3.6%
	2	1.5	3.4%
	3	1.3	2.9%
	4	1.2	2.8%
	5	1.2	2.7%
4	measures	11.8	20.8%
	1	1.5	3.3%
	2	1.3	3.0%
	3	1.2	2.8%
	4	1.2	2.7%
	5	1.2	2.7%
5	measures	13.1	21.5%
	1	1.6	3.3%
	2	1.4	2.9%
	3	1.3	2.8%
	4	1.2	2.5%
	5	1.2	2.4%
6	measures	13.6	22.0%
	1	1.7	3.5%
	2	1.3	2.8%
	3	1.3	2.8%
	4	1.2	2.5%
	5	1.2	2.5%

Grade	Component	Eigenvalue	Explained Variance
7	measures	15.3	24.1%
	1	1.6	3.3%
	2	1.3	2.8%
	3	1.2	2.6%
	4	1.2	2.5%
	5	1.2	2.4%
8	measures	13.8	21.6%
	1	1.7	3.5%
	2	1.3	2.5%
	3	1.3	2.5%
	4	1.2	2.4%
	5	1.1	2.3%
11	measures	14.1	22.0%
	1	1.5	3.1%
	2	1.5	2.9%
	3	1.3	2.6%
	4	1.2	2.5%
	5	1.2	2.4%

Table 5.2.2 Results from PCA of Residuals – Mathematics

Grade	Component	Eigenvalue	Explained Variance
3	measures	15.7	24.00%
	1	1.9	3.80%
	2	1.6	3.30%
	3	1.5	3.00%
	4	1.4	2.90%
	5	1.4	2.80%
4*	measures	16.9	23.50%
	1	1.9	3.40%
	2	1.7	3.20%
	3	1.5	2.80%
	4	1.4	2.60%
	5		

Grade	Component	Eigenvalue	
5	measures	17.6	24.20%
	1	1.9	3.40%
	2	1.7	3.10%
	3	1.6	2.80%
	4	1.4	2.60%
	5	1.4	2.50%
6	measures	19.8	25.40%
	1	1.7	3.00%
	2	1.5	2.60%
	3	1.5	2.50%
	4	1.4	2.30%
	5	1.3	2.20%
7	measures	18.8	24.50%
	1	1.7	3.00%
	2	1.6	2.70%
	3	1.5	2.50%
	4	1.4	2.50%
	5	1.4	2.40%
8	measures	18.7	23.70%
	1	1.8	3.00%
	2	1.6	2.60%
	3	1.5	2.40%
	4	1.3	2.20%
	5	1.3	2.10%
11	measures	21.5	26.30%
	1	2	3.30%
	2	1.7	2.80%
	3	1.4	2.40%
	4	1.4	2.30%
	5	1.4	2.30%

*Only four components with eigenvalues greater than one were extracted.

Table 5.2.3 Results from PCA of Residuals – Science

Grade	Component	Eigenvalue	Explained Variance
5	measures	13.4	21.1%
	1	1.6	3.1%
	2	1.4	2.7%
	3	1.3	2.5%
	4	1.2	2.3%
	5	1.1	2.3%
8	measures	15.8	20.9%
	1	1.7	2.9%
	2	1.5	2.5%
	3	1.3	2.1%
	4	1.2	2.0%
	5	1.2	2.0%
11	measures	16.8	21.9%
	1	1.9	3.1%
	2	1.4	2.4%
	3	1.3	2.1%
	4	1.3	2.1%
	5	1.2	2.0%

Local Independence: Local independence (LI) is a fundamental assumption of Rasch Measurement. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. Many indicators of LI are framed by the form of local independence proposed by McDonald (1979) that the conditional covariances of all pairs of item responses, conditioned on the abilities, are required to be equal to zero.

Residual item correlations provided in WINSTEPS for each item pair were used to assess local dependence among the NeSA items. Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It should be noted that the raw score residual correlation essentially corresponds to Yen's *Q3* index, a popular LI statistic. The expected value for the *Q3* statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected *Q3* values should be approximately −0.02 for the NeSA tests (since most of the NeSA tests had more than 50 core items). Index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Tables 5.2.4 – 5.2.6 show the summary statistics—median, interquartile range (IQR), minimum, maximum, and several percentiles (P10, P25, P50, P75, P90)—

for all the residual correlations for each test. The total number of item pairs (*N*) and the number of pairs with the residual correlations greater than 0.20 are also reported in this table. The median residual correlations were slightly negative and the values were close to −0.02. The vast majority of the correlations were very small; suggesting local item independence generally holds for the NeSA reading, mathematics, and science assessments.

Table 5.2.4 Summary of Item Residual Correlations for NeSA-R

	Reading						
Statistics	3	4	5	6	7	8	11
N	990	990	1128	1128	1128	1225	1225
Median	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
IQR	0.03	0.02	0.02	0.02	0.02	0.02	0.02
Minimum	-0.08	-0.07	-0.08	-0.11	-0.08	-0.08	-0.08
P10	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
P25	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
P50	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
P75	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
P90	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Maximum	0.29	0.07	0.14	0.12	0.07	0.12	0.14
>0.20	1	0	0	0	0	0	0

Table 5.2.5 Summary of Item Residual Correlations for NeSA-M

	Mathematics						
Statistics	3	4	5	6	7	8	11
N	1225	1485	1485	1653	1653	1770	1770
Median	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
IQR	0.03	0.03	0.03	0.03	0.02	0.02	0.03
Minimum	-0.09	-0.08	-0.09	-0.08	-0.09	-0.09	-0.12
P10	-0.05	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04
P25	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03
P50	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
P75	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
P90	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Maximum	0.44	0.47	0.33	0.17	0.33	0.21	0.27
>0.20	6	8	4	0	2	1	3

Table 5.2.6 Summary of Item Residual Correlations for NeSA-S

	Science		
Statistics	5	8	11
N	1225	1770	1770
Median	-0.02	-0.02	-0.02
IQR	0.02	0.02	0.02
Minimum	-0.07	-0.09	-0.09
P10	-0.04	-0.04	-0.04
P25	-0.03	-0.03	-0.03
P50	-0.02	-0.02	-0.02
P75	-0.01	-0.01	-0.01
P90	0.00	0.01	0.01
Maximum	0.14	0.09	0.27
>0.20	0	0	1

Item Fit: WINSTEPS provides two item fit statistics (infit/weighted and outfit/unweighted) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic with each statistic having an expected value of 1 and a different variance for each mean square or as a standardized statistic (Zstd with an expected mean = 0 and expected variance = 1).

MnSq values are more difficult to interpret due to an asymmetrical distribution and unique variance, while Zstd values are more oriented toward standardized statistical significance. Though both are informative, the Zstd values are less likely to be sensitive to the large sample sizes and have better distributional properties (Smith, Schumacker, & Bush, 1998). In the case of the NeSA-RMS, the sample sizes can be considered large (n > 5,000). The outfit statistic tends to be affected more by unexpected responses far from the person, item, or rating scale category measure (i.e., it is more sensitive to outlying, off-target, and low information responses that are very informative with regard to fit). The infit statistic tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., with more information, but contributing little to the understanding of fit

The expected MnSq value is 1.0 and can range from 0 to positive infinity. Values greater than 1.0 can be interpreted as indicating the presence of noise or lack of fit between the responses and the measurement model. Values less than 1.0 can be interpreted as item consistency or overfitting (too predictable and/or too much redundancy). Rules of thumb regarding “practically significant” MnSq values vary from author to author. More conservative users might prefer items with MnSq values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values from 0.5 to 1.5. In the results below, values outside of 0.7 to 1.3 are given practical importance.

The expected Zstd value is 0.0 with an expected *SD* of 1.0 and can effectively range from -9.99 to +9.99 in WINSTEPS. Fit values greater than 0.0 can be interpreted as indicating the presence of noise or lack of fit between the items and the model (underfitting). Fit values lower than 0.0 can be interpreted as item redundancy or overfitting items (too predictable and/or too much redundancy). Rules of thumb regarding "practically significant" Zstd values vary from author to author. More conservative users might prefer items with Zstd values that range from -2 to +2. Others believe reasonable test results can be achieved with values from -3 to +3. In the results below, values outside of -3 to +3 are given practical importance.

Table 5.2.7 lists the summary statistics of infit and outfit mean square statistics for the NeSA reading, mathematics, and science tests, including the mean, *SD*, and minimum and maximum values. The number of items within the range of [0.7, 1.3] is also reported in Table 5.2.7. As can be seen, the mean values for both fit statistics were close to 1.00 for all tests. Most of the items had infit values falling in the range of [0.7, 1.3]. Though more outfit values fell outside this range than infit values, it is not surprising given that the infit statistic mutes the effects of anomalous response by extreme students.

Table 5.2.8 lists the summary statistics of infit and outfit Zstd statistics for the NeSA reading, mathematics, and science tests, including the mean, *SD*, and minimum and maximum values. The number of items within the range of [-3, +3] is also reported in Table 5.2.8. As can be seen, the mean values for both fit statistics were variable, ranging from -1.22 to 0.65. The fact that 16 of the 17 infit means were negative and 15 of the 17 outfit means were negative suggests that on average the data overfit the Rasch model, i.e. the data were a bit more consistent than expected by the probabilistic model.

Table 5.2.7 Summary of Infit and Outfit Mean Square Statistics for 2016 NeSA Tests

		Infit Mean Square					Outfit Mean Square				
		Mean	*SD*	MIN	MAX	[0.7, 1.3]	Mean	*SD*	MIN	MAX	[0.7, 1.3]
Reading	3	1.02	0.09	0.80	1.24	45/45	1.00	0.17	0.72	1.50	43/45
	4	0.99	0.10	0.66	1.18	44/45	1.00	0.15	0.46	1.28	43/45
	5	1.01	0.11	0.67	1.23	47/48	0.99	0.17	0.53	1.47	45/48
	6	1.02	0.13	0.71	1.48	47/48	1.01	0.17	0.58	1.29	44/48
	7	1.00	0.11	0.84	1.20	48/48	0.99	0.18	0.67	1.49	45/48
	8	1.02	0.11	0.76	1.21	50/50	1.02	0.19	0.57	1.62	44/50
	11	1.01	0.11	0.79	1.28	50/50	1.01	0.18	0.69	1.38	45/50
Mathematics	3	1.00	0.08	0.82	1.19	50/50	0.99	0.13	0.65	1.33	46/50
	4	1.01	0.11	0.80	1.25	55/55	1.03	0.20	0.72	1.70	50/55
	5	1.01	0.11	0.85	1.38	54/55	1.03	0.20	0.69	1.75	49/55
	6	1.01	0.11	0.79	1.27	58/58	1.01	0.19	0.68	1.50	52/58
	7	1.01	0.10	0.84	1.31	57/58	1.01	0.17	0.70	1.47	52/58
	8	1.01	0.10	0.82	1.24	60/60	1.02	0.19	0.66	1.69	56/60
	11	0.99	0.11	0.74	1.25	60/60	1.00	0.20	0.61	1.54	53/60

		Infit Mean Square					Outfit Mean Square				
		Mean	*SD*	MIN	MAX	[0.7, 1.3]	Mean	*SD*	MIN	MAX	[0.7, 1.3]
Science	5	1.00	0.10	0.78	1.20	50/50	1.00	0.16	0.68	1.34	47/50
	8	1.00	0.10	0.79	1.24	60/60	0.99	0.15	0.67	1.33	57/60
	11	0.99	0.10	0.76	1.19	60/60	0.98	0.16	0.61	1.28	58/60

Table 5.2.8 Summary of Infit and Outfit Z STD Statistics for 2016 NeSA Tests

		Infit Z STD					Outfit Z STD				
		Mean	*SD*	MIN	MAX	[-3.0, 3.0]	Mean	*SD*	MIN	MAX	[-3.0, 3.0]
Reading	3	1.45	7.48	-9.90	9.90	8/45	0.17	7.95	-9.90	9.90	8/45
	4	0.32	8.14	-9.90	9.90	6/45	0.77	8.19	-9.90	9.90	7/45
	5	0.80	8.19	-9.90	9.90	4/48	-0.24	8.01	-9.90	9.90	7/48
	6	1.35	8.17	-9.90	9.90	5/48	1.20	8.07	-9.90	9.90	6/48
	7	-0.41	8.43	-9.90	9.90	5/48	-1.00	8.29	-9.90	9.90	7/48
	8	1.25	8.23	-9.90	9.90	3/50	0.86	8.07	-9.90	9.90	10/50
	11	0.61	7.78	-9.90	9.90	10/50	0.47	7.76	-9.90	9.90	10/50
Mathematics	3	-0.10	7.61	-9.90	9.90	11/50	-0.43	6.99	-9.90	9.90	11/50
	4	0.43	8.14	-9.90	9.90	8/55	0.24	7.77	-9.90	9.90	9/55
	5	-0.40	7.92	-9.90	9.90	6/55	-0.08	7.72	-9.90	9.90	10/55
	6	-0.48	8.12	-9.90	9.90	7/58	-0.77	7.84	-9.90	9.90	7/58
	7	-0.16	8.04	-9.90	9.90	10/58	-0.24	7.76	-9.90	9.90	9/58
	8	0.49	8.24	-9.90	9.90	4/60	0.51	7.73	-9.90	9.90	10/60
	11	-0.52	8.25	-9.90	9.90	7/60	-0.49	7.41	-9.90	9.90	11/60
Science	5	-0.04	7.77	-9.90	9.90	11/50	-0.08	7.85	-9.90	9.90	9/50
	8	0.21	8.00	-9.90	9.90	9/60	-0.52	7.95	-9.90	9.90	8/60
	11	-0.39	8.09	-9.90	9.90	10/60	-0.70	8.03	-9.90	9.90	10/60

5.3 RASCH ITEM STATISTICS

WINSTEPS 3.91.02 program (Linacre, 2016) was used for item calibration. The characteristics of calibration samples are reported in Chapter Three. These samples only include the students who attempted the tests. All omits (no response) and multiple responses (more than one response selected) were scored as incorrect answers (coded as 0s) for calibration.

As noted earlier, the Rasch model expresses item difficulty (and student ability) in units referred to as *logits* rather than on the proportion-correct metric. Large negative logits represent easier items while large positive logits represent more difficult items. The logit metrics is an interval scale, meaning that two items with logit difficulties of 0.0 and +1.0 have the same difference in difficulty as two items with logit difficulties of +3.0 and +4.0.

Appendices J, K, L, and M report the Rasch calibration summaries and logit difficulties for all the operational items. Table 5.3.1 summarizes the Rasch logit difficulties of the operational items on each test. The minimum and maximum values and standard deviations suggest that the NeSA items covered a relatively wide range of difficulties. It is important to note that the logit difficulty values presented have not been linked to a common scale of measurement. Therefore, the relative magnitude of the statistics across subject areas and grades cannot be compared. The item pool was then updated with the item statistics.

Table 5.3.1 Summary of Rasch Item Difficulties for NeSA-R, NeSA-M, and NeSA-S

	Grade	*N*	Mean	*SD*	Min	Max	Range
Reading	**3**	45	-0.08	0.73	-1.63	1.19	2.83
	4	45	-0.25	0.62	-1.61	1.16	2.76
	5	48	-0.18	0.70	-1.58	1.14	2.72
	6	48	-0.27	0.72	-1.68	1.17	2.85
	7	48	-0.16	0.82	-1.74	1.65	3.39
	8	50	-0.45	0.75	-1.43	1.04	2.47
	11	50	-0.69	0.71	-2.14	0.95	3.09
Mathematics	**3**	50	-0.44	0.62	-1.66	1.03	2.69
	4	55	-0.44	0.61	-1.71	0.90	2.60
	5	55	-0.37	0.65	-1.65	0.95	2.60
	6	58	-0.39	0.61	-1.67	1.12	2.78
	7	58	-0.44	0.55	-1.49	0.76	2.25
	8	60	-0.50	0.61	-1.48	0.90	2.37
	11	60	-0.52	0.55	-1.71	0.53	2.24
Science	**5**	50	-0.73	0.72	-2.11	0.72	2.83
	8	60	-0.69	0.67	-2.37	0.78	3.14
	11	60	-0.65	0.66	-2.16	0.89	3.05

6. EQUATING AND SCALING

As discussed earlier in Chapter 2, the 2016 test forms were constructed with items that were either field tested, or used operationally on a previously administered NeSA test. NeSA assessments are constructed each year allowing each NeSA assessment to be different from the previous year's assessment. To ensure that all forms for a given grade and content area provide comparable scores, and to ensure the passing standards across different administrations are equivalent, the new operational items need to be placed on the bank scale via equating to bring the 2016 NeSA raw-score-to-Rasch-ability scale to the previous operational scale. When the new 2016 NeSA tests are placed on the bank's scale, the resulting scale scores for the new test form will be the same as the scale scores of the previous operational form such that students performing at the same level of (underlying) achievement should receive the same score (i.e., scale score). The resulting scale scores will be used for score reporting and performance level classification. Once operational items are equated, field test items are then placed on the bank scale and are then ready for future operational use.

This chapter begins with a summary of the entire NeSA equating procedures. This is followed by a scaling analysis that transforms raw scores to scale scores that represent the same skill level on every test form. Some summary results of the state scale score performance are also provided.

6.1 EQUATING

The equating design employed for NeSA is often referred to as a common-item non-equivalent groups (CINEG) design, which uses a set of anchor items that appear on two forms to adjust for differences in test difficulty across years. As discussed earlier, the 2016 NeSA test forms were constructed with items from previous administrations. The items were previously either field-test or operational items. If the item difficulty estimated from the previous administrations are within estimation error for the current administration, the entire set of the 2016 NeSA operational items can serve as the linking set. This means that the raw to scale score conversion tables can be established prior to the operational administration. This is often referred to as the pre-equating process because it is conducted before the operational test is administered. The most appealing feature of the pre-equating process, when applicable, is its ability to facilitate immediate score reporting for tests which have tight reporting windows.

However, it may not be appropriate to assume that the operational items will maintain their relative difficulty across administrations. The same item can perform differently across administrations due to changes in the item's position or changes in the students' experiences. Once the 2016 operational test data was available, DRC Psychometric Services staff, together with NDE, evaluated the item difficulty equivalence using a post-equating check procedure (Robust *Z*) to identify items that show significant difficulty changes from the bank values. If no unstable items are identified, the 2016 equating process would result in the pre-equating solution. On the other hand, if an item or items are found to be outside the normal estimation error, a post-equated solution would be used. The sub-set of 2016 operational items, with those identified items excluded, was used as the set to estimate the link constant to map the 2016 test to the bank scale. This equating process is known as the post-equating because the equating occurs after the administration of the operation test and the raw-to-scale-score conversion is generated based on the operational test data.

As part of the post-equating check procedures, DRC Psychometric Services staff evaluated the item difficulty equivalence by comparing the old banked item calibration (called pre-calibration) with a new unanchored calibration of the 2016 data (called post-calibration). The evaluations were conducted for each grade and content area, using statistical methods.

DRC Psychometric Services employed the Robust *Z* statistic (Huynh, 2000; Huynh & Rawls, 2009) for the post-equating check. This method focuses on the correlations between the pre- and post-calibrated item difficulties, and the ratio of standard deviations (*SD*) between the two calibrations. The correlation between the two estimates of item difficulty should be 0.95 or higher and the ratio of standard deviations between the two sets of estimates of the item difficulty should range between 0.90 and 1.10 (Huynh & Meyer, 2010). To detect inconsistent item difficulty estimates, a critical value for the Robust *Z* statistic of ±1.645 was used.

Table 6.1.1 contains these statistics of correlation and SD ratio for the 2016 NeSA-M test. The item difficulty correlation for Grade 5 is the only statistic that falls below the criteria defined above. Appendices N – P contain the same statistics for each grade and content combination.

Table 6.1.1 NeSA-M Pre- and Post-Equating Comparison

	Grade						
	3	4	5	6	7	8	11
Correlation	0.96	0.94*	0.97	0.98	0.97	0.97	0.96
SD pre	0.62	0.62	0.65	0.61	0.55	0.61	0.55
SD post	0.62	0.57	0.60	0.58	0.52	0.57	0.57
SD Ratio	1.01	0.92	0.92	0.95	0.94	0.94	1.04

*The Grade 4 correlation was the only value that didn't meet the Robust Z criteria

Across all three content areas, the test forms with values below the ideal ranges of Robust *Z* correlation, or *SD* ratio values were further evaluated by the NDE in determining whether to include items that exceeded the Robust *Z* critical value of ±1.645 in the linking set used for the post-equating. Items that exceeded the Robust *Z* critical value were then deleted, one item at a time, until both the item difficulty correlation and the *SD* ratio fell within the prescribed limits.

To summarize the 2016 NeSA test equating solutions, NDE decided to adopt a post-equating results for NeSA-M Grade 4, NeSA-R grades 3, 4, 5, and 11. For these tests, test equating was adjusted by excluding the items exceeding the critical value until the Robust Z criteria were met. A new raw-to-scale-score conversion table was created for these tests. For the other grades and content areas, NDE decided to use a pre-equating solution, keep the whole set of operational items in the linking set and then apply to the existing raw-to-scale-score conversion table.

6.2 SCALING

The purpose of a scaling analysis is to create a scale score for test reporting. The basic score on any test is the raw score, which is the number of items answered correctly or the total score points earned. However, the raw score alone does not present a wide-ranging picture of test performance because it is not on an equal-interval scale and can be interpreted only in terms of a particular set of items. Since a given raw score may not represent the same skill level on every test form, scale scores were assigned to each raw score point to adjust for slight shifts in item difficulties and permit valid comparison across all test administrations within a particular content area.

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 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 Q 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 2016, Appendix R for NeSA-M for 2016 and Appendix S for NeSA-S 2016. 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 scale scores.

The scale chosen, for all grades and content areas of the NeSA assessment, range from 0 to 200. The value of 0 is reserved for students who were not tested or were otherwise invalidated. Thus, 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 calculation. It is possible that a future form will be easy enough that the upper limit of 200 is not invoked even for a perfect paper or could be difficult enough that the lower limit is not invoked.

As part of its deliberations concerning defining the performance levels, the State Board of Education specified that the *Meets the Standards* performance level have a scale score of 85 and that the *Exceeds the Standards* level have a scale score of 135. The logit standards defining the performance levels were adopted by the SBE per the standard setting and standard validation completed in 2010 for NeSA-R, in 2011 for NeSA-M, and in 2012 for NeSA-S.

Complete documentation of all standard setting events are presented in separate documents and are placed on the Nebraska State Department of Education website labeled:

2010 NeSA-Reading Standard Setting Technical Report, http://www.education.ne.gov/Assessment/pdfs/2010_NeSA_Reading_Standard_Setting_Tech_%20Report.pdf ,

2011 NeSA-Mathematics Standard Setting Technical Report, http://www.education.ne.gov/Assessment/pdfs/2011_NeSA_Math_Standard_Setting_Tech_Report.pdf

and *2012 NeSA-Science Standard Setting Technical Report*, http://www.education.ne.gov/Assessment/pdfs/Final_NeSA_Science_Standard_Setting_Tech_Report_October_2012.pdf

Given the scale score and the logit standards defining the performance level, it is sufficient to define the final scale score metric. To ensure proper rounding on all future forms, the calculations used 84.501 and 134.501 as the scale score performance standards. The arithmetic was done using logits rounded to four decimals and the final constants for the slope and intercept of the transformation were rounded to five. Scale scores are rounded to whole numbers.

The transformation to scale scores is:

1. SS = $a + b$ * logit where:
2. $b = \frac{134.501 - 84.501}{x_E - x_M}$ where x_E is the logit for *Exceeds Standards* and x_M is the logit for *Meets Standards*.
3. $a = 84.501 - bx_M$ or $a = 134.501 - bx_E$.

Calculations of the slopes and intercepts for all grades of the NeSA-R scale score conversion are given in Table 6.2.1, for NeSA-M 6.2.2, and for NeSA-S 6.2.3. The raw-to-scale conversions are provided in Appendices Q, R, and S.

Table 6.2.1 NeSA-R 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	-0.5168	1.2340	1 to 84	85-134	135 to 200	28.55837	99.25997
4	-0.5117	0.8591	1 to 84	85-134	135 to 200	36.47505	103.16528
5	-0.4122	0.8560	1 to 84	85-134	135 to 200	39.42751	100.75302
6	-0.4331	0.8924	1 to 84	85-134	135 to 200	37.72161	100.83823
7	-0.5104	0.7855	1 to 84	85-134	135 to 200	38.58471	104.19271
8	-0.4812	0.8712	1 to 84	85-134	135 to 200	36.97131	102.29159
11	-0.4103	0.8508	1 to 84	85-134	135 to 200	39.64793	100.76854

Table 6.2.2 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	-0.6	1.1000	1 to 84	85-134	135 to 200	29.41176	102.15706
4	-0.6	1.2000	1 to 84	85-134	135 to 200	27.77778	101.17667
5	-0.57	1.1597	1 to 84	85-134	135 to 200	28.90675	100.98685
6	-0.47	1.1816	1 to 84	85-134	135 to 200	30.27367	98.73862
7	-0.45	1.2500	1 to 84	85-134	135 to 200	29.41176	97.74529
8	-0.4	1.3000	1 to 84	85-134	135 to 200	29.41176	96.2747
11	-0.29	1.1000	1 to 84	85-134	135 to 200	35.97122	94.94165

Table 6.2.3 NeSA-S 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
5	-0.4971	1.0580	1 to 84	85-134	135 to 200	32.15095	100.49331
8	-0.4543	1.0378	1 to 84	85-134	135 to 200	33.50958	99.73252
11	-0.5407	1.3130	1 to 84	85-134	135 to 200	26.97256	99.09502

Complete frequency distributions of the state scale scores for the NeSA-R, NeSA-M, and NeSA-S are provided in Appendices Q, R, and S as part of the raw-to-scale-score conversion tables. A simple summary of the reading, mathematics, and science distributions can be found in Tables 6.2.4, 6.2.5, and 6.2.6.

Table 6.2.4 2016 NeSA-R State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	21887	120.2	31.9	98	119	141
4	23039	123.4	36.7	96	122	149
5	22689	129.6	40.0	101	132	157
6	22915	125.4	39.9	98	124	152
7	22598	129.7	39.3	102	133	156
8	22220	119.1	37.6	92	121	143
11	21400	111.4	41.8	84	114	139

Table 6.2.5 2016 NeSA-M State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
3	22028	114.8	35.2	89	113	137
4	23143	112.7	33.5	88	110	134
5	22799	112.1	35.6	85	108	134
6	22997	111.8	39.2	82	108	138
7	22716	108.6	37.9	81	105	134
8	22320	105.3	36.9	77	101	129
11	21357	104.7	47.0	69	103	138

Table 6.2.6 2016 NeSA-S State Scale Score Summary, All Students

Grade	Count	Scale Score		Quartile		
		Mean	S.D.	First	Second	Third
5	22798	105.2	32.6	83	104	126
8	22328	103.8	34.4	79	102	127
11	21366	103.7	27.8	84	104	122

7. FIELD TEST ITEM DATA SUMMARY

As noted in Chapter Two, in addition to the operational items, field test items were embedded in all content areas and grade level assessments in order to expand the item pool for future form development. Field test items are items being administered for the first time to gather statistical information. These items do not count toward an individual student's score. All field tested items were analyzed statistically following classical item analysis methods including proportion correct, point-biserial correlation, and DIF.

7.1 CLASSICAL ITEM STATISTICS

Indices known as classical item statistics included the item *p*-value and the point-biserial correlations for MC items. For MC items, the *p*-value reflects the proportion of students who answered the item correctly. In general, more capable students are expected to respond correctly to easy items and less capable students are expected to respond incorrectly to difficult items. The primary way of detecting such conditions is through the point-biserial correlation coefficient for dichotomous (MC) items. The point-biserial correlation will be positive if the total test mean score is higher for the students who respond correctly to MC items and negative when the reverse is true.

The traditional statistics are computed for each NeSA-R field test item in Appendix F, for NeSA-M in Appendix G and for NeSA-S in Appendix H. Tables 7.1.1, 7.1.2, and 7.1.3 provide summaries of the distributions of item proportion correct and point-biserial correlations. For future form construction, items with negative point-biserial correlations are never considered for operational use. Items with correlations less than 0.2 or proportion correct less than 0.3 or greater 0.9 are avoided when possible. In reading the following tables, the heading ≤ 0.1 descrbes items between 0.0 and 0.1, and the heading ≤ 0.2 descirbes items between 0.1 and 0.2, etc.

Table 7.1.1 Summary of Statistics for NeSA-R 2016 Multiple Choice Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	0	3	3	11	11	10	9	2	1	0.582	50
4	0	0	1	4	8	4	14	9	6	3	0.641	49
5	0	0	2	3	4	10	11	9	8	2	0.632	49
6	0	0	2	3	7	8	10	10	10	0	0.629	50
7	0	0	1	3	11	14	5	9	6	1	0.602	50
8	0	0	0	1	5	7	10	17	10	0	0.685	50
11	0	0	1	7	5	4	13	9	10	1	0.634	50

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	2	4	8	22	14	0	0	50
4	1	5	8	17	17	1	0	49
5	2	5	14	19	9	0	0	49
6	3	7	10	18	10	2	0	50
7	5	6	9	18	12	0	0	50
8	1	3	13	19	14	0	0	50
11	0	6	9	14	15	6	0	50

Table 7.1.2 Summary of Statistics for NeSA-M 2016 Multiple Choice Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	1	0	1	3	9	6	14	12	4	0.707	50
4	0	0	0	4	4	5	5	14	12	6	0.708	50
5	0	0	0	4	6	5	11	11	12	1	0.671	50
6	0	0	1	2	5	8	12	13	8	1	0.667	50
7	0	0	0	5	10	13	10	8	3	1	0.589	50
8	0	0	1	2	10	11	12	9	4	1	0.607	50
11	0	2	2	8	15	7	7	9	0	0	0.511	50

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	1	4	4	13	23	5	0	50
4	0	1	6	17	20	4	2	50
5	0	1	12	15	17	5	0	50
6	0	0	6	17	19	8	0	50
7	1	0	6	8	25	10	0	50
8	0	0	8	13	20	8	1	50
11	3	3	6	13	16	9	0	50

Table 7.1.3 Summary of Statistics for NeSA-S 2016 Multiple Choice Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
5	0	0	1	1	1	4	4	5	20	14	0.798	50
8	0	0	4	6	9	12	7	7	4	1	0.553	50
11	0	2	3	9	16	8	6	6	0	0	0.484	50

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
5	2	9	12	15	12	0	0	50
8	2	4	12	16	15	1	0	50
11	7	8	13	15	7	0	0	50

Table 7.1.4 Summary of Statistics for NeSA-R 2016 Evidence Based Selected Response Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	0	1	2	1	0	0	0	0	0	0.331	4
4	0	0	1	2	0	0	2	0	0	0	0.469	5
5	0	0	0	2	2	2	0	1	0	0	0.499	7
6	0	0	0	0	2	2	2	0	0	0	0.565	6
7	0	0	1	1	4	2	0	0	0	0	0.451	8
8	0	1	1	0	0	0	4	0	0	0	0.507	6
11	0	0	2	0	1	0	3	0	1	0	0.527	7

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	0	1	1	1	1	0	0	4
4	0	0	2	1	1	1	0	5
5	0	1	2	2	2	0	0	7
6	0	0	1	2	3	0	0	6
7	1	0	1	3	3	0	0	8
8	1	1	1	0	2	1	0	6
11	0	1	1	1	1	2	1	7

Table 7.1.5 Summary of Statistics for NeSA-R 2016 Multi-Select Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	0	0	0	0	2	2	1	0	0	0	0.543	5
4	0	0	0	1	2	0	1	1	0	0	0.532	5
5	0	0	0	0	0	1	1	0	1	0	0.676	3
6	0	0	0	0	1	0	2	1	1	0	0.686	5
7	0	0	0	0	0	1	2	1	1	0	0.690	5
8	0	0	0	0	0	1	2	2	0	0	0.661	5
11	0	0	0	0	1	1	1	0	0	0	0.561	3

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	0	0	1	1	3	0	0	5
4	0	0	1	3	0	1	0	5
5	0	0	0	1	1	1	0	3
6	0	0	1	1	3	0	0	5
7	0	0	0	2	2	1	0	5
8	0	0	0	1	3	1	0	5
11	0	0	0	2	1	0	0	3

Table 7.1.6 Summary of Statistics for NeSA-R 2016 R Technology Enhanced Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3	1	0	0	2	2	2	1	2	1	0	0.520	11
4	0	1	0	1	1	1	3	2	2	0	0.608	11
5	0	0	0	1	1	2	1	5	1	0	0.654	11
6	0	0	0	1	2	0	1	4	1	0	0.638	9
7	0	0	0	0	0	2	2	2	1	0	0.684	7
8	0	1	0	1	2	3	2	0	0	0	0.484	9
11	0	0	0	1	2	0	2	1	3	0	0.636	9

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3	2	0	1	5	2	1	0	11
4	0	0	1	3	6	1	0	11
5	0	0	1	5	5	0	0	11
6	0	0	0	4	5	0	0	9
7	0	0	0	1	4	2	0	7
8	1	0	5	2	1	0	0	9
11	0	1	0	2	5	1	0	9

Table 7.1.7 Summary of Statistics for NeSA-R 2016 R Text Dependent Analysis Field Test Items

	Item Proportion Correct											
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9	Mean	Total
3												0
4												0
5	0	0	4	1	0	0	0	0	0	0	0.298	5
6	0	0	4	1	0	0	0	0	0	0	0.274	5
7	0	0	2	3	0	0	0	0	0	0	0.313	5
8	0	0	0	5	0	0	0	0	0	0	0.339	5
11	0	0	0	5	0	0	0	0	0	0	0.359	5

	Item Point-biserial Correlation							
Grade	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	>0.6	Total
3								0
4								0
5	0	0	0	4	1	0	0	5
6	0	0	1	4	0	0	0	5
7	0	0	0	1	4	0	0	5
8	0	0	0	0	4	1	0	5
11	0	0	0	0	1	4	0	5

7.2 DIFFERENTIAL ITEM FUNCTIONING

DIF occurs when examinees with the same ability level but different group memberships do not have the same probability of answering an item correctly. This pattern of results may suggest the presence of *item bias*. Items exhibiting DIF were referred to content specialists to determine possible bias. No statistical procedure should be used as a substitute for rigorous, hands-on reviews by content and bias specialists. The statistical results can help organize the review so the effort is concentrated on the most problematic cases. Further, no items should be automatically rejected simply because a statistical method flagged them or accepted because they were not flagged.

For MC items, the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) for detecting DIF is a commonly used technique in educational testing. The procedure as implemented by DRC contrasts a focal group with a reference group. While it makes no practical difference in the analysis which group is defined as the focal group, the group most apt to be disadvantaged by a biased measurement is typically defined as the focal group. In these analyses, the focal group was female for gender-based DIF and minority for ethnicity-based DIF; reference groups were male and white, respectively.

To assist the review committees in interpreting the analyses, the items are assigned a severity code based on the magnitude of the MH statistic. Items classified as A+ or A- have little or no statistical indication of DIF. Items classified as B+ or B- have some indication of DIF but may be judged to be acceptable for future use. Items classified as C+ or C- have strong evidence of DIF and should be reviewed and possibly rejected from the eligible item pool. The plus sign indicates that the item favors the focal group and a minus sign indicates that the item favors the reference group. Tables 7.2.1 – 7.2.3 show summaries of the DIF statistics. The first column defines the focal group. Appendices T, U, and V provide more summary information on DIF analysis.

Table 7.2.1 Summary of DIF by Code for NeSA-R Multiple Choice 2016 Field Test

Grade 3	A+	A-	B+	B-	C+	C-	FT Items
Female	27	22	1	0	0	0	50
Black	11	36	0	1	0	2	50
Hispanic	14	32	0	4	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 4	A+	A-	B+	B-	C+	C-	FT Items
Female	26	20	3	0	0	0	49
Black	7	33	0	7	0	2	49
Hispanic	12	32	0	4	0	1	49
Asian	0	0	0	0	0	0	49
American Indian/Alaskan Native	0	0	0	0	0	0	49

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	25	22	1	1	0	0	49
Black	11	33	0	4	0	1	49
Hispanic	10	34	0	4	0	1	49
Asian	0	0	0	0	0	0	49
American Indian/Alaskan Native	0	0	0	0	0	0	49

Grade 6	A+	A-	B+	B-	C+	C-	FT Items
Female	31	17	1	0	0	1	50
Black	7	34	0	7	0	2	50
Hispanic	17	28	0	4	0	1	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 7	A+	A-	B+	B-	C+	C-	FT Items
Female	30	15	2	3	0	0	50
Black	11	36	0	3	0	0	50
Hispanic	18	30	0	1	0	1	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	34	13	2	0	0	1	50
Black	9	38	0	3	0	0	50
Hispanic	14	32	0	3	0	1	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	26	19	4	0	1	0	50
Black	1	34	0	14	0	1	50
Hispanic	9	35	0	6	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Table 7.2.2 Summary of DIF by Code for NeSA-R Evidence Based Selected Response 2016 Field Test

Grade 3	A+	A-	B+	B-	C+	C-	FT Items
Female	2	2	0	0	0	0	4
Black	2	2	0	0	0	0	4
Hispanic	0	4	0	0	0	0	4
Asian	0	0	0	0	0	0	4
American Indian/Alaskan Native	0	0	0	0	0	0	4

Grade 4	A+	A-	B+	B-	C+	C-	FT Items
Female	4	1	0	0	0	0	5
Black	1	3	0	1	0	0	5
Hispanic	0	5	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	1	6	0	0	0	0	7
Black	0	7	0	0	0	0	7
Hispanic	2	5	0	0	0	0	7
Asian	0	0	0	0	0	0	7
American Indian/Alaskan Native	0	0	0	0	0	0	7

Grade 6	A+	A-	B+	B-	C+	C-	FT Items
Female	4	1	0	1	0	0	6
Black	0	6	0	0	0	0	6
Hispanic	1	4	0	1	0	0	6
Asian	0	0	0	0	0	0	6
American Indian/Alaskan Native	0	0	0	0	0	0	6

Grade 7	A+	A-	B+	B-	C+	C-	FT Items
Female	5	3	0	0	0	0	8
Black	1	6	0	0	0	1	8
Hispanic	2	6	0	0	0	0	8
Asian	0	0	0	0	0	0	8
American Indian/Alaskan Native	0	0	0	0	0	0	8

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	5	1	0	0	0	0	6
Black	0	5	0	0	0	1	6
Hispanic	2	3	0	1	0	0	6
Asian	0	0	0	0	0	0	6
American Indian/Alaskan Native	0	0	0	0	0	0	6

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	2	3	2	0	0	0	7
Black	0	5	0	0	0	2	7
Hispanic	2	5	0	0	0	0	7
Asian	0	0	0	0	0	0	7
American Indian/Alaskan Native	0	0	0	0	0	0	7

Table 7.2.3 Summary of DIF by Code for NeSA-R Multiple Select 2016 Field Test

Grade 3	A+	A-	B+	B-	C+	C-	FT Items
Female	2	3	0	0	0	0	5
Black	0	4	0	1	0	0	5
Hispanic	1	4	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 4	A+	A-	B+	B-	C+	C-	FT Items
Female	1	4	0	0	0	0	5
Black	0	3	0	1	0	1	5
Hispanic	3	2	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	2	1	0	0	0	0	3
Black	0	3	0	0	0	0	3
Hispanic	1	2	0	0	0	0	3
Asian	0	0	0	0	0	0	3
American Indian/Alaskan Native	0	0	0	0	0	0	3

Grade 6	A+	A-	B+	B-	C+	C-	FT Items
Female	4	1	0	0	0	0	5
Black	1	1	0	3	0	0	5
Hispanic	1	4	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 7	A+	A-	B+	B-	C+	C-	FT Items
Female	2	2	1	0	0	0	5
Black	0	3	0	2	0	0	5
Hispanic	0	5	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	3	2	0	0	0	0	5
Black	1	2	0	2	0	0	5
Hispanic	0	5	0	0	0	0	5
Asian	0	0	0	0	0	0	5
American Indian/Alaskan Native	0	0	0	0	0	0	5

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	2	1	0	0	0	0	3
Black	1	0	0	1	0	1	3
Hispanic	1	2	0	0	0	0	3
Asian	0	0	0	0	0	0	3
American Indian/Alaskan Native	0	0	0	0	0	0	3

Table 7.2.4 Summary of DIF by Code for NeSA-R Technology Enhanced 2016 Field Test

Grade 3	A+	A-	B+	B-	C+	C-	FT Items
Female	8	3	0	0	0	0	11
Black	0	11	0	0	0	0	11
Hispanic	3	8	0	0	0	0	11
Asian	0	0	0	0	0	0	11
American Indian/Alaskan Native	0	0	0	0	0	0	11

Grade 4	A+	A-	B+	B-	C+	C-	FT Items
Female	7	4	0	0	0	0	11
Black	0	6	0	3	0	2	11
Hispanic	3	7	0	0	0	1	11
Asian	0	0	0	0	0	0	11
American Indian/Alaskan Native	0	0	0	0	0	0	11

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	8	3	0	0	0	0	11
Black	0	6	0	5	0	0	11
Hispanic	1	10	0	0	0	0	11
Asian	0	0	0	0	0	0	11
American Indian/Alaskan Native	0	0	0	0	0	0	11

Grade 6	A+	A-	B+	B-	C+	C-	FT Items
Female	7	2	0	0	0	0	9
Black	3	4	0	1	0	1	9
Hispanic	3	6	0	0	0	0	9
Asian	0	0	0	0	0	0	9
American Indian/Alaskan Native	0	0	0	0	0	0	9

Grade 7	A+	A-	B+	B-	C+	C-	FT Items
Female	6	1	0	0	0	0	7
Black	0	5	0	1	0	1	7
Hispanic	1	5	0	1	0	0	7
Asian	0	0	0	0	0	0	7
American Indian/Alaskan Native	0	0	0	0	0	0	7

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	5	2	2	0	0	0	9
Black	1	6	0	1	0	1	9
Hispanic	4	5	0	0	0	0	9
Asian	0	0	0	0	0	0	9
American Indian/Alaskan Native	0	0	0	0	0	0	9

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	8	1	0	0	0	0	9
Black	0	4	0	3	0	2	9
Hispanic	1	8	0	0	0	0	9
Asian	0	0	0	0	0	0	9
American Indian/Alaskan Native	0	0	0	0	0	0	9

Table 7.2.5 Summary of DIF by Code for NeSA-M 2016 Multiple Choice Field Test

Grade 3	A+	A-	B+	B-	C+	C-	FT Items
Female	21	26	0	3	0	0	50
Black	8	28	1	8	0	5	50
Hispanic	18	27	0	4	0	1	50
Asian	6	3	0	1	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 4	A+	A-	B+	B-	C+	C-	FT Items
Female	27	21	0	2	0	0	50
Black	7	34	0	6	0	3	50
Hispanic	12	36	0	2	0	0	50
Asian	3	5	1	0	0	1	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	27	21	1	1	0	0	50
Black	8	27	1	11	0	3	50
Hispanic	15	28	1	4	0	2	50
Asian	3	3	0	1	0	1	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 6	A+	A-	B+	B-	C+	C-	FT Items
Female	26	24	0	0	0	0	50
Black	14	28	0	7	0	1	50
Hispanic	22	25	0	3	0	0	50
Asian	3	5	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 7	A+	A-	B+	B-	C+	C-	FT Items
Female	25	23	1	0	0	1	50
Black	8	38	0	3	0	1	50
Hispanic	11	39	0	0	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	26	21	2	1	0	0	50
Black	12	29	0	7	0	0	50
Hispanic	10	39	1	0	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	26	21	2	1	0	0	50
Black	12	29	0	7	0	0	50
Hispanic	10	39	1	0	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Table 7.2.6 Summary of DIF by Code for NeSA-S Multiple Choice 2016 Field Test

Grade 5	A+	A-	B+	B-	C+	C-	FT Items
Female	28	21	0	1	0	0	50
Black	5	20	2	15	0	8	50
Hispanic	8	31	0	9	0	2	50
Asian	0	6	0	3	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 8	A+	A-	B+	B-	C+	C-	FT Items
Female	23	22	0	5	0	0	50
Black	10	30	0	8	0	2	50
Hispanic	13	34	0	1	0	2	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

Grade 11	A+	A-	B+	B-	C+	C-	FT Items
Female	18	28	1	3	0	0	50
Black	16	30	0	4	0	0	50
Hispanic	15	35	0	0	0	0	50
Asian	0	0	0	0	0	0	50
American Indian/Alaskan Native	0	0	0	0	0	0	50

7.3 Exploring the Dimensionality of New Item Types

The new test blue-print for the English Language Arts (ELT) test that will begin administration in spring of 2017 requires the use of a text-dependent analysis (TDA) writing prompt. Five prompts per grade were field tested in spring 2016. Along with the TDA writing prompt, other new item formats were introduced that include: Evidence-based-selected-response (EBSR), Technology enhanced (TE) and Multi-select (MS). With the introduction of any new item type it is always prudent to determine if the new item types are measuring the same underlying construct as the majority of the test items that follow the multiple choice (MC) format.

This investigation was undertaken using the principle component analysis of residuals that is available in Table 23 of Winsteps (Linacre, 2016). In Rasch model analyses the modeled effects of person ability and item difficulty are removed from the actual response matrix. The remaining information, the residual matrix is then subject to a principal component analysis (PCA). In theory the extraction of the ability and difficulty variance from the response matrix should leave only uncorrelated residuals that contain no useful information. Any patterns in the residual information can be identified by using factor loadings available in the PCA analysis as well as looking for outliers in the intercorrelation matrix of the residuals. Both of these statistics are available in the Table 23 output from Winsteps.

In Spring 2016 five TDA prompts were filed-tested for each grade (Grades 5-8). The prompts were randomly administered with between 4,000 and 5,000 students taking each form. A sub-sample of the TDA responses were scored using a 4-point holistic rubric. The number of scored responses for each prompt is shown in the third column of Table 1. This number ranged from approximately 1600 to 2600. Columns 4 to 8 in table 1 show the results of the PCA analysis. Column 4 indicated the variance explained by the measures for all operational and field test items. These values indicate that there is a dominant first factor. This corresponds to an eigen value over 20. Column 5 contains the eigen value of the first remaining factor after the variance of the measures have been removed. The values range between 1.77 and 2.18. These values are not far outside the plausible range for residuals from simulated unidimensional data (Smith, 1996). This suggests that there is no strong second factor.

The factor loadings for the TDA items are found in column 6. These factor loadings range from -0.08 to +0.15, clearly indicating that the TDA items are measuring the same underlying construct as is being measured by the test as a whole.

Because the PCA analysis included all FT items, it was also possible to look at the dimensionality of other new item types. There were a total of 26 EBSR items included on the 20 FT forms across the 4 grades. There was at least one ESBR item on each form and sometimes two. The factor loadings for these items are shown in columns seven and eight of Table 1. There were five of the items with a factor loading that exceeded |.40| (Shaded gold in the table.). There were six items with a factor loading that fell between |.40| and |.30| (Shaded yellow in the table.). Finally, there were two items with a factor loading that fell between |.30| and |.25| (Shaded green in the table.). In all there were a total of 13 of the 26 EBSR items tested that had large factor loadings. This may suggest these items and not measuring the same underlying construct as the predominance of MC items that make up the test.

The results of the technology enhanced (TE) items are found in columns 9 through 11. There were between 1 and 3 TE items on each form for a total of 36 items. One of the items had a factor loading that exceeded |.50| (Shaded red in the table.). Finally, there were five items with a factor loading that fell between |.30| and |.25| (Shaded green in the table.). Overall the TE items appeared to be closer to the underlying construct measured by the test then the EBSR items.

The results of the multi select (MS) items are found in columns 12 and 13. There were between 1 or 2 MS items on each form for a total of 18 items. Overall the MS items appeared to be closer to the underlying construct measured by the test then the EBSR or the TE items. There were two items with a factor loading that fell between |.40| and |.30| (Shaded yellow in the table.).

Table 7.3.1 New Item Type Dimensionality

					Factor Loadings for New Item Types							
Grade	Form	Number of TDA Responses Scored	Variance Explained by Measures	First Remaining Contrast	TDA Item	EBSR 1	EBSR 2	TE 1	TE 2	TE 3	MS 1	MS 2
5	1	1872	26.9%	2.00	-0.03	0.39	0.28	0.01			-0.11	
5	2	1824	27.0%	2.08	-0.06	0.41		0.13	0.06	0.01		
5	3	1808	28.5%	1.77	0.05	0.06		0.13	0.05		0.24	
5	4	1600	24.7%	1.88	-0.05	0.48		-0.19	0.09		-0.10	
5	5	1588	27.1%	1.82	0.15	-0.23		-0.27	-0.05	0.04		
6	1	2046	27.0%	2.02	0.00	-0.44	-0.31	-0.17			0.08	
6	2	1964	30.1%	1.74	-0.04	-0.29		-0.25	-0.16	-0.02		
6	3	2045	24.4%	1.97	0.05	-0.15		0.22	-0.06		-0.05	
6	4	2329	26.4%	1.80	0.06	0.00		-0.14	-0.03		0.05	
6	5	2640	26.6%	2.05	0.00	0.22		-0.04			-0.01	0.10
7	1	2178	27.1%	1.94	0.04	-0.34		-0.14	-0.01		-0.14	
7	2	1801	28.5%	2.03	-0.03	0.38	0.23	-0.05			-0.31	
7	3	1697	26.3%	1.98	0.06	-0.21	-0.10	0.27			-0.17	
7	4	1725	21.5%	1.98	0.07	-0.30	-0.16	0.25			-0.09	
7	5	2067	26.8%	1.86	0.03	-0.12		0.13	0.01		-0.09	
8	1	1598	27.2%	2.18	0.07	0.02		0.15			-0.34	-0.01
8	2	2141	29.4%	2.02	-0.04	0.48	0.11	0.05	-0.06			
8	3	1732	23.2%	1.96	-0.08	0.41		0.11	0.11		0.02	
8	4	1862	25.6%	2.02	-0.02	-0.08		0.54	-0.16	0.01		
8	5	2267	22.6%	2.11	-0.02	0.35		-0.29			0.15	-0.15

7.4 Nebraska Text Dependent Analysis Field Test Handscoring Process for Holistic Scoring

The Nebraska Text Dependent Analysis (TDA) Field Test consisted of five new items and passages for each grade 5-8, and 11, scored holistically using a new TDA rubric.

Rangefinding

After receiving student responses from the 2016 NE TDA Field Test, DRC's Performance Assessment Services (PAS) staff reviewed the responses for each of the five items per grade (25 items total) and assembled them into rangefinding sets that exemplified the range of different score points available for each item. Copies of these sets were then made for each member of the rangefinding committees. DRC's staff then travelled to Lincoln, NE (May 2 – May 6, 2016) and facilitated the rangefinding sessions. The rangefinding committees consisted of Nebraska educators and Nebraska Department of Education (NDE) staff members.

The rangefinding meeting began in a joint session with an overview of the ELA transition process and a discussion of the TDA holistic rangefinding process; along with guidelines for the consensus scoring of the assembled responses. A joint review of the new TDA rubric was also presented by an NDE representative. The group then broke into five separate grade-specific committees, each consisting of nine NE educators and an NDE representative. Each committee then reviewed the passage and the prompt for the first of five items to be scored. Following this review and discussion, each committee consensus scored 40 – 50 responses selected by DRC PAS staff from the 2016 TDA Field Test.

Initially, each student response was read aloud and then discussed by all members of the group equally to ensure that everyone was interpreting the new holistic TDA rubric consistently and uniformly. Following the discussions and once a consensus was achieved, the scores were agreed upon for each response. The first set of 25 responses was discussed at length and consensus scored using this method. Committee members then went on to score the remaining responses independently and discuss responses when initial agreement was not achieved.

Discussions of student responses focused on the use of rubric language to ensure that the committee members remained focused on the specific requirements of each score point in the TDA rubric. For each student response, committee members' scores were recorded and DRC PAS staff made notes regarding the justifications and scoring decisions made by the committee members. Nebraska educators also made individual scoring notes for each response. Responses for which there was strong agreement among committee members were identified as potential anchor papers to be used in the Scoring Guides for training DRC readers. This information was used by the Scoring Directors during training. This same process was then repeated for each of the remaining four items for each grade.

At the end of each day of rangefinding, the DRC grade-level facilitators and NDE representatives met to conduct an inter-grade calibration. A sample of committee scored student responses from each adjacent grade were discussed and compared to ensure consistent interpretation of the rubric between grade-level committees, consistency of scoring between the grade levels, and the application of increasingly appropriate levels of rigor for each grade level. (NDE representatives from the grade 3 and 4 Pilot Test also participated in this process).

Training Material Creation

DRC's PAS staff assembled the committee scored rangefinding responses into sets to be used for training readers. Responses that the rangefinding committee selected as relevant in terms of the scoring concepts they illustrate were annotated and included as anchor papers in a scoring guide. These anchor papers, along with the holistic rubric served as the readers' constant reference throughout the project. Training and qualifying sets were then assembled using the remaining student responses that were reviewed and scored by rangefinding committee members. Responses were selected for training to show readers the spectrum for each score point and to highlight some of the TDA scoring characteristics.

Readers

The readers were chosen by the project managers from a pool, consisting of experienced individuals who are proven successful readers and leaders, and who have strong backgrounds in Nebraska assessment. All scoring personnel are required to sign confidentiality agreements before any training or handling of secure materials begins. DRC retains a pool of experienced readers from year to year and every reader who works on a Nebraska project must have a four-year degree from an accredited institution. DRC readers work at an hourly rate and are evaluated on accuracy; not by how many responses they read per hour/day. For the field test scoring, each grade had a team of ten readers and a Scoring Director.

Training

Representatives from NDE were on site at the Plymouth, Minnesota Scoring Center (May19– June 3, 2016) to collaborate with DRC Scoring Directors and staff during the entire scoring window. A representative from NDE worked cooperatively to review and discuss all of the training materials, and to oversee and make necessary scoring decisions throughout the project. NDE and DRC representatives identified one of the five items per grade to use as a qualifying item on which to begin training readers. A scoring guide, training set, and qualifying set were produced for readers to train and qualify. Training began with the Scoring Director providing an intensive review of the TDA holistic scoring rubric and the anchor papers in the scoring guide. Next, readers practiced by independently scoring the responses in the training set. The Scoring Director led a thorough discussion of the practice set papers in a small-group setting. Once the anchor set and training set papers were thoroughly discussed, each reader was required to demonstrate understanding of the scoring criteria by qualifying with acceptable agreement to the true scores. Once a reader was qualified at a grade level, subsequent items for that grade required each reader to thoroughly understand the scoring guide for each item and to adequately complete the corresponding practice set.

Scoring

Student responses are scored blindly and independently by multiple readers using DRC's handscoring system. Readers are not able to see demographic information pertaining to the student being scored, nor are they able to see any of the other scores given by any other reader. Each reader is required to apply the holistic scoring rubric to a given TDA response and is instructed to avoid any bias in their scoring decisions. All 25 items were scored, with a minimum of 1,500 responses per item completed.

Quality Control

Monitoring and Read-Behinds.

Scoring Directors conducted routine read-behinds for every member of their team and provided feedback and assistance to their readers.

Statistical Handscoring Reports

Numerous quality control reports were produced on demand or run daily in order to maintain high standards of scoring accuracy. The Inter-Rater Reliability Report and Score Point Distribution Report were especially helpful in analyzing scoring data and maintaining high standards of scoring quality.

Table 7.4.1 Inter-Rater Reliability and Score Point Distribution Results for Nebraska TDA Field Test 2016

			Inter-Rater Reliability			Score Point Distribution				
Grade	Item	Total Scored	%EX	%AD	% EX +AD	%1	%2	%3	%4	%NS
5	748963	2,077	94	6	100	73	16	1	0	9
5	749019	1,748	89	11	100	54	38	3	0	4
5	751234	1,825	95	5	100	76	15	2	0	7
5	751281	2,099	95	5	100	75	15	1	0	8
5	751404	2,251	97	3	100	68	13	1	0	16
6	748979	2,352	90	10	100	68	21	2	0	8
6	748994	2,651	95	5	100	81	11	1	0	7
6	749691	2,394	100	0	100	74	5	0	0	20
6	754876	2,405	97	3	100	73	10	1	0	16
6	754877	3,055	92	8	100	67	21	2	0	10

			Inter-Rater Reliability			Score Point Distribution				
Grade	Item	Total Scored	%EX	%AD	% EX +AD	%1	%2	%3	%4	%NS
7	751288	1,989	92	8	100	57	25	3	0	15
7	751328	2,097	90	9	99	51	32	7	1	9
7	751438	2,548	95	5	100	60	23	2	0	14
7	751533	2,054	95	5	100	69	16	2	0	13
7	751546	2,441	91	9	100	66	20	3	0	11
8	749026	1,863	94	6	100	55	22	4	1	18
8	749100	2,360	93	7	100	61	23	4	0	13
8	749109	2,100	86	14	100	49	39	8	1	4
8	749491	2,045	90	10	100	59	22	3	0	15
8	749601	2,364	93	7	100	54	31	5	1	9
11	751747	1,646	92	8	100	46	31	10	2	11
11	751815	1,809	90	10	100	44	33	10	3	10
11	752247	1,957	87	13	100	45	29	10	2	15
11	754843	2,215	94	6	100	47	31	9	2	11
11	754856	1,813	85	15	100	47	32	10	2	8

8. RELIABILITY

This chapter addresses the reliability of NeSA-Alt test scores. According to Mehrens and Lehmann (1975) reliability is defined as:

> … the degree of consistency between two measures of the same thing. (p. 88).

8.1 COEFFICIENT ALPHA

The ability to measure consistently is a necessary prerequisite for making appropriate interpretations (i.e., showing evidence of valid use of results). Conceptually, reliability can be referred to as the consistency of the results between two measures of the same thing. This consistency can be seen in the degree of agreement between two measures on two occasions. Operationally, such comparisons are the essence of the mathematically defined reliability indices.

All measures consist of an accurate, or true, component and an inaccurate, or error, component. Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical environment and changes in examinee disposition may increase error and decrease reliability. This is the fundamental premise of traditional reliability analysis and measurement theory. Stated explicitly, this relationship can be seen as the following:

$$\textit{Observed Score} = \textit{True Score} + \textit{Error} \tag{8.1}$$

To facilitate a mathematical definition of reliability, these components can be rearranged to form the following ratio:

$$\textit{Reliability} = \frac{\textbf{\textit{TrueScoreVariance}}}{\textbf{\textit{ObservedScoreVariance}}} = \frac{\textbf{\textit{TrueScoreVariance}}}{\textbf{\textit{TrueScoreVariance}} + \textbf{\textit{ErrorScoreVariance}}} \tag{8.2}$$

When there is no error, the reliability is true score variance divided by true score variance, which equals 1. However, as more error influences the measure, the error component in the denominator of the ratio increases. As a result, the reliability decreases.

The reliability index used for the 2016 administration of the NeSA was the Coefficient Alpha α (Cronbach, 1951). Acceptable α values generally range in the mid to high 0.80s to low 0.90s. The total test Coefficient Alpha reliabilities of the whole population are presented in Table 8.1.1 for each grade and content area of the NeSA. The table contains test length in total number of items (*L*), test reliabilities, and traditional standard errors of measurement (*SEM*). As can be seen in the table, all reading, mathematics, and science forms for grades 3-11 have Coefficient Alphas in the high 0.80s or low 0.90s. Overall, these α values provide evidence of good reliability.

Table 8.1.1 Reliabilities and Standard Errors of Measurement

	Grade	*L*	Reliability	*SEM*
Reading	3	45	0.90	2.8
	4	45	0.87	2.9
	5	48	0.88	2.9
	6	48	0.89	2.9
	7	48	0.88	2.9
	8	50	0.88	3.0
	11	50	0.90	2.9
Mathematics	3	50	0.92	3.0
	4	55	0.93	3.1
	5	55	0.93	3.1
	6	58	0.94	3.2
	7	58	0.94	3.2
	8	60	0.93	3.3
	11	60	0.95	3.2
Science	5	50	0.89	3.0
	8	60	0.91	3.3
	11	60	0.91	3.3

Reliability estimates for subgroups based on gender, ethnicity, special education status, limited English proficiency status, and food program eligibility status are also computed and reported in Appendix W. Results show fairly high reliability indices for all subpopulations in the high 0.80s to low 0. 90s across grades and content areas, which indicates that the NeSA is not only reliable for the population as a whole, but it is also reliable for subpopulations of interest under NCLB. Appendix X present α for the content strands. Given that α is a function of test length, the smaller item counts for the content standards result in lower values of α which is to be expected. Overall, these two sets of values provide evidence of good reliability.

8.2 STANDARD ERROR OF MEASUREMENT

The traditional *SEM* uses the information from the test along with an estimate of reliability to make statements about the degree to which error influences individual scores. The *SEM* is based on the premise that underlying traits, such as academic achievement, cannot be measured exactly without a perfectly precise measuring instrument. The standard error expresses unreliability in terms of the raw-score metric. The *SEM* formula is provided below:

$$SEM = SD\sqrt{1 - reliability}. \qquad (8.3)$$

This formula indicates that the value of the *SEM* depends on both the reliability coefficient and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the

SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the *SEM* would be 0.0. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). *SEM*s were calculated for each NeSA grade and content area using raw scores and displayed in Table 8.1.1.

8.3 CONDITIONAL STANDARD ERROR OF MEASUREMENT (CSEM)

The preceding discussion reviews the traditional approach to judging a test's consistency. This approach is useful for making overall comparisons between alternate forms. However, it is not very useful for judging the precision with which a specific student's score is known. The Rasch measurement models provide "conditional standard errors" that pertain to each unique ability estimate. Therefore, the *CSEM* may be especially useful in characterizing measurement precision in the neighborhood of a score level used for decision-making—such as cut scores for identifying students who meet a performance standard.

The complete set of conditional standard errors for every obtainable score can be found in Appendices Q, R and S as part of the raw-to-scale-score conversions for each grade and content area. Values were derived using the calibration data file described in Chapter Six and are on the scaled score metric. The magnitudes of *CSEM*s across the score scale seemed reasonable for most NeSA tests that the values are lower in the middle of the score range and increase at both extremes (i.e., at smaller and larger scale scores). This is because ability estimates from scores near the center of the test scoring range are known much more precisely than abilities associated with extremely high or extremely low scores. Table 8.3.1 reports the minimum *CSEM* of the scale score associated with the zero total test score (Min *CSEM*), the maximum *CSEM* of the scale score associated with the perfect total test score (Max *CSEM*), *CSEM* at the cuts of Below and Meets performance levels (*CSEM* B/M), and *CSEM* at the cuts of Meets and Exceeds performance levels (*CSEM* M/E) for each grade and content area. *CSEM* values at the cut score were generally associated with smaller *CSEM* values, indicating that more precise measurement occurs at these cuts.

Table 8.3.1 *CSEM* of the Scale Scores for 2016 NeSA Tests

	Grade	Min *CSEM*	Max *CSEM*	*CSEM* B/M	*CSEM* M/E
Reading	3	9	52	9	11
	4	11	67	11	13
	5	12	72	12	13
	6	12	69	12	13
	7	12	71	12	13
	8	11	68	11	13
	11	12	73	12	15
Mathematics	3	9	54	9	11
	4	8	51	8	10
	5	8	53	8	10
	6	8	55	8	11
	7	8	54	8	11
	8	8	54	8	11
	11	10	66	10	13
Science	5	10	59	10	13
	8	9	61	9	12
	11	7	49	7	11

8.4 DECISION CONSISTENCY AND ACCURACY

When criterion-referenced tests are used to place the examinees into two or more performance classifications, it is useful to have some indication of how accurate or consistent such classifications are. Decision consistency refers to the degree to which the achievement level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision accuracy describes the extent to which achievement-level classification decisions based on the administered test form would agree with the decisions that would be made on the basis of a perfectly reliable test. In a standards-based testing program there should be great interest in knowing how consistently and accurately students are classified into performance categories.

Since it is not feasible to repeat NeSA testing in order to estimate the proportion of students who would be reclassified in the same achievement levels, a statistical model needs to be imposed on the data to project the consistency or accuracy 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. These approaches are fairly complex, and the cited sources contain details regarding the statistical models used to calculate decision consistency from the single NeSA administration.

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 three 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 yield similar results to the Livingston and Lewis method (1995) and the method by Stearns and Smith (2007).

The results for the overall consistency across all three achievement levels are presented in Tables 8.4.1 – 8.4.3. The tabled values, derived using the program *BB-Class* (Brennan, 2004), show that consistency values across the two methods are generally very similar. Across all content areas, the overall decision consistency ranged from the mid 0.80s to the low 0.90s while the decision accuracy ranged from the high 0.80s to the mid 0.90s. If a parallel test were administered, at least 85% or more of students would be classified in the same way. Dichotomous decisions using the Meets cuts (Below/Meets) generally have the highest consistency values and exceeded 0.90 in all cases. The pattern of decision accuracy across different cuts is similar to that of decision consistency.

Table 8.4.1 NeSA-R Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Reading	3	0.94	0.91	0.92	0.87	0.94	0.91	0.92	0.88
	4	0.93	0.89	0.91	0.85	0.93	0.89	0.91	0.85
	5	0.94	0.90	0.92	0.85	0.94	0.90	0.92	0.86
	6	0.94	0.90	0.91	0.86	0.94	0.90	0.91	0.86
	7	0.94	0.90	0.92	0.86	0.94	0.90	0.92	0.86
	8	0.93	0.89	0.91	0.85	0.93	0.89	0.91	0.85
	11	0.93	0.90	0.90	0.86	0.93	0.90	0.90	0.87

Table 8.4.2 NeSA-M Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Math	3	0.94	0.92	0.91	0.89	0.94	0.92	0.91	0.90
	4	0.94	0.93	0.92	0.90	0.94	0.93	0.92	0.91
	5	0.94	0.94	0.91	0.91	0.94	0.94	0.91	0.91
	6	0.94	0.94	0.92	0.91	0.94	0.94	0.92	0.92
	7	0.94	0.94	0.91	0.92	0.94	0.94	0.91	0.92
	8	0.93	0.94	0.90	0.92	0.93	0.94	0.90	0.92
	11	0.94	0.95	0.92	0.92	0.94	0.95	0.92	0.92

Table 8.4.3 NeSA-S Decision Consistency Results

Content Area	Grade	Livingston & Lewis				Hanson & Brennan			
		Decision Accuracy		Decision Consistency		Decision Accuracy		Decision Consistency	
		Meets	Exceeds	Meets	Exceeds	Meets	Exceeds	Meets	Exceeds
Science	5	0.92	0.91	0.89	0.88	0.92	0.92	0.89	0.88
	8	0.92	0.93	0.89	0.90	0.92	0.93	0.89	0.90
	11	0.93	0.93	0.90	0.90	0.93	0.93	0.91	0.90

9. VALIDITY

As defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), “Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests” (p. 11). The validity process involves the collection of a variety of evidence to support the proposed test score interpretations and uses. This entire technical report describes the technical aspects of the NeSA tests in support of their score interpretations and uses. Each of the previous chapters contributes important evidence components that pertain to score validation: test development, test scoring, item analysis, Rasch calibration, scaling, and reliability. This chapter summarizes and synthesizes the evidence based on the framework presented in *The Standards*.

9.1 EVIDENCE BASED ON TEST CONTENT

Content validity addresses whether the test adequately samples the relevant material it purports to cover. The NeSA for grades 3 through 11 is a criterion-referenced assessment. The criteria referenced are the Nebraska reading and mathematics content standards. Each assessment was based on and was directly aligned to the Nebraska statewide content standards to ensure good content validity.

For criterion-referenced, standards-based assessment, the strong content 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 Appendices K, L, and M, as well as the success of the Bookmark and Contrasting Groups standard setting processes (in the separate *2010 NeSA-R Standard Setting Technical Report, 2011 NeSA-M Standard Setting Technical Report and 2012 NeSA-S 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. Contrasting Groups participants, using their knowledge and experience with their students, placed their students in a corresponding Performance Level.

9.2 EVIDENCE BASED ON INTERNAL STRUCTURE

As described in the *Standards* (2014), internal-structure evidence refers to the degree to which the relationships between test items and test components conform to the construct on which the proposed test interpretations are based.

Item-Test Correlations: Item-test correlations are reviewed in Chapter Four. All values are positive and of acceptable magnitude.

Rasch Measurement Dimensionality: Results from principle components analyses are presented in Chapter Five. The NeSA reading, mathematics, and science tests were essentially unidimensional, providing evidence supporting interpretations based on the total scores for the respective NeSA tests.

Strand Correlations: Correlations and disattenuated correlations between strand scores within each content area are presented below. This data can also provide information on score dimensionality that is part of internal-structure evidence. As noted in Chapter Two and also in Table 9.2.1, the NeSA-R tests have two strands (denoted by R.1 and R.2), the NeSA-M tests have four strands (denoted by M.1, M.2, M.3, and M.4), and the NeSA-S have four strands (denoted by S.1, S.2, S.3, and S.4) for each grade and content area.

For each grade, Pearson's correlation coefficients between these strands are reported in Tables 9.2.2.a through 9.2.2.g. The intercorrelations between the strands within the content areas are positive and generally range from moderate to high in value.

Table 9.2.1 NeSA Content Strands

Content	Code	Strand
Reading	R.1	Vocabulary
	R.2	Comprehension
Mathematics	M.1	Number Sense
	M.2	Geometric/Measurement
	M.3	Algebraic
	M.4	Data Analysis/Probability
Science	S.1	Inquiry, the Nature of Science, and Technology
	S.2	Physical Science
	S.3	Life Science
	S.4	Earth and Space Science

Table 9.2.2.a Correlations between Reading and Mathematics Strands for Grade 3

Grade 3	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.76	–				
M.1	0.65	0.66	–			
M.2	0.60	0.61	0.69	–		
M.3	0.59	0.63	0.72	0.62	–	
M.4	0.60	0.62	0.67	0.57	0.62	–

Table 9.2.2.b Correlations between Reading and Mathematics Strands for Grade 4

Grade 4	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.73	–				
M.1	0.61	0.66	–			
M.2	0.60	0.64	0.78	–		
M.3	0.53	0.57	0.73	0.66	–	
M.4	0.48	0.53	0.58	0.54	0.50	–

Table 9.2.2.c Correlations between Reading, Mathematics, and Science Strands for Grade 5

Grade 5	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	0.74	–								
M.1	0.63	0.69	–							
M.2	0.55	0.59	0.70	–						
M.3	0.57	0.61	0.74	0.61	–					
M.4	0.62	0.66	0.73	0.59	0.65	–				
S.1	0.64	0.71	0.64	0.55	0.57	0.63	–			
S.2	0.60	0.66	0.59	0.54	0.53	0.58	0.65	–		
S.3	0.59	0.65	0.55	0.48	0.49	0.55	0.62	0.63	–	
S.4	0.59	0.63	0.58	0.52	0.52	0.58	0.64	0.66	0.64	–

Table 9.2.2.d Correlations between Reading and Mathematics Strands for Grade 6

Grade 6	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.75	–				
M.1	0.60	0.65	–			
M.2	0.61	0.65	0.75	–		
M.3	0.64	0.70	0.78	0.76	–	
M.4	0.60	0.66	0.73	0.73	0.76	–

Table 9.2.2.e Correlations between Reading and Mathematics Strands for Grade 7

Grade 7	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.76	–				
M.1	0.61	0.68	–			
M.2	0.54	0.60	0.71	–		
M.3	0.62	0.71	0.82	0.70	–	
M.4	0.58	0.66	0.73	0.64	0.73	–

Table 9.2.2.f Correlations between Reading, Mathematics, and Science Strands for Grade 8

Grade 8	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	0.77	–								
M.1	0.59	0.67	–							
M.2	0.58	0.65	0.75	–						
M.3	0.61	0.69	0.79	0.74	–					
M.4	0.60	0.67	0.73	0.71	0.72	–				
S.1	0.64	0.71	0.63	0.62	0.64	0.62	–			
S.2	0.61	0.67	0.62	0.62	0.62	0.61	0.65	–		
S.3	0.65	0.71	0.62	0.62	0.62	0.62	0.68	0.71	–	
S.4	0.60	0.66	0.61	0.61	0.60	0.59	0.64	0.70	0.71	–

Table 9.2.2.g Correlations between Reading, Mathematics, and Science Strands for Grade 11

Grade 11	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	0.74	–								
M.1	0.48	0.56	–							
M.2	0.59	0.68	0.67	–						
M.3	0.61	0.69	0.71	0.82	–					
M.4	0.61	0.69	0.61	0.75	0.76	–				
S.1	0.60	0.69	0.54	0.67	0.67	0.65	–			
S.2	0.66	0.72	0.55	0.69	0.69	0.68	0.70	–		
S.3	0.67	0.72	0.54	0.67	0.67	0.67	0.69	0.75	–	
S.4	0.61	0.66	0.48	0.62	0.62	0.61	0.65	0.73	0.71	–

The correlations in Tables 9.2.2.a through 9.2.2.g are based on the observed strand scores. These observed-score correlations are weakened by existing measurement error contained within each strand. As a result, disattenuating the observed correlations can provide an estimate of the relationships between strands if there is no measurement error. The disattenuated correlation coefficients can be computed from the observed correlations (reported in Tables 9.2.2.a – 9.2.2.g) and the reliabilities for each strand (Spearman, 1904, 1910). Disattenuated correlations very near 1.00 might suggest that the same or very similar constructs are being measured. Values somewhat less than 1.00 might suggest that different strands are measuring slightly different aspects of the same construct. Values markedly less than 1.00 might suggest the strands reflect different constructs.

Tables 9.2.3.a through 9.2.3.g show the corresponding disattenuated correlations for the 2016 NeSA tests for each grade. Given that none of these strands has perfect reliabilities (see Chapter Eight), the disattenuated strand correlations are higher than their observed score counterparts. Some within-content-area correlations are very high (e.g., above 0.95), suggesting that the within-content-area strands might be measuring essentially the same construct. This, in turn, suggests that some strand scores might not provide unique information about the strengths or weaknesses of students.

On a fairly consistent basis, the correlations between the strands within each content area were higher than the correlations between strands across different content areas. In general, within-content-area strand correlations were mostly greater than 0.90, while across-content-area strand correlations generally ranged from 0.75 to 0.92. Such a pattern is expected since the two content area tests were designed to measure different constructs.

Table 9.2.3.a Disattenuated Strand Correlations for Reading and Mathematics: Grade 3

Grade 3	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.96	–				
M.1	0.83	0.78	–			
M.2	0.81	0.76	0.87	–		
M.3	0.82	0.80	0.93	0.85	–	
M.4	0.89	0.86	0.94	0.84	0.94	–

Table 9.2.3.b Disattenuated Strand Correlations for Reading and Mathematics: Grade 4

Grade 4	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	0.99	–				
M.1	0.81	0.78	–			
M.2	0.83	0.78	0.94	–		
M.3	0.80	0.76	0.96	0.91	–	
M.4	0.91	0.89	0.95	0.92	0.93	–

Table 9.2.3.c Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 5

Grade 5	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	1.00	–								
M.1	0.84	0.79	–							
M.2	0.84	0.78	0.91	–						
M.3	0.86	0.79	0.95	0.89	–					
M.4	0.93	0.86	0.93	0.87	0.94	–				
S.1	0.96	0.93	0.82	0.82	0.83	0.90	–			
S.2	0.93	0.87	0.78	0.81	0.78	0.85	0.95	–		
S.3	0.92	0.87	0.73	0.74	0.74	0.82	0.93	0.96	–	
S.4	0.91	0.84	0.76	0.78	0.77	0.85	0.93	0.99	0.97	–

Table 9.2.3.d Disattenuated Strand Correlations for Reading and Mathematics: Grade 6

Grade 6	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	1.00	–				
M.1	0.82	0.79	–			
M.2	0.84	0.79	0.95	–		
M.3	0.86	0.83	0.95	0.94	–	
M.4	0.84	0.81	0.94	0.93	0.95	–

Table 9.2.3.e Disattenuated Strand Correlations for Reading and Mathematics: Grade 7

Grade 7	R.1	R.2	M.1	M.2	M.3	M.4
R.1	–					
R.2	1.00	–				
M.1	0.81	0.81	–			
M.2	0.79	0.79	0.93	–		
M.3	0.83	0.83	0.97	0.92	–	
M.4	0.84	0.86	0.94	0.92	0.95	–

Table 9.2.3.f Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 8

Grade 8	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	1.00	–								
M.1	0.80	0.81	–							
M.2	0.80	0.80	0.94	–						
M.3	0.83	0.84	0.97	0.93	–					
M.4	0.87	0.87	0.97	0.96	0.96	–				
S.1	0.94	0.93	0.83	0.83	0.85	0.89	–			
S.2	0.86	0.85	0.80	0.81	0.80	0.85	0.91	–		
S.3	0.92	0.90	0.79	0.81	0.80	0.86	0.95	0.95	–	
S.4	0.87	0.84	0.79	0.81	0.79	0.84	0.91	0.97	0.97	–

Table 9.2.3.g Disattenuated Strand Correlations for Reading, Mathematics and Science: Grade 11

Grade 11	R.1	R.2	M.1	M.2	M.3	M.4	S.1	S.2	S.3	S.4
R.1	–									
R.2	0.97	–								
M.1	0.75	0.77	–							
M.2	0.79	0.80	0.94	–						
M.3	0.78	0.78	0.98	0.95	–					
M.4	0.85	0.84	0.90	0.93	0.92	–				
S.1	0.90	0.91	0.85	0.89	0.87	0.90	–			
S.2	0.91	0.87	0.81	0.85	0.84	0.88	0.97	–		
S.3	0.93	0.88	0.80	0.83	0.82	0.87	0.96	0.98	–	
S.4	0.90	0.85	0.74	0.81	0.79	0.84	0.95	0.99	0.98	–

9.3 EVIDENCE RELATED TO THE USE OF THE RASCH MODEL

Since the Rasch model is the basis of all calibration, scaling, and linking analyses associated with the NeSA, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met as well as the fit between the model and test data. As discussed at length in Chapter Five, the underlying assumptions of Rasch models were essentially met for all the NeSA data, indicating the appropriateness of using the Rasch models to analyze the NeSA data.

In addition, the Rasch model was also used to link different operational NeSA tests across years. The accuracy of the linking also affects the accuracy of student scores and the validity of score uses. DRC Psychometric Services staff conducted verifications to check the accuracy of the procedures, including item calibration, conversions from the raw score to the Rasch ability estimate, and conversions from the Rasch ability estimates to the scale scores.

10. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69-81.

Andrich, D. (1988). *Rasch models for measurement*. Newberry Park, CA: Sage.

Brennan, R. L. (2004). BB-Class (Version 1.0). [Computer software] Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement & Assessment. Retrieved from www.education.uiowa.edu/casma.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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

Fischer, G., & Molenaar, I. (1995). *Rasch models : Foundations, recent developments, and applications.* New York, NY: Springer.

Hambleton, R. K., & Novick, M. R. (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,* 345-359.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices*, *10*(2), 33-41.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Huynh, H. (2000). Guidelines for Rasch linking for PACT. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H., & Rawls, A. (2009). A comparison between Robust *z* and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In E. V. Smith, Jr., & G. E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing*. (pp. 429-442). Maple Grove, MN: JAM Press.

Huynh, H., & Meyer, P. (2010). Use of Robust z in detecting unstable items in item response theory models. *Practical Research, Assessment, and Evaluation*, 15, (2).

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). 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. M. (2009). *A user's guide to WINSTEPS MININSTEP Rasch-model computer programs*. Chicago, IL: Winsteps.com

Linacre, J. M. (2016). Winsteps® Rasch measurement computer program (V3.90). Beaverton, OR: Winsteps.com.

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.

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

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research*, *14*, 21-38.

Mead, R. J. (1976). *Assessing the fit of data to the Rasch model through the analysis of residuals*. Unpublished doctoral dissertation. Chicago, IL: 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.

Mehrens, W. A., & Lehmann, I. J. (1975) *Standardized tests in education* (2nd ed.). New York, NY: Holt, Rinehart, and Winston.

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.

Spearman C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72-101.

Spearman C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271-295.

Smith, E. V., Jr., & Smith, R. M. (Eds.). (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199-218.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Stearns, M., & Smith R. M. (2008). Estimation of classification consistency indices for complex assessments: Model based approaches. *Journal of Applied Measurement*, *9*, 305-315.

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, DC: Council of Chief State School Officers.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problem* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith, Jr., & R. M. Smith (Eds.) *Introduction to Rasch measurement* (pp. 25-47). Maple Grove, MN: JAM Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure of sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

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

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