North Dakota State Assessments

Fall 2014 Administration Final Technical Report

Submitted to the
North Dakota Department of Public Instruction

September 9, 2015

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

Developed and published under contract with North Dakota Department of Public Instruction by CTB/McGraw-Hill LLC, 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2014 by North Dakota Department of Public Instruction. Based on a template copyright 1996 © CTB/McGraw-Hill LLC. All rights reserved. Only North Dakota State educators and citizens may copy, download and/or print the document, located online at: http://www.dpi.state.nd.us/testing/asses/index.shtm. Any other use or reproduction or distribution of this document, in whole or in part, requires the prior written permission of North Dakota Department of Public Instruction and the publisher.

Table of Contents

TABLE OF CONTENTS ..... **I**

TABLE OF TABLES ..... **III**

TABLE OF FIGURES ..... **IV**

PART 1: OVERVIEW **5**

USES OF TEST SCORES 5
CRITERION-REFERENCED SCORES REPORTED 6
Test-Level Scores *6*
Content Standard and Benchmark Scores *7*
The Use of Content Standard and Benchmark Scores *7*

PART 2: TEST CONSTRUCTION **8**

ESTABLISHING THE CONTENT RATIONALE 8
Designing Assessment Specifications *8*
ITEM SELECTION 8
Development Procedures for the CTB Item Pool *8*
Development Procedures for the North Dakota Item Pool *9*
Selection Procedures for the North Dakota State Assessment *9*

PART 3: ADMINISTRATION AND SCORING **15**

NORTH DAKOTA SCORING PROCESS 15
Selection of Scoring Evaluators *15*
Handscoring Training Process *16*
Monitoring the Scoring Process *18*
Security *19*
INTER-RATER RELIABILITY RESULTS 19
SUMMARY 20

PART 4: TEST VALIDITY **21**

MINIMIZATION OF CONSTRUCT-IRRELEVANT VARIANCE AND UNDER-REPRESENTATION 21
CONVERGENT VALIDITY 22
Principal Components Analysis *22*

PART 5: TEST RELIABILITY **24**

TEST RELIABILITY INDICES 24
STANDARD ERROR OF MEASUREMENT 25
DECISION ACCURACY AND CONSISTENCY 26

PART 6: TEST AND ITEM STATISTICS **30**

TEST-LEVEL STATISTICS 30
ITEM-LEVEL STATISTICS 30

PART 7: SCALING AND LINKING **35**

SCALING 35
ITEM RESPONSE THEORY 35
MODEL FIT 36
LINKING METHOD AND RESULTS 37
Anchor Evaluation *38*
Vertical Scale *39*

Copyright © 2014 by North Dakota Department of Public Instruction

Scoring Procedures 40
Lowest and Highest Obtainable Scale Scores 41

PART 8: FAIRNESS 45

Minimizing Bias through Careful Test Development 45
Evaluating Bias through Differential Item Functioning Statistics 45

PART 9: SUMMARY OF OPERATIONAL TEST RESULTS 50

Scale Score Distribution Summary 50
Performance Level Distribution Summary 50

REFERENCES 55

APPENDIX A: NORTH DAKOTA 2014 SCORING TABLES 58

Copyright © 2014 by North Dakota Department of Public Instruction

Table of Tables

TABLE 2.1: TOTAL NUMBER OF ITEMS AND TOTAL NUMBER OF SCORE POINTS MEASURING EACH NORTH DAKOTA SCIENCE CONTENT STANDARD 11

TABLE 3.1: INTER-RATER RELIABILITY, SCIENCE 19

TABLE 4.1: PRINCIPAL COMPONENTS ANALYSIS FOR SCIENCE 23

TABLE 5.1: RELIABILITY COEFFICIENTS AND SEMS FOR THE TOTAL POPULATION, SCIENCE 27
TABLE 5.2: TEST RELIABILITY COEFFICIENTS AND SEMS BY SUBGROUP, SCIENCE 28
TABLE 5.3: DECISION ACCURACY AND CONSISTENCY AT ACHIEVEMENT CUT POINTS 29

TABLE 6.1: NDSA MEANS, STANDARD DEVIATIONS FOR *P*-VALUES, ITEM-TOTAL CORRELATION 31
TABLE 6.2: ITEM STATISTICS: GRADE 4 32
TABLE 6.3: ITEM STATISTICS: GRADE 8 33
TABLE 6.4: ITEM STATISTICS: GRADE 11 34

TABLE 7.1: ITEM FIT STATISTICS FOR MISFITTING ITEMS 42
TABLE 7.2: LOSS AND HOSS VALUES BY GRADE 42
TABLE 7.3: NUMBER AND PERCENT OF STUDENTS AT THE HOSS AND PENULTIMATE HOSS, BY GRADE FOR SCIENCE 42

TABLE 8.1: SUMMARY OF THE NUMBER OF ITEMS FLAGGED FOR DIF IN SCIENCE NDSA 48
TABLE 8.2: 2013 NDSA LINN-HARNISCH STATISTICS: FLAGGED ITEMS, SCIENCE 49

TABLE 9.1: MEAN AND STANDARD DEVIATION OF SCALE SCORE FOR ALL STUDENTS, STUDENTS DISAGGREGATED BY SUBGROUP, SCIENCE 51
TABLE 9.2: MEAN AND STANDARD DEVIATION OF SCALE SCORES FOR 2006–2014 BY GRADE—CENSUS (DOES NOT INCLUDE HOME-EDUCATION STUDENTS) 52
TABLE 9.3: PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL FOR ALL STUDENTS, STUDENTS DISAGGREGATED BY SUBGROUP, SCIENCE 53
TABLE 9.4: PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL FOR 2006–2014 BY GRADE: SCIENCE—STATE . 54

Copyright © 2014 by North Dakota Department of Public Instruction

Table of Figures

FIGURE 5.1: STANDARD ERROR OF MEASUREMENT CURVES FOR SCIENCE 29

FIGURE 7.1: SCIENCE, GRADE 8, ITEM #34, CR ITEM (3 LEVELS) 43
FIGURE 7.2: COMPARISON OF TEST CHARACTERISTICS CURVES (TCCS) BETWEEN 2013 AND 2014 SCIENCE NDSAS 44

Copyright © 2014 by North Dakota Department of Public Instruction

Part 1: Overview

Historically, North Dakota's testing program comprised norm-referenced assessments. In 2001, the North Dakota Department of Public Instruction contracted with CTB/McGraw-Hill to expand the testing program to include both norm-referenced and criterion-referenced tests for grades 4, 8, and 12. Then, in the spring of 2002, North Dakota administered assessments in Reading, Language Arts, and Mathematics to students in grades 4, 8, and 12. In 2002, the high school test was administered in the fall instead of the spring.

In the fall of 2004, North Dakota administered North Dakota State Assessments (NDSA) assessments in Reading, Language Arts, and Mathematics to students in grades 3 through 8 and grade 11. Grade-level tests were administered so that the NDSA would be in compliance with the No Child Left Behind Act of 2001 (NCLB). Beginning in 2004, NDSA was a custom assessment comprising items from CTB's item pool and, unlike previous years, was based solely on criterion-referenced scores. Students received test scores on the Reading and Mathematics NDSA as well as percent-correct scores for each content standard and benchmark. Students only received number-correct scores at the content-standard and benchmark levels on the Language Arts NDSA until 2011, when students began receiving Language Arts test-level scores as well.

In 2006, the Science NDSA was administered for the first time. Like Reading, Language Arts, and Mathematics in past administrations, test-level and content standard-level scores were reported for the Science NDSA.

This report provides a technical overview of the Science assessment of the 2014 North Dakota State Assessment. As such, it presents evidence for the validity of the 2014 NDSA scores. This first section, Part 1, provides a brief historical overview of the NDSA assessments and also discusses the uses of NDSA test scores. Part 2 focuses on the test development process used to create the NDSA. The discussions in this section are important to the content-related validity of the NDSA scores. Part 3 discusses the scoring of constructed-response items, as well as the results of the inter-rater reliability studies. Part 4 discusses the concept of validity, including construct-related validity. Validity hinges on the uses of test scores. In this part, the assumption is tested that the content-area NDSAs measure only one construct. For example, the grade-level Science NDSA should measure one primary dimension (Science). Part 5 discusses test reliability, including decision accuracy and consistency. Part 6 reviews the test- and item-level characteristics. Part 7 presents the scaling and linking procedures, as well as the results of these procedures. Part 8 contains an overview of the statistical and developmental processes used to ensure fairness of the NDSA for all examinees. Part 9 is an overview of student demographics as well as past and current student achievement. Some analyses in this document are based on the calibration sample while others are based on census data. The type of data used for particular analyses is indicated throughout this technical report.

Uses of Test Scores

The purpose of the NDSA Science test scores is to measure student achievement on the Science content and achievement standards in North Dakota. The tests are designed so that comparisons

Copyright © 2014 by North Dakota Department of Public Instruction

of student achievement can be made by student, by classrooms within a school and school district, by schools within the state, and by school districts within the state. Test results are used by the state for accountability purposes under the No Child Left Behind Act of 2001.

The test scores may be used to classify students, schools, school districts, and the state with respect to the level of achievement each shows in each content area. Classification is based on the level of achievement demonstrated on the NDSA for each content area. The primary purpose of this document is to present prior evidence bearing on the interpretation of the NDSA Science scores as individual achievement scores in the content area of Science. As such, this document will present only limited evidence for the uses of the individual achievement NDSA scores. It is recommended that policy and program evaluation studies be conducted in accordance with the *Standards for Educational and Psychological Testing* (AERA, 2014) to support some of the uses of the NDSA scores.

Criterion-Referenced Scores Reported

Performance in Science is reported in multiple ways to enhance understanding of student achievement. For the test as a whole, an overall scale score with an associated level of achievement based on student performance is reported for Science. At the content standard and benchmark levels, percent-correct scores for each benchmark are reported. These scores, in varying ways, indicate a student's ability in Science.

In part, the rationale for the claim that the NDSA scores are credible individual achievement scores is that the NDSAs were devised by using items similar to the questions and activities teachers use to teach their students. These items, selected from CTB's item pool, were aligned to the North Dakota Content Standards by CTB content experts. North Dakota teachers then reviewed each item for alignment to the North Dakota Content Standards from which the NDSAs were developed. Detailed descriptions of the process used to delineate the knowledge, skills, and abilities, including content limits and descriptions for each content area, are beyond the scope of this document.

Test-Level Scores

At the test level, two types of scores are reported to indicate student achievement on the NDSA: (1) a scale score and (2) its associated level of achievement.

1. Scale Score

A scale score indicating a student's total performance on the Science NDSA is provided. This overall scale score is a quantification of the achievement being measured by the Science test. The validation of the test scores is focused on gathering contextual evidence that supports the test's construct: achievement of state content standards. Psychometric validation of this construct consists of *prima facie* evidence. Generally, alignment of items with the North Dakota content standards provides support for this *prima facie* evidence.

2. Level of Achievement

A student's performance on the NDSA is reported in one of four levels of achievement: *Novice*, *Partially Proficient*, *Proficient*, or *Advanced*. These levels are based on the student's

Copyright © 2014 by North Dakota Department of Public Instruction

scale score. Using the North Dakota content and performance standards to guide them, North Dakota teachers recommended the cut scores for the levels of achievement in Science at the Bookmark Standard Setting Workshop in March 2007 (refer to *North Dakota State Assessment Bookmark Standard Setting Technical Report 2007 for Grades 4, 8, and 11 Science*). These recommended cut scores reflect teachers' expectations of what North Dakota students should know and be able to do in each grade and content area. Thus, there is reason to believe the claim that the NDSAs are indeed reflecting the abilities intended by the North Dakota legislature and the Department of Public Instruction (DPI).

Content Standard and Benchmark Scores

In addition to scale scores and achievement levels, content standards and benchmark scores are reported at five reporting levels: state, school district, school, class, and student. The content standard score measures the aggregate information within the content standard. The content standards comprise benchmarks. A benchmark measures specific information within a content standard. A percent-correct score is reported for each content standard and benchmark.

The Use of Content Standard and Benchmark Scores

The purpose of reporting content standard scores on the NDSA is to present an indication of a student's achievement on each content standard. The Content Standard Performance Report provides an indication of the extent to which the student's overall achievement has the desired breadth of the North Dakota Content Standards. Teachers may use the percent-correct scores on each content standard as indicators of strengths and weaknesses, but other evidence, such as homework, class participation, diagnostic test scores, or observation, should be used as corroboration.

Likewise, the purpose of reporting benchmark subscores on the NDSA is to show how students perform on each benchmark. The diagnostic value of these subscores is limited for two reasons: (a) there are a limited number of items measuring each benchmark and (b) items measuring each benchmark may vary in difficulty across benchmarks and across years. Therefore, although the metric is consistent for interpretation purposes (percent correct), some benchmarks may have easier items than others.

Copyright © 2014 by North Dakota Department of Public Instruction

Part 2: Test Construction

Establishing the Content Rationale

The NDSA for Science is based on the North Dakota Content Standards and Benchmarks for Science. These content standards were developed in North Dakota by committees of North Dakota educators at the elementary, secondary, and higher education levels. Consultants from the Mid-Continent Regional Educational Laboratory (McREL) supported the process.

Designing Assessment Specifications

The 2014 NDSA Science reused test forms from the 2013 NDSA. The 2013 NDSA comprises items from the 2004 through 2012 NDSAs as well as CTB's item pool. A portion of the 2004 through 2012 NDSA items compose the anchor set. This set of items is used to place the 2013 NDSA on the NDSA scale. All of the items on the NDSA are aligned to the North Dakota Content Standards and Benchmarks for Science.

Item Selection

In this section, the development procedures used in creating CTB's item pool are discussed first, followed by the selection procedures used in selecting items from CTB's item pool.

Development Procedures for the CTB Item Pool

The CTB item pool comprises secure items from alternate forms of *TerraNova* (CTB/McGraw-Hill, 1996, 2000, 2006), the CoreLink Item Bank, and various state assessments. Each of these items was developed by professional item writers, has gone through CTB's development process, and has been field-tested. Item content was documented using curriculum guides and major basals or textbooks so that the items were sound and appropriate for the grade levels. Field-testing of items incorporated teachers' comments about the accuracy, validity, and grade-level appropriateness of the items. All assessment materials were carefully reviewed for content and editorial accuracy. Artists and designers worked with the item writers during development to ensure graphic and textual consistency.

The items in this pool have been screened and passed rigorous criteria for bias and statistical fit. The bias reduction studies incorporated four aspects:

- a focus on validity during content development
- the use of specific guidelines to avoid bias and cultural or gender stereotyping in test items, with analysis to ensure that each item met guidelines [*Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993) and *Guidelines for Bias-Free Publishing* (McGraw-Hill, 1983)]
- a review of test materials by a wide range of education professionals representing various ethnic and special student groups to identify possible bias in language, subject matter, or representation of people
- the application of differential item functioning (DIF) analysis

Copyright © 2014 by North Dakota Department of Public Instruction

Statistical fit was reviewed using item response theory (IRT). A three-parameter logistic model was used to review each multiple-choice item, and a two-parameter logistic model was used for each constructed-response item. The effect of each item within the set of items was studied to confirm that the set of items was cohesive and functioned well together for the intended assessment purposes.

Development Procedures for the North Dakota Item Pool

The North Dakota item pool comprised items written for and aligned to the North Dakota benchmarks. Each year, CTB developed a number of new items for field-testing at each grade. CTB assessment editors ran a gap analysis of the benchmarks and item types in the North Dakota pool. They selected benchmarks for development and began the item-writing process.

All North Dakota items were developed by professional item writers, went through CTB's development process, and were field-tested. Prior to field testing, the newly developed items went through a content and bias review by North Dakota educators and DPI staff. The new items were reviewed in terms of content alignment and item-level bias. CTB's assessment editors adjusted the new development to reflect the work that occurred during the content and bias review.

Selection Procedures for the North Dakota State Assessment

The selection procedure for the NDSA occurred in three phases. In the first phase, CTB assessment editors, working with CTB's research team, initially selected items for the 2013 operational form. In the second phase, CTB's assessment editors selected a group of newly developed items that passed through content and bias review for field-testing on the 2013 form. In the third phase, CTB's assessment editors provided the final selection to DPI for approval.

Initial Selection

CTB development experts performed the initial match of items to the North Dakota content standards and benchmarks. The selection of items was conducted using CTB's Automated Test Assembly (ATA) software. The use of ATA software allowed CTB assessment editors to construct alternate forms of a test, balancing psychometric and content considerations.

The 2013 NDSA may be considered an alternate form to the 2004 through 2012 NDSAs because the 2013 NDSA was placed on the same test scale as the earlier forms. To do this, item anchor sets were selected from the item sets administered in 2004 through 2012. The anchors were used to equate the 2014 NDSA to the baseline 2006 NDSA for Science. The results of the equating may be found in Part 7 of this technical report. Using ATA software, these anchor sets had to meet the following criteria:

- The anchor set had to be at least 25% of the overall test length.
- No changes could be made to the anchor items between years.
- The test characteristic curve (TCC) for the 2013 anchor set could vary by no more than 5% from the baseline-year anchor set's TCC.

Copyright © 2014 by North Dakota Department of Public Instruction

- The percentage of items measuring each content standard had to be comparable to the percentage of items measuring each content standard in the full test of the baseline-year NDSA. The coverage could only vary by 10% for any one content strand.
- The anchor items had to be placed throughout the entire test, not grouped in one section of the test. The anchor items were placed in close proximity to their placement on the 2004 through 2012 tests.
- To the degree possible, the grade-level TCCs should progress upward because the NDSA is on a vertical scale.

The selection for the overall test was completed in a manner similar to the selection of the anchor set. CTB's ATA software was again used. The overall 2013 form was pre-equated to the baseline-year form using CTB item statistics. Note that these items were recalibrated using data from North Dakota students. The pre-equating based on CTB item parameters was conducted for the benefit of the post-equating procedure. Using ATA software, the overall test had to meet the following criteria:

- The operational tests had to comprise approximately 55 items (including anchor items).
- The test characteristic curve (TCC) for the 2013 NDSA could vary by no more than 5% from the baseline-year anchor set's TCC.
- The percentage of items measuring each content standard could vary by no more than 10% between the 2013 NDSA and the baseline-year NDSA.

Content and Bias Review

Following the item selection process, content editors from CTB reviewed all items on the NDSA to confirm that the North Dakota standards and benchmarks were comprehensively covered (i.e., Science standards and benchmarks were assessed to the degree possible). The ND DPI also reviewed the selections in order to approve the match and selection of items made by CTB development experts.

Prior to 2013, all items were reviewed in content and bias review meetings composed of representatives from the ND DPI, teachers from North Dakota, and content editors from CTB. During these reviews, all items on the NDSA were reviewed to confirm that the North Dakota standards and benchmarks were comprehensively covered (i.e., Science standards and benchmarks were assessed to the degree possible). Further, these groups reviewed each potential item for cognitive alignment with Norman Webb's Depth of Knowledge levels and made recommendations about which items to accept in the test core and augmented supplements of the tests. The groups also examined items for possible sources of bias, including those related to gender and ethnicity/race.

Final Selections

Following the review by DPI, CTB assessment editors made necessary modifications both to items and to the test selections. These modifications were approved by the CTB research staff. CTB sent second and final pages of the NDSA test books to ND DPI for approval. Both sets of pages were approved by the ND DPI staff.

Copyright © 2014 by North Dakota Department of Public Instruction

Number of Items and Points Measuring Content Standards and Benchmarks

The number of items and total score points measuring each content standard and benchmark for the NDSA are summarized in Table 2.1 for Science.

Table 2.1: Total Number of Items and Total Number of Score Points Measuring Each North Dakota Science Content Standard

Grade	Code	Standard / Benchmark	# of points	# of items
	A	**SCIENCE CONCEPTS AND INQUIRY**	**17**	**15**
	1	**Standard 1: Unifying Concepts**	**3**	**3**
	3.01	3.1 Identify repetitive changes	3	3
	2	**Standard 2: Science Inquiry**	**14**	**12**
	3.01	3.1 Select science tools for investigations	4	3
	3.02	3.2 Ask scientific investigation questions	5	5
	3.03	3.3 Record observations of investigations	5	4
	B	**PHYSICAL SCIENCE**	**9**	**9**
	3	**Standard 3: Physical Science**	**9**	**9**
	3.01	3.1 ID physical properties: solids, liquid	4	4
	3.02	3.2 Identify a force as push or pull	1	1
	3.03	3.3 Describe magnetic attraction/repulsion	1	1
	3.04	3.4 Explain how vibration produces sound	1	1
	3.05	3.5 Describe direction and motion of light	1	1
		Items not reporting to Benchmarks	1	1
	C	**LIFE SCIENCE**	**12**	**12**
	4	**Standard 4: Life Science**	**12**	**12**
	3.01	3.1 Identify functions of an organism's parts	3	3
4	3.02	3.2 Describe life cycles: plants and animals	6	6
	3.03	3.3 Identify the needs of living things	3	3
	D	**EARTH AND SPACE SCIENCE**	**8**	**8**
	5	**Standard 5: Earth and Space Science**	**8**	**8**
	3.01	3.1 Identify measureable weather conditions	1	1
	3.02	3.2 Relate properties to uses of materials	1	1
	3.03	3.3 Identify ways in which rocks break down	2	2
	3.04	3.4 Identify properties of soil	1	1
	3.05	3.5 Explain how stars are like the sun	2	2
		Items not reporting to Benchmarks	1	1
	E	**SCIENCE AND CULTURE**	**11**	**11**
	6	**Standard 6: Science and Technology**	**3**	**3**
	3.01	3.1 Identify ways technology solves problems	3	3
	7	**Standard 7: Science and Other Areas**	**7**	**7**
	3.01	3.1 Identify ways to prevent spread of germs	2	2
	3.02	3.2 Identify benefits of recycling, reducing	4	4
		Items not reporting to Benchmarks	1	1
	8	**Standard 8: History and Nature of Science**	**1**	**1**
	3.01	3.1 Identify uses of science in careers, life	1	1

Copyright © 2014 by North Dakota Department of Public Instruction

Table 2.1: Total Number of Items and Total Number of Score Points Measuring Each North Dakota Science Content Standard (cont'd)

Grade	Code	Standard / Benchmark	# of points	# of items
	A	**SCIENCE CONCEPTS AND INQUIRY**	**18**	**18**
	1	**Standard 1: Unifying Concepts**	**7**	**7**
	7.01	7.1 Use models to explain science principles	1	1
	7.02	7.2 Identify components of body & ecosystems	2	2
	7.03	7.3 Identify examples of feedback mechanisms	1	1
	7.04	7.4 Relate form to function	1	1
		Items not reporting to Benchmarks	2	2
	2	**Standard 2: Science Inquiry**	**11**	**11**
	7.01	7.1 Communicate science investigation results	3	3
		Items not reporting to Benchmarks	8	8
	B	**PHYSICAL SCIENCE**	**10**	**10**
	3	**Standard 3: Physical Science**	**10**	**10**
	7.01	7.1 Explain transfer of energy across forms	6	6
		Items not reporting to Benchmarks	4	4
	C	**LIFE SCIENCE**	**11**	**11**
	4	**Standard 4: Life Science**	**11**	**11**
	7.01	7.1 Explain functions of the cell	1	1
	7.02	7.2 ID levels of organization: living systems	1	1
	7.03	7.3 Identify characteristics of reproduction	1	1
8	7.04	7.4 Identify organism/environment interaction	4	4
	7.05	7.5 Classify organisms into taxonomic groups	3	3
	7.06	7.6 Relate adaptations to survival	1	1
	D	**EARTH AND SPACE SCIENCE**	**8**	**7**
	5	**Standard 5: Earth and Space Science**	**8**	**7**
	7.01	7.1 Identify factors that affect climate	5	4
	7.02	7.2 Explain how seasons affect organisms	1	1
	7.03	7.3 Identify renewable/nonrenewable resources	2	2
	E	**SCIENCE AND CULTURE**	**10**	**9**
	6	**Standard 6: Science and Technology**	**3**	**3**
	7.01	7.1 Identify influences of technology	1	1
	7.02	7.2 Identify society/technology interactions	1	1
	7.03	7.3 Identify technology benefits/consequences	1	1
	7	**Standard 7: Science and Other Areas**	**4**	**4**
	7.02	7.2 Identify factors that may cause disease	1	1
	7.03	7.3 Explain the effects of overpopulation	1	1
	7.04	7.4 Explain science impact on food technology	2	2
	8	**Standard 8: History and Nature of Science**	**3**	**2**
	7.01	7.1 Explain role: human qualities in science	2	1
		Items not reporting to Benchmarks	1	1

Copyright © 2014 by North Dakota Department of Public Instruction

Table 2.1: Total Number of Items and Total Number of Score Points Measuring Each North Dakota Science Content Standard (cont'd)

Grade	Code	Standard / Benchmark	# of points	# of items
	A	**SCIENCE CONCEPTS AND INQUIRY**	**13**	**12**
	1	**Standard 1: Unifying Concepts**	**3**	**2**
	10.01	9-10.1 Use models to explain sci. principles	1	1
	10.02	9-10.2 Describe component interactions	2	1
	2	**Standard 2: Science Inquiry**	**10**	**10**
	10.01	9-10.1 Explain: new ideas from sci investig	1	1
	10.02	9-10.2 Use safety equipment and precautions	1	1
	10.03	9-10.3 Identify investig concepts/questions	2	2
	10.04	9-10.4 Formulate a testable hypothesis	1	1
	10.05	9-10.5 Identify variables, control, constants	1	1
	10.06	9-10.6 Design, conduct guided investigation	1	1
	10.08	9-10.8 Analyze data in tables, charts, graphs	3	3
	B	**PHYSICAL SCIENCE**	**15**	**13**
	3	**Standard 3: Physical Science**	**15**	**13**
	10.01	9-10.1 Classify elements: similar properties	1	1
	10.02	9-10.2 Classify physical/chemical changes	3	3
	10.03	9-10.3 Identify Law of Conservation of Matter	1	1
11	10.04	9-10.4 Construct a model of an atom	4	2
	10.05	9-10.5 Identify reactants/product in reaction	1	1
	10.06	9-10.6 Distinguish balanced chem. equations	1	1
	10.07	9-10.7 Use Newton's Laws to describe motion	2	2
	10.08	9-10.8 Desc. kinetic/potential energy changes	1	1
	10.10	9-10.10 Describe series and parallel circuits	1	1
	C	**LIFE SCIENCE**	**9**	**9**
	4	**Standard 4: Life Science**	**9**	**9**
	10.01	9-10.1 Relate cell function to cell structure	1	1
	10.02	9-10.2 Relate functions of cells to cell type	1	1
	10.04	9-10.4 Relate DNA, genes, and chromosomes	1	1
	10.05	9-10.5 Explain DNA changes: genetic variation	1	1
	10.06	9-10.6 Compare results of mitosis and meiosis	1	1
	10.07	9-10.7 Apply genetic concepts: predict traits	1	1
	10.08	9-10.8 Relate natural selection/evol. conseq.	1	1
	10.10	9-10.10 Explain energy, org: trophic pyramids	1	1
	10.11	9-10.11 Explain flow matter/energy: ecosystem	1	1

Copyright © 2014 by North Dakota Department of Public Instruction

Table 2.1: Total Number of Items and Total Number of Score Points Measuring Each North Dakota Science Content Standard (cont'd)

Grade	Code	Standard / Benchmark	# of points	# of items
	D	**EARTH AND SPACE SCIENCE**	**7**	**7**
	5	**Standard 5: Earth and Space Science**	**7**	**7**
	10.01	9-10.1 Explain the Big Bang Theory	1	1
	10.02	9-10.2 Relate atmosph change: plant evolution	1	1
	10.04	9-10.4 Relate physical proc: envir, society	3	3
	10.06	9-10.6 Explain effects: humans on environment	1	1
		Items not reporting to Benchmarks	1	1
	E	**SCIENCE AND CULTURE**	**14**	**14**
	6	**Standard 6: Science and Technology**	**5**	**5**
11	10.01	9-10.1 Use technologies to solve a problem	3	3
	10.02	9-10.2 Explain principles used in technology	1	1
	10.03	9-10.3 Explain impacts: emerging technology	1	1
	7	**Standard 7: Science and Other Areas**	**2**	**2**
	10.01	9-10.1 Relate pers. health: lifestyle factors	1	1
	10.02	9-10.2 ID factors that affect populations	1	1
	8	**Standard 8: History and Nature of Science**	**7**	**7**
	10.01	9-10.1 ID theoretical/applied scientist roles	4	4
	10.05	9-10.5 Explain: attitudes influence science	2	2
		Items not reporting to Benchmarks	1	1

Copyright © 2014 by North Dakota Department of Public Instruction

Part 3: Administration and Scoring

In this section, the scoring process used for North Dakota is described first, followed by the results of the inter-rater reliability study conducted on the handscoring of North Dakota constructed-response items.

Scoring of North Dakota test items adheres to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) 4.18, 4.20, and 6.8. Each of these Standards is cited in this part of the report. Standard 4.18 provides some general guidance for scoring test items:

> *Procedures for scoring and if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.*

This section of the report explains the procedures used for scoring the North Dakota constructed-response items. The scoring criteria used for each item are not presented in this section to preserve the integrity of the items for future use.

North Dakota Scoring Process

Multiple-choice items were scored by CTB using electronic scanning equipment. Constructed-response items were scored by human raters who were trained by CTB.

Selection of Scoring Evaluators

AERA, APA, & NCME (2014) Standard 4.20 specifies:

> *The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.*

CTB/McGraw-Hill strives to develop a highly qualified, experienced core of evaluators through a rigorous recruitment process so that the integrity of all projects is appropriately maintained.

Recruitment

The North Dakota 2014 project was staffed with a large number of returning evaluators and team leaders who had previous experience with North Dakota and other handscoring projects. Kelly

Copyright © 2014 by North Dakota Department of Public Instruction

Services also recruited new team leaders and evaluators for employment. Recruitment sources included advertisements in newspapers in Indianapolis, Indiana.

CTB requires that all evaluators and team leaders possess a bachelor's degree or higher. Kelly Services carefully screened all new applicants and required them to produce either a transcript or a copy of the degree. Kelly Services also required a one- to two-hour interview/screening process. Individuals who did not present proper documentation or who had less than desirable work records were eliminated during this process. Kelly Services verified that 100% of all potential evaluators met the degree requirement. All experienced evaluators and team leaders had already successfully completed the screening process.

The Interview Process

All potential evaluators completed a pre-interview activity. For some parts of the pre-interview activity, applicants were shown examples of test responses and were supplied with a scoring guide. In a brief introduction, they became acquainted with the application of a rubric. After the introduction, applicants applied the scoring guide to score the sample responses. The applicant's scores were used for discussion during the interview process to determine the applicant's trainability, as well as his/her ability to understand and implement the standards set forth in the sample scoring guide.

Kelly Services interviewed each applicant and determined the applicant's suitability for a specific content area and grade level. Applicants with strong leadership skills were questioned further to determine whether they were qualified to be team leaders.

When Kelly Services determined applicants were qualified, the applicants were recommended for employment. All assignments were made according to availability and suitability. Before being hired, all employees were required to read, agree to, and sign a nondisclosure agreement outlining the CTB/McGraw-Hill business ethics and security procedures.

Handscoring Training Process

Training Material Development

All materials necessary for scoring were developed by CTB. These materials include the scoring guides and training papers used to complete the handscoring of constructed-response and extended-response items (writing essays and performance events).

North Dakota operational items were previously field-tested or were selected from *TerraNova* Multiple Assessment forms. Prior to actual scoring, handscoring supervisors assembled materials based on the rubrics. Student answer documents were randomly sampled to ensure that a representative sample of possible responses was used. Supervisors selected anchor papers and training papers and recommended clarifications to rubrics. Training materials for new items field-tested in 2013 were also submitted for ND DPI review.

Copyright © 2014 by North Dakota Department of Public Instruction

Training and Qualifying Procedures

Handscoring involves training and qualifying team leaders and evaluators, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities. An explanation of the training and qualification procedures follows.

All readers were trained and qualified in a specific Rater Item Block (RIB) consisting of one item to be scored. Evaluators were trained using the following steps:

- Reviewing constructed-response items
- Reviewing rubrics
- Reviewing anchor papers
- Explaining scoring strategies, followed by a question-and-answer period
- Scoring a training set, followed by sharing established scores
- Qualifying Round 1
- Qualifying Round 2 (if necessary)
- Explaining condition codes and sensitive paper procedures
- Explaining unscannable image procedures

All evaluators were trained and qualified using the same procedures and criteria. In order to score an item, readers must have met the specific standards for that item. The qualification standards were:

- 6-point item: 70% qualification
- 5-point item: 70% qualification
- 4-point item: 70% qualification
- 3-point item: 80% qualification
- 2-point item: 90% qualification
- 1-point item: 95% qualification

Copyright © 2014 by North Dakota Department of Public Instruction

Monitoring the Scoring Process

AERA, APA, & NCME (2014) Standard 6.8 states:

> *Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.*

This subsection explains the monitoring procedures that CTB uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored. Detailed scoring rubrics are available for all CR items, which specify the criteria for scoring those CR items. The rubrics will not be presented here in order to preserve the integrity of the items for use in future North Dakota forms.

Daily Accuracy Checks

Throughout the course of handscoring, calibration sets of pre-scored papers (checksets/validity sets) were administered daily to each scorer to monitor scoring accuracy and to maintain a consistent focus on the established rubrics and guidelines. Checksets were executed via imaging software that provided images in such a way that the reader did not know when a checkset was administered.

In addition to the checkset process, CTB's handscoring protocol included the use of read-behinds. The read-behind was another valuable rater-reliability monitoring technique that allowed a team leader to review a reader's scored documents, providing feedback and counseling as appropriate.

Approximately 10% of operational Science papers were scored by a second reader to establish inter-rater reliability statistics for all constructed-response items. This procedure is called a "double-blind read" because the second reader does not know the first reader's score. For field test items, approximately 20% were scored by a second reader.

Recalibration of Raters

Once qualified, readers were required to maintain accuracy standards throughout the project. These requirements were assessed at the item level primarily through each reader's daily checkset performance, as well as agreement rates with other readers on double-blind reads and targeted read-behinds with team leaders. Data monitors generated reports daily that flagged any readers falling below the established quality standards for any item, providing insight on reader scoring trends (such as difficulty with a particular score point). These reports were shared with handscoring supervisors. Those readers identified in the reports received additional coaching, training, reviews, targeted read-behinds, or additional checksets. Readers who did not meet standards with these initial corrective actions were administered another validation (recalibration) round. Failure to recalibrate resulted in dismissal from the scoring assignment. This process was in place throughout the entire handscoring window.

Copyright © 2014 by North Dakota Department of Public Instruction

Security

Security guards were on site whenever employees were present in the building. All employees were issued photo identification badges and were required to wear them in plain view at all times. Visitors and employees who forgot their badges were issued visitors' badges and were required to wear them in plain view. All employees and visitors were subject to inspection of their personal effects.

Inter-Rater Reliability Results

Approximately 10% of the operational papers in Science were scored independently by a second reader. The statistics for the inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the two readers was examined.

For each item, a weighted kappa was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring. These weighted kappa values are presented in Table 3.1. To aid in the interpretation of kappa, the following cutoffs have been suggested (Landis & Koch, 1977; Altman, 1991):

Kappa Value	Strength of Agreement
0	None
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

As shown in Table 3.1, raters demonstrated at or above 99% *Perfect or Adjacent* agreement for all Science items. The weighted kappa statistic indicates that there was close to good or very good inter-rater agreement for all Science items.

Table 3.1: Inter-Rater Reliability, Science

Grade	Item #	# Points	% Perfect	% Perfect or Adjacent	Weighted Kappa
4	33	2	100	100	1.00
4	34	2	95	99	0.90
4	35	1	86	100	0.71
8	33	2	82	99	0.63
8	34	2	89	100	0.89
8	35	1	78	100	0.57
11	33	2	86	99	0.87
11	34	1	87	100	0.67
11	35	3	86	99	0.93

Copyright © 2014 by North Dakota Department of Public Instruction

Summary

The information presented in this section of the report summarizes the steps taken by CTB to ensure accuracy in the handscoring process. The inter-rater reliability statistics demonstrate that the items are scored reliably. These efforts by CTB address multiple best practices of the testing industry but are particularly related to AERA, APA, & NCME (2014) Standards 4.18, 4.20, and 6.8.:

- Standard 4.18— Procedures for scoring and if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
- Standard 4.20— The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.
- Standard 6.8— Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

Copyright © 2014 by North Dakota Department of Public Instruction

Part 4: Test Validity

The degree to which evidence and theory support accurate interpretations of test scores centers around the test scores' use; this is the heart of test validity. Validity is one of the most important components of any testing program. The analyses and narrative information found in this technical report constitute *validity evidence* that supports the intended uses and interpretations of the test scores for the North Dakota State Assessment (NDSA). The following text is from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) [hereafter referred to as "the *Standards*"]:

> *Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers. (p. 22)*

The validity of a testing program hinges on the use of the test scores. In gathering evidence related to the validity of the test scores, the first step should be to examine the uses of the test scores. In this section, some possible uses of the NDSA test scores are examined.

The following sections (Parts 4–9) provide technical evidence that the NDSA measures what it is intended to measure: the achievement of students on the North Dakota State Standards. It should be noted that any technical report cannot anticipate all possible interpretations and uses of test scores within a state. Rather, this technical report is designed to provide validity evidence for the purposes for which the test was originally designed. It is recommended, therefore, that policy and program evaluation studies be conducted in accordance with the *Standards* to provide additional support that NDSA scores are being used and are useful for their intended purposes.

Construct validity—the meaning of test scores and the inferences they support—is the central concept underlying the NDSA validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity.

Minimization of Construct-Irrelevant Variance and Under-Representation

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs being measured by a test. For example, when test items are not administered under standardized conditions (e.g., one administration may be timed, but another administration may be untimed), differences in student performance related to different administration conditions may result. Careful specification of the content and review of the items representing that content are first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data such as fit indexes, is used as a flag to investigate potential instances of construct-irrelevance variance.

Copyright © 2014 by North Dakota Department of Public Instruction

Construct under-representation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. The NDSA is designed to represent the North Dakota Content Standards.

Minimization of construct-irrelevant variance and construct under-representation is addressed in the following steps of the test development process: (a) test specifications, (b) item writing, (c) content and bias reviews of items, (d) field-testing, (e) test construction, and (f) calibration.

Convergent Validity

Convergent validity is a subtype of construct validity that can be estimated by the extent to which measures of constructs that theoretically should be related to each other are, in fact, observed as being related to each other. Analyses of the internal structure of a test can indicate the extent to which the relationships among test items conform to the construct the test purports to measure. For example, the NDSA Science test is designed to measure a single overall construct—Science achievement; therefore, the items comprising the Science NDSA should measure only Science achievement.

This technical report summarizes additional statistics that contribute to construct validity. The internal consistency coefficient (stratified alpha) reported in Part 5 is a measure of item homogeneity. In order for a group of items to be homogeneous, they must measure the same construct and represent the same content domain. Because IRT models were used to calibrate test items and to report student scores, item fit is also relevant to construct validity. The extent to which test items function as the IRT model prescribes is relevant to the validation of test scores. As will be shown in Part 7, very few items were flagged for poor model-data fit.

Principal Components Analysis

As another measure of construct validity, CTB examined the unidimensionality of each grade-level NDSA test. One of the underlying assumptions of the IRT models used to scale the Science NDSA is that the tests being calibrated are unidimensional; that is, items comprising the NDSA in each grade measure a single content domain (Reckase, 1979). For example, Science items should measure Science ability and not measure Reading skills. The unidimensionality assumption is, in practice, a testable hypothesis that is commonly evaluated through Principal Components Analysis (PCA). This analysis seeks evidence that there exists a single primary factor, the first principal component, which accounts for much of the relationship among test items. The presence of a single or dominant component suggests that a test is sufficiently unidimensional (i.e., it measures one underlying construct).

A PCA, using the correlation matrix, was conducted on each grade. A large, first principal component is evident in each analysis. It is common to have additional eigenvalues greater than 1.0, which may suggest the presence of other factors.

All of the NDSA grade tests exhibit large, first principal components, as evidenced in Table 4.1. The first principal component accounts for more than 12% of the test variance in all grades. To further investigate the unidimensionality of the Science tests, CTB explored the ratio of the first eigenvalue to the second eigenvalue (see Table 4.1). These ratios show that the first eigenvalue is

Copyright © 2014 by North Dakota Department of Public Instruction

at least four times as large as the second eigenvalue for all the grades. This difference in magnitude indicates that one factor appears to be dominant on the Science tests. This evidence supports the claim that there is a dominant dimension underlying the items in each test and that scores from each test represent performance primarily determined by that ability.

Table 4.1: Principal Components Analysis for Science

Grade	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Grade 4			
First Component	6.742	12.26%	12.26%
Second Component	1.620	2.95%	15.20%
Ratio (First/Second)	4.163		
Grade 8			
First Component	7.679	13.96%	13.96%
Second Component	1.545	2.81%	16.77%
Ratio (First/Second)	4.969		
Grade 11			
First Component	7.430	13.51%	13.51%
Second Component	1.674	3.04%	16.55%
Ratio (First/Second)	4.438		

Copyright © 2014 by North Dakota Department of Public Instruction

Part 5: Test Reliability

Reliability refers to the consistency of the students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly to the same population of individuals under similar conditions. Often, however, it is impractical to administer multiple forms of the test, and reliability is therefore estimated on a single administration of the test. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across items within a test during a single test administration (Crocker & Algina, 1986). Reliability is a necessary but insufficient condition of validity.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) indicate that

> *Typically, developers and distributors of tests have responsibility for obtaining and reporting evidence for reliability/precision (e.g., appropriate standards errors, reliability or generalized coefficients, or test information functions)* (p. 40).

In accordance with the *Standards*, the reliability of each NDSA test is investigated in a variety of ways: reliability of raw scores, IRT-based conditional standard error of measurement, and decision consistency of achievement level classifications.

Test Reliability Indices

The stratified alpha is an indication of the reliability of the test. It measures the internal consistency of a test comprising multiple-choice and constructed-response items. The stratified alpha accounts for computation error in the covariance matrix due to multiple-item formats by stratifying items by the number of score points attributed to them. Like the more common coefficient alpha or the Kuder-Richardson 20, the stratified alpha is a lower-bound estimate of test reliability. The number of items on the test influences reliability indices—longer tests can be expected to be more reliable than shorter tests.

The reliability coefficient is a ratio of the variance of true scores to observed scores, with the values ranging from 0 to 1. The closer the value of the reliability coefficient is to 1, the more consistent the scores, where 1.00 refers to a perfectly consistent test. As a rule of thumb, reliability coefficients that are equal to or greater than 0.85 are considered acceptable for tests of moderate lengths.

Total test reliability measures, such as the stratified alpha and the standard error of measurement, consider the consistency (reliability) of performance over all test questions in a given form, the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations.

The reliability coefficients and the standard errors of measurement (SEMs) for the NDSA are reported in Table 5.1 and Table 5.2. Table 5.1 reports the stratified alpha and coefficient alpha

Copyright © 2014 by North Dakota Department of Public Instruction

for the calibration sample. Table 5.2 shows the reliability indices for the following subgroups of students:

- Gender groups, including both male and female students;
- Ethnic groups, including White, Asian, American Indian, Hispanic, and Black students;
- Limited English Proficiency (LEP) students;
- Students eligible to receive a free/reduced-price lunch under the National School Lunch Program[1];
- Students with disabilities qualifying under Section 504[2] and receiving services identified in an Accommodation Plan; and
- Students with disabilities receiving instruction under an Individualized Education Plan[3] (IEP).

For most reported groups of students, test reliability coefficients were high and the standard errors of measurement were low, indicating good test internal consistency (reliability) for the total population and subpopulations of North Dakota examinees.

Note that the reliability coefficients and standard error of measurement were computed for all ethnicity subgroups, including those with a little more than 100 students (e.g., Asian students). The statistics for such a small group of students should be interpreted with caution.

Standard Error of Measurement

The reliability of reported test scores can be characterized by the IRT standard errors associated with the individual test scores. The standard error of measurement may be used to determine the range within which a student's true score is likely to fall. An observed score is not a student's true score; rather, it is an estimate of a student's true score. Sixty-eight percent of the time a student's observed score obtained from a single test administration would fall within one SEM of the student's true score were the test administered on other occasions. Similarly, 95% of the time, a student's observed score obtained from a single test administration would fall within two SEMs of the student's true score were the test administered on other occasions.

[1] The National School Lunch Program (NSLP) is a federally assisted meal program operating in public and nonprofit private schools and residential child care institutions. It provides nutritionally balanced, low-cost or free lunches to children each school day. The program was established under the National School Lunch Act, signed by President Harry Truman in 1946. In this study, eligibility to receive a free/reduced price lunch is considered a substitute variable for a student's family socioeconomic status.

[2] To qualify under Section 504 of the Rehabilitation Act of 1973 a student must:

- Be determined to have a physical or mental impairment that *substantially limits* one or more major life activities, including learning and behavior;
- Have a record of having such an impairment; OR
- Be regarded as having such impairment.

[3] A plan mandated by the Individuals with Disabilities Education Act (IDEA) that requires public schools to develop an IEP for every student with a disability who is found to meet the federal and state requirements for special education.

Copyright © 2014 by North Dakota Department of Public Instruction

Figure 5.1 displays the conditional SEM curves for each grade for Science. The estimates of measurement error tend to be higher at the low and high ends of the scale score range. The measurement error increases when there are few observations at a particular ability level. Generally, there are few students with extreme scores, and these score levels cannot be estimated as accurately as levels toward the middle of the ability range. Figure 5.1 demonstrates that the tests are designed so that measurement error is minimized in the middle of the scale range, where the majority of students are located.

Decision Accuracy and Consistency

The *Standards* make reference to the following measurement consideration on classification consistency:

> *Some authorities have proposed that a semantic distinction be made between "reliability of scores" and "degree of agreement in classification." The former term would be reserved for analysis of score variation under repeated measurement. The term* 'classification consistency*' rather than reliability would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee's score to the cut score* (p. 30).

Decision Accuracy: Decision accuracy is defined as the extent to which the actual classifications of test takers into various achievement levels agree with classifications that would be made on the basis of their true scores (Livingston & Lewis, 1995). That is, decision consistency refers to the agreement between two observed scores, whereas decision accuracy refers to the agreement between the observed score and the true score.

Decision Consistency: Classification consistency or decision consistency is defined as the extent to which the classifications of students in a particular achievement level agree on the basis of two independent administrations of the test or one administration of two parallel test forms. It is often logistically infeasible, as well as expensive, to obtain data from repeated administrations of a test, be it re-administration of the same test or administration of a parallel form. Therefore, a common practice is to estimate decision consistency from one administration of a test.

The Livingston-Lewis (1995) methodology was used to calculate decision accuracy statistics based on the 2014 NDSA results. The Livingston-Lewis procedure utilizes a beta-binomial model that requires two steps: (a) fitting proportion-correct true scores to a four-parameter beta distribution and (b) using the binomial distribution to estimate classification accuracy and consistency. All calculations for decision accuracy and consistency are based on census data.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. These results, as well as the results of the overall accuracy and consistency of classification across all cut points, are reported in Table 5.3. To evaluate decisions at specific cut points, the joint distribution of all of the performance levels is collapsed into a dichotomized distribution around that specific cut point. For example, the dichotomization at the cut point between the *Partially Proficient* and *Proficient* classifications was formed, and students were either considered *Partially Proficient* or

Copyright © 2014 by North Dakota Department of Public Instruction

Proficient. By analyzing this dichotomization, we can estimate how accurately and consistently the *Proficient* cut score separated students into these two categories. As shown in Table 5.3, all accuracy and consistency statistics conditioned on individual cut points are at or above 89% and 84%, respectively. The accuracy and consistency statistics across all points were at or above 73% and 64%, respectively.

Table 5.1: Reliability Coefficients and SEMs for the Total Population, Science

Grade	Number of Items	Number of Score Points	Stratified Alpha	Stratified Alpha SEM	Coefficient Alpha	Coefficient Alpha SEM
4	55	57	0.85	2.89	0.85	2.88
8	55	57	0.87	2.99	0.87	2.97
11	55	58	0.86	3.33	0.87	3.20

Copyright © 2014 by North Dakota Department of Public Instruction

Table 5.2: Test Reliability Coefficients and SEMs by Subgroup, Science

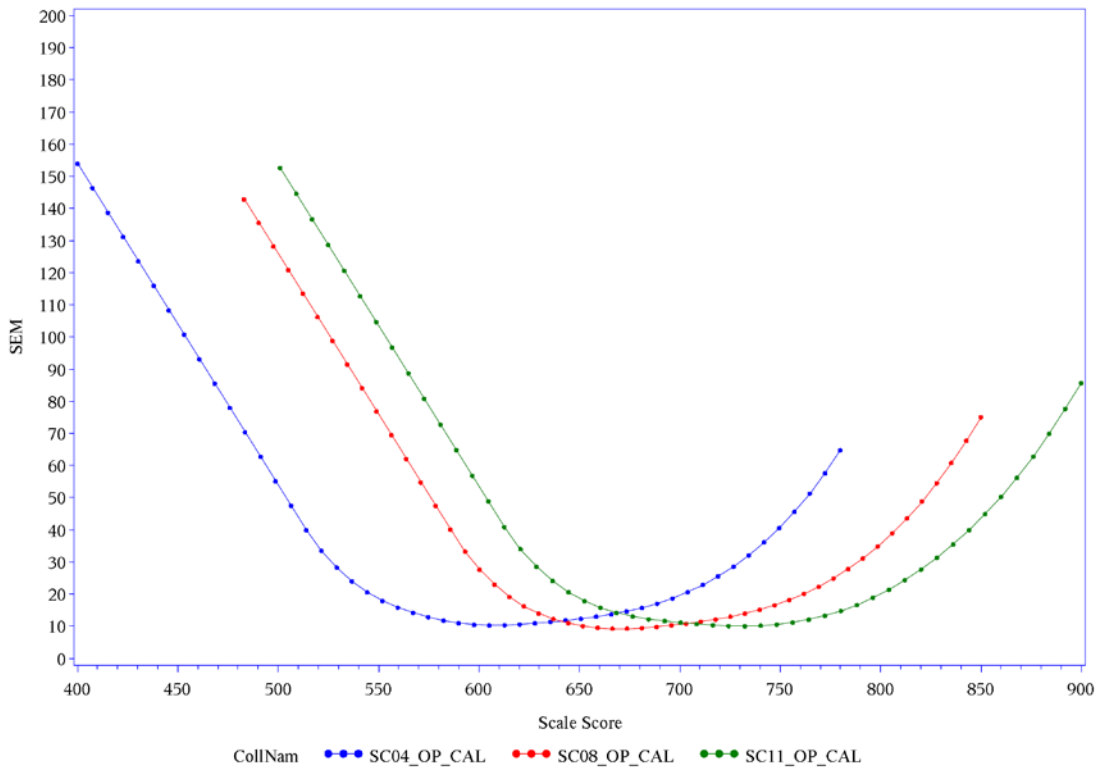
Grade	Number of Items	Number of Points	Subgroup	Category	N Count	Coefficient Alpha	Coefficient Alpha SEM
4	55	57	Gender	Male	4492	0.85	2.85
				Female	4047	0.84	2.90
			Ethnicity	White	6170	0.82	2.80
				Asian	116	0.90	3.02
				Am Indian	891	0.84	3.13
				Hispanic	323	0.84	3.07
				Black	317	0.86	3.12
			LEP	Yes	202	0.85	3.39
			NSLP	Yes	2689	0.86	3.04
			504 Plan	Yes	145	0.83	2.90
			IEP	Yes	991	0.86	3.13
8	55	57	Gender	Male	4040	0.88	2.93
				Female	3866	0.86	3.00
			Ethnicity	White	5902	0.84	2.91
				Asian	102	0.92	3.03
				Am Indian	758	0.87	3.23
				Hispanic	284	0.85	3.17
				Black	260	0.90	3.20
			LEP	Yes	215	0.82	3.41
			NSLP	Yes	2327	0.87	3.16
			504 Plan	Yes	185	0.83	2.95
			IEP	Yes	923	0.86	3.28
11	55	58	Gender	Male	3824	0.88	3.15
				Female	3623	0.86	3.23
			Ethnicity	White	5662	0.86	3.16
				Asian	117	0.92	3.29
				Am Indian	498	0.85	3.34
				Hispanic	172	0.86	3.32
				Black	202	0.88	3.38
			LEP	Yes	114	0.74	3.45
			NSLP	Yes	1513	0.88	3.30
			504 Plan	Yes	185	0.86	3.26
			IEP	Yes	653	0.85	3.39

Copyright © 2014 by North Dakota Department of Public Instruction

Table 5.3: Decision Accuracy and Consistency at Achievement Cut Points

		Accuracy				Consistency			
		Novice/ Part. Prof.	Part. Prof./ Prof.	Prof./ Adv.	Across All Cut Points	Novice/ Part. Prof.	Part. Prof./ Prof.	Prof./ Adv.	Across All Cut Points
Science	4	0.94	0.89	0.91	0.73	0.91	0.84	0.87	0.64
	8	0.95	0.89	0.91	0.75	0.93	0.85	0.88	0.66
	11	0.94	0.89	0.92	0.75	0.92	0.84	0.89	0.65

Figure 5.1: Standard Error of Measurement Curves for Science



Copyright © 2014 by North Dakota Department of Public Instruction

Part 6: Test and Item Statistics

In this section, summary test statistics for each NDSA grade are presented. These are followed by item-level statistics for each grade.

Test-Level Statistics

Table 6.1 presents the mean and standard deviations of the *p*-values and item-total correlations (point-biserial correlation for multiple-choice [MC] items). The statistics reported in this table are based on the total population of North Dakota students. The mean *p*-value is the average of all item *p*-values of a specific grade. The mean item-total correlation is the average of all item-total correlations of a specific grade. The *p*-value is a measure of item difficulty. Simply, the *p*-value for MC items is the number of students correctly responding to an item divided by the total number of students attempting the item.

The average *p*-values for Science ranged from 0.69 to 0.75 (see Table 6.1). Test scores tend to be most accurate when their average *p*-values are in the mid 0.50s to low 0.70s. High average *p*-values indicate that there may be few difficult items on the test (i.e., the test is easy for the students) and possible ceiling effects on the tests.

The item-total correlation is a measure of item discrimination. It is the correlation between an item and the total test score. It indicates how well an item differentiates between low- and high-achieving students. In general, items with item-total correlations below 0.15 are considered to be poorly discriminating. As is demonstrated in Table 6.1, the average item-total correlation was well above 0.15 for each grade.

Item-Level Statistics

Tables 6.2 through 6.4 present item-level *p*-values and item-total test correlations for each grade. It is important that one examine the range of *p*-values and not just the average *p*-value for a test to draw a conclusion about the capability of the test to measure well. It is desirable for the test to measure well throughout the range of skills present at a given grade. That is, it is important that the items measure the performance of both low-scoring and high-scoring students as well as students in the center of the distribution. Having a range of item difficulties also helps to prevent floor and/or ceiling effects; that is, the test does not have large numbers of students at the minimum or maximum possible scores. Examination of the individual *p*-values (see Tables 6.2 through 6.4) shows that the NDSAs have items with most *p*-values ranging from the middle 0.20s to the 0.90s. A broad range of *p*-values suggests that the items measure throughout the range of skills at a given grade. Tables 6.2 through 6.4 also show that almost all items had item-total test correlations above the 0.15 threshold.

Copyright © 2014 by North Dakota Department of Public Instruction

Table 6.1: NDSA Means, Standard Deviations for *P*-Values, Item-Total Correlation

Content Area	Grade	Mean *P*-Value (SD)	Mean Item-Total Correlation (SD)	Total Items	Total Points
Science	4	0.75 (.18)	0.29 (.07)	55	57
	8	0.72 (.18)	0.32 (.08)	55	57
	11	0.69 (.17)	0.31 (.08)	55	58

Copyright © 2014 by North Dakota Department of Public Instruction

Table 6.2: Item Statistics: Grade 4

Science					
Item	***P*-Value**	**Item-Total Test Correlation**	**Item**	***P*-Value**	**Item-Total Test Correlation**
1	0.60	0.23	31	0.42	0.22
2	0.88	0.23	32	0.45	0.27
3	0.80	0.33	33	0.88	0.20
4	0.81	0.35	34	0.92	0.30
5	0.93	0.19	35	0.39	0.29
6	0.95	0.13	41	0.66	0.27
7	0.92	0.22	42	0.82	0.25
8	0.88	0.19	43	0.81	0.38
9	0.81	0.41	44	0.59	0.31
10	0.86	0.36	45	0.49	0.24
11	0.91	0.39	46	0.67	0.20
12	0.92	0.26	47	0.80	0.39
13	0.54	0.22	48	0.35	0.23
14	0.91	0.45	49	0.62	0.37
15	0.91	0.40	50	0.50	0.36
16	0.60	0.28	51	0.84	0.39
17	0.77	0.42	52	0.64	0.21
18	0.96	0.30	53	0.72	0.40
19	0.81	0.27	54	0.90	0.33
20	0.71	0.25	55	0.50	0.28
21	0.75	0.30	56	0.95	0.28
22	0.88	0.22	57	0.74	0.32
23	0.65	0.29	58	0.84	0.34
24	0.84	0.26	59	0.77	0.26
25	0.96	0.29	60	0.61	0.34
26	0.40	0.20			
27	0.83	0.34			
28	0.97	0.21			
29	0.47	0.25			
30	0.88	0.31			

Copyright © 2014 by North Dakota Department of Public Instruction

Table 6.3: **Item Statistics: Grade 8**

Science					
Item	***P*-Value**	**Item-Total Test Correlation**	**Item**	***P*-Value**	**Item-Total Test Correlation**
1	0.84	0.23	31	0.59	0.41
2	0.96	0.22	32	0.85	0.40
3	0.64	0.38	33	0.15	0.32
4	0.40	0.23	34	0.72	0.26
5	0.77	0.12	35	0.45	0.29
6	0.79	0.28	41	0.60	0.13
7	0.68	0.27	42	0.61	0.27
8	0.91	0.33	43	0.87	0.46
9	0.80	0.28	44	0.66	0.24
10	0.80	0.20	45	0.73	0.42
11	0.78	0.23	46	0.44	0.26
12	0.77	0.29	47	0.52	0.39
13	0.88	0.31	48	0.76	0.44
14	0.92	0.34	49	0.54	0.28
15	0.85	0.44	50	0.79	0.43
16	0.79	0.33	51	0.40	0.31
17	0.81	0.41	52	0.48	0.19
18	0.95	0.19	53	0.86	0.38
19	0.68	0.38	54	0.83	0.45
20	0.87	0.37	55	0.39	0.26
21	0.86	0.36	56	0.79	0.28
22	0.78	0.45	57	0.81	0.34
23	0.48	0.32	58	0.91	0.30
24	0.91	0.37	59	0.52	0.31
25	0.73	0.33	60	0.60	0.30
26	0.68	0.25			
27	0.84	0.33			
28	0.77	0.43			
29	0.96	0.33			
30	0.66	0.35			

Copyright © 2014 by North Dakota Department of Public Instruction

Table 6.4: **Item Statistics: Grade 11**

Science					
Item	***P*-Value**	**Item-Total Test Correlation**	**Item**	***P*-Value**	**Item-Total Test Correlation**
1	0.96	0.23	31	0.80	0.33
2	0.72	0.26	32	0.51	0.36
3	0.93	0.34	33	0.38	0.38
4	0.77	0.13	34	0.29	0.41
5	0.79	0.32	35	0.43	0.50
6	0.85	0.11	41	0.93	0.38
7	0.87	0.32	42	0.25	0.34
8	0.87	0.29	43	0.59	0.40
9	0.78	0.29	44	0.61	0.32
10	0.64	0.35	45	0.61	0.38
11	0.82	0.18	46	0.67	0.39
12	0.88	0.22	47	0.41	0.30
13	0.78	0.22	48	0.45	0.28
14	0.90	0.35	49	0.37	0.25
15	0.77	0.34	50	0.50	0.40
16	0.67	0.34	51	0.69	0.38
17	0.80	0.31	52	0.55	0.42
18	0.87	0.35	53	0.70	0.29
19	0.60	0.24	54	0.82	0.30
20	0.76	0.20	55	0.81	0.32
21	0.90	0.30	56	0.69	0.20
22	0.77	0.41	57	0.55	0.36
23	0.85	0.26	58	0.65	0.25
24	0.56	0.37	59	0.83	0.19
25	0.86	0.34	60	0.66	0.31
26	0.62	0.45			
27	0.82	0.28			
28	0.62	0.28			
29	0.57	0.38			
30	0.61	0.35			

Copyright © 2014 by North Dakota Department of Public Instruction

Part 7: Scaling and Linking

Over 99% of all North Dakota students who took the Science tests were included in the calibration sample. In this section, first a general overview of scaling is provided, and then the item response theory (IRT) models used for calibrating the data are discussed. Next, the model-data fit for the North Dakota data is addressed. If the IRT model fits the empirical item response distributions for the population for which generalizations are going to be made (i.e., North Dakota students), then the claim that the scores are valid indicators of an underlying ability is strengthened. The methods used to link the NDSA results to the NDSA scale and student scoring procedures are discussed. Finally, the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for the NDSA are presented.

Scaling

The purpose of scaling a test is to enhance its validity by increasing the comparability of test takers' scores. For all test configurations, scale scores are based on performance on all the items in the test. To obtain this scale score, number-correct (raw) scoring is used. This procedure produces maximum-likelihood trait estimates based on raw scores (Yen, 1984).

Scaling and linking of complex assessment data were performed using PARDUX (Burket, 1995), which is proprietary software developed by CTB/McGraw-Hill. PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both multiple-choice (MC) items and constructed-response (CR) items. In PARDUX, items are calibrated based on IRT, using the 3PL model (Lord & Novick, 1968) for MC items and the 2PPC model (Muraki, 1992; Yen, 1993) for CR items. PARDUX is also used to link the scales developed by two calibrations through the use of common items and the characteristic curve alignment procedure of Stocking and Lord (1983).

Item Response Theory

A marginal maximum-likelihood procedure was used to simultaneously estimate the item parameters using the 3PL and 2PPC IRT models (Bock & Aitkin, 1981; Thissen, 1982). Under the 3PL model, the probability that a student with trait or scale score θ will respond correctly to multiple-choice item *j* is

$$P_j(\theta) = c_j + (1 - c_j)/[1 + \exp(-1.7a_j(\theta - b_j))]. \qquad (1)$$

In equation (1), a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-scoring student. Under the 2PPC model, the probability that a student with trait or scale score θ will respond in category *k* to partial-credit item *j* is

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}), \qquad (2)$$

Copyright © 2014 by North Dakota Department of Public Instruction

where $z_{jk} = (k-1)f_j - \sum_{i=0}^{k-1} g_{ji}$, and $g_{j0} = 0$ for all j.

The summary output of the 3PL and 2PPC models is in two different metrics. The location and discrimination parameters for the MC items are in the traditional 3PL metric and are labeled *b* and *a*, respectively. In the 2PPC model, *f* and *g* (gamma) are analogous to *b* and *a*, where alpha is the discrimination parameter and gamma over alpha (*g/f*) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL parameters *b* and *a* are not directly comparable to the 2PPC parameters *f* and *g*; however, they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f/1.7$ (Burket, 1995). As a result of this procedure, the MC and CR items are placed on the same scale. Note that for the 2PPC model there are $m_j - 1$ (where m_j is a score level *j*) independent *g*'s and one *f*, for a total of m_j independent parameters estimated for each item, while there is one *a* and one *b* per item in the 3PL model.

Model Fit

A procedure developed by Yen (1981) was used to assess model-to-data fit for all test items. In this procedure, students' ranks are ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells, with 10% of the sample in each cell. Each item *j* in each decile *i* has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item *j* in category *k*. The observed proportion O_{ijk} is also tabulated for each decile and approximates the chi-square statistic according to the following equation:

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij}(O_{ijk} - E_{ijk})^2}{E_{ijk}}.$$

Q_{1j} should be approximately chi-square distributed with degrees of freedom (*DF*) equal to the number of "independent" cells, $10(m_j-1)$, minus the number of estimated parameters. For the 3PL model, $m_j = 2$, so $DF = 10(2-1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 10$. Since *DF* differs between MC and CR items and between CR items with different score levels m_j, Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and cutoff values for flagging an item based on Z_j have been developed and were used to identify items for the item review. The cutoff value is (*N*/1500 x 4) for a given test, where *N* is the sample size.

The *Z*-statistic is an index of the degree to which observed proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. These values are computed for ten intervals corresponding to deciles of the theta

Copyright © 2014 by North Dakota Department of Public Instruction

distribution. The *Z*-statistic is used to characterize item fit. The critical value of *Z* is different for each grade because it is dependent on sample size.

In Science, one item (item 34) was flagged for poor fit in grade 8. Table 7.1 shows the chi-square statistic and the *Z*-statistic for the flagged item. The average percent across ten cells of observed percentage correct and predicted percentage correct is also provided. The difference between the observed and predicted percentages provides an indication of how well the modeled response curves reflect the empirical curves.

The flagged item was examined more closely by studying its item characteristic curve (ICC) at each non-zero score point. The ICC models the relationship between an examinee's performance on an item and the examinee's underlying ability. In almost all cases for which model misfit occurs, relatively fewer students occupy the scale score ranges where the misfit is prominent. Poor fit may occur in one region of the underlying ability distribution where there are relatively few students at that particular point in the distribution. The model functions well for the flagged items in the middle of the theta distribution where the majority of students perform.

Figure 7.1 presents the ICC for the misfitting item (item 34) on the grade 8 Science NDSA. This is a 2-point, CR item. As shown, there is poor fit at the lower and upper ends of the distribution for Levels 1, 2, and 3 (Score Points 0, 1, and 2, respectively).

The smooth line in each of these figures represents the predicted relationship between examinee performance on the item and examinee ability, and the jagged line represents the observed relationship.[4] Large differences between the two lines indicate poor fit. Each figure also shows the distribution of scale scores so that the fit between observed and predicted performance at different ability levels can be interpreted in light of the number of examinees at each level.

With a large number of observations, items may be flagged for small but statistically significant differences that may not be of practical importance. In the case of the one NDSA item flagged for misfit, the differences do not seem to be of practical importance. Because the understanding of the causes of misfit are the subject to continuing investigation, misfitting items that have content validity are often retained for use in an assessment and monitored over a period of usage. A large number of misfitting items in an assessment would indicate that caution should be exercised in the interpretation of the overall score.

Linking Method and Results

After item calibrations, all Science items were linked to the 2006 NDSA Science scale using the Stocking and Lord (1983) procedure. For the linking, 19–21 MC items per form from the 2004 through 2012 NDSA served as anchors. Linking was performed using a test characteristic curve (TCC) method proposed by Stocking & Lord (1983) and implemented using PARDUX (Burket, 1995).

[4] For constructed-response items, there will be one graph for each score level. For example, a 2-point item will have three score-level graphs, one each for 0, 1, and 2 score points. An overall ICC across score levels is also provided.

Copyright © 2014 by North Dakota Department of Public Instruction

Anchor Evaluation

The same anchor sets from 2013 were re-used in the 2014 administration. The evaluation of anchor items is a routine part of the linking process. The reliability of the anchor set is critical to the longitudinal stability of a testing program because performance on the anchor items helps a state evaluate student growth in relation to the content standards. Anchor item functioning is investigated using techniques to diagnose item parameter drift (IPD). From the outset, the causes of IPD should be divided into two categories: construct-relevant and construct-irrelevant. In examining growth across years, states will want to capture construct-relevant IPD because this is a change in ability that is purposeful. For example, states may mandate a new curriculum to align to a change in standards (Bock, Muraki & Pfeiffenberger, 1988; Demars, 2004). Over time, we may expect to see performance on items change as teachers begin implementing the new curriculum. We may see positive or negative changes as a new curriculum is implemented with some items getting more difficult and some getting easier. This type of change assumes that test forms are being equated over several years, not just over a single year.

At the state level, we do not expect to see rapid changes in item performance over the course of a year. If an item is shared between two contiguous years, it seems unlikely that a change in curriculum could be implemented to such a degree across the state that it would improve performance on an item. In these cases, it is more likely that we are capturing construct-irrelevant IPD. This type of IPD includes item exposure (Glas, 2000), multidimensionality (Li & Reckase, 2009) and the following noted by Kolen & Brennan (2004):

- ✓ Position of the common item on the two forms to be equated has changed
- ✓ The number of items on the test increased or decreased
- ✓ An anchor item has been edited from one form to another form
- ✓ Answer choices in a multiple-choice item rearranged
- ✓ Changes in the font used on the test or in the pagination used
- ✓ Changes in timing of the test between one form and another form
- ✓ Changes in motivation conditions between one form and another form
- ✓ Changes in the answer sheet design between one form and another form
- ✓ Items scrambled for security purposes
- ✓ Change in the calculator use policy (or dictionaries or word processors)
- ✓ Section pre-equating occurs in which pre-equating and operational sections appeared in different positions in different forms

Despite the best efforts of test developers, the changes listed above may result in an increase in student achievement that does not reflect construct-relevant growth.

Unfortunately, it is not always possible to separate construct-related growth from construct-irrelevant drift. Miller and Fitzpatrick (2009) advise that if construct-irrelevant factors cannot explain observed IPD, then assume that the drift is related to the construct being measured. If the drift is related to the construct, the item should not be removed from the equating coefficient.

Following Miller and Fitzpatrick (2009), CTB does not recommend removing an item from the anchor set unless the item is clearly performing differently than expected and due to factors extraneous to the construct. An anchor item should not be removed based solely on its impact on

Copyright © 2014 by North Dakota Department of Public Instruction

score trends. And, the removal of an anchor item should not adversely affect the content structure of the anchor set such that a different construct is being measured.

The anchor items were evaluated immediately following the calibration and linking of the NDSA. For the North Dakota Science assessments, four statistical methods are used to evaluate anchor items: (1) iterative linking (Candell & Drasgow, 1988) using Stocking and Lord's (1983) test characteristic curve method; (2) delta plots (Angoff, 1972, 1982; Dorans and Holland, 1993); (3) Lord's chi-square criterion (p. 223, Lord, 1980); and (4) differences between the item-ability regression curves.

While removing an anchor item based solely on statistical criteria can simplify decision making, it may inadvertently change the content coverage, shift scale score distributions, and affect the classification of students into different proficiency levels. This is especially true when the number of anchor items in the set is small. Before an anchor item may be dropped from the anchor set, the adequacy of the content coverage must be evaluated. For the NDSA, anchor items are considered for exclusion under the following conditions:

1. Items flagged using the TCC method are considered for exclusion when the correlation between the input and estimated item parameters is below 0.80 for the *a*-parameter and below 0.90 for the *b*-parameter. If the exclusion of an outlying anchor item increases the *a*-parameter correlation to above 0.80 or increases the *b*-parameter correlation to above 0.90, then the anchor is a candidate for removal.
2. An anchor is a candidate for removal when the item is flagged using both the delta-plot and Lord's chi-square OR when the item is flagged on three of the seven statistics considered when examining the large differences between the IRT regression curves.
3. Removal of the item may not significantly alter the content distribution of the anchor set. The distribution of items across the content standards must remain within 10% of the 2006 item distribution for Science.
4. Beginning in 2011, a new guideline was added. It is important to recognize that differential item performance in two test administrations does not necessarily indicate item flaws and may be affected by population differences, differences in teaching strategies, curriculum changes, etc. Therefore, CTB recommends that DPI consider item content-related factors in addition to statistical evidence of differential item performance in two test administrations.

Item 40 in grade 4 was observed with a *p*-value of 0.37 in the 2010 administration, and 0.50 in the 2014 administration. Given that the correlation between the *a*-parameter inputs and estimates and those for the *b*-parameter correlation at grade 4 were within the prescribed criteria (at or above 0.80 for *a* and 0.90 for *b*), item 40 would not be considered for removal based on being flagged by the Stocking and Lord Method alone.

Vertical Scale

In the baseline test administration year, the NDSA operational Science assessments were linked to the respective *TerraNova* scales. This linking allows North Dakota to take advantage of the

Copyright © 2014 by North Dakota Department of Public Instruction

vertical properties of the *TerraNova* scale. The underlying vertical scale enables comparisons to be made across grades directly on the test scale. When making cross-grade comparisons, it is important that comparisons be limited to contiguous grades. While the vertical scale that underlies a content area measures a single content domain (e.g., Science), the aspects of the content domain being measured will vary across the grades. Patz and Hanson (2002) noted, "The validity of inferences based on such a linking [i.e., vertical scale] will depend on the reliability or strength of the link, and this reliability will generally diminish with the distance between levels under most vertical scaling research designs." The reliability of the link diminishes partly because the content being assessed changes as the grade levels become further apart. For this reason, it is inappropriate to compare performance on the Science NSDA from grade 4 to grade 8 to grade 11 using scale scores, even though the Science NDSA is linked to the *TerraNova* scale.

Scoring Procedures

North Dakota students were scored using the number correct (NC) scoring method. This method considers how many score points a student obtained on a test in determining his or her score. That is, two students with the same number of score points on the test will receive the same score regardless of which items they answered correctly. In this method, the number correct (or raw) score on the test is converted to a scale score by means of a conversion table. This traditional scoring method is often preferred for its conceptual simplicity and familiarity.

Because the 2014 administration re-used test forms from the 2013 administration, it was proposed to re-use the 2013 scoring tables in 2014 as well. Appropriateness of reusing the table was examined by comparing obtained TCCs between 2013 and 2014 after applying the final Stocking-Lord equating constants. As illustrated in Figure 7.2, the 2013 and 2014 TCCs overlap well throughout the ability continuum at each of the three grades. The matched TCCs support reusing the 2013 scoring tables in the 2014 administration. The scoring tables are presented in Appendix A.

In the 2013 administration, the final item parameters in the scale score metric were used to produce raw score-to-scale score conversion tables for the North Dakota tests. An inverse TCC method was employed. The scoring tables were created using CTB/McGraw-Hill's FLUX program. The inverse of the TCC procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points (Yen, 1984). All North Dakota tests have a maximum raw score higher than 57 points. In the inverse TCC method, a student's trait estimate is taken to be the trait value that has an expected raw score equal to the student's observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order approximation to number correct maximum likelihood estimates (MLE), showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

The inverse of the TCC method relies on the following equation:

Copyright © 2014 by North Dakota Department of Public Instruction

$$\sum_{i=1}^{n} v_i x_i = \sum_{i=1}^{n} v_i E(X_i \mid \tilde{\theta}),$$

where

x_i is a student's observed raw score on item *i;*

v_i is a weight specified in a scoring process (v_i = 1 if no weights are specified); and

$\tilde{\theta}$ is a trait estimate.

Lowest and Highest Obtainable Scale Scores

A maximum likelihood procedure cannot produce scale score estimates for students with perfect scores or scores below the level expected by guessing. Also, although maximum likelihood estimates are available for students with extreme scores other than zero or perfect, occasionally these estimates have standard errors of measurement that are very large, and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure. These values, which are set separately by grade, are called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). Table 7.2 gives the LOSS and HOSS values used for each grade of the Science NDSAs.

To further examine the tests for the possibility of ceiling effects, the distribution of student performance on the scale scores was examined. Table 7.3 summarizes the number of students obtaining the HOSS and the penultimate HOSS (one scale score below the HOSS) on the NDSA for Science. As the table shows, there are very few students in any grade achieving the HOSS for the content areas, suggesting that there is room for student growth.

Copyright © 2014 by North Dakota Department of Public Instruction

Table 7.1: Item Fit Statistics for Misfitting Items

Content	Grade	Item	Chi Square	DF	Total N	Z	Observed	Predicted	Obs-Pred
Science	8	34	155.74	17	7924	23.79	0.72	0.72	0.00

Table 7.2: LOSS and HOSS Values by Grade

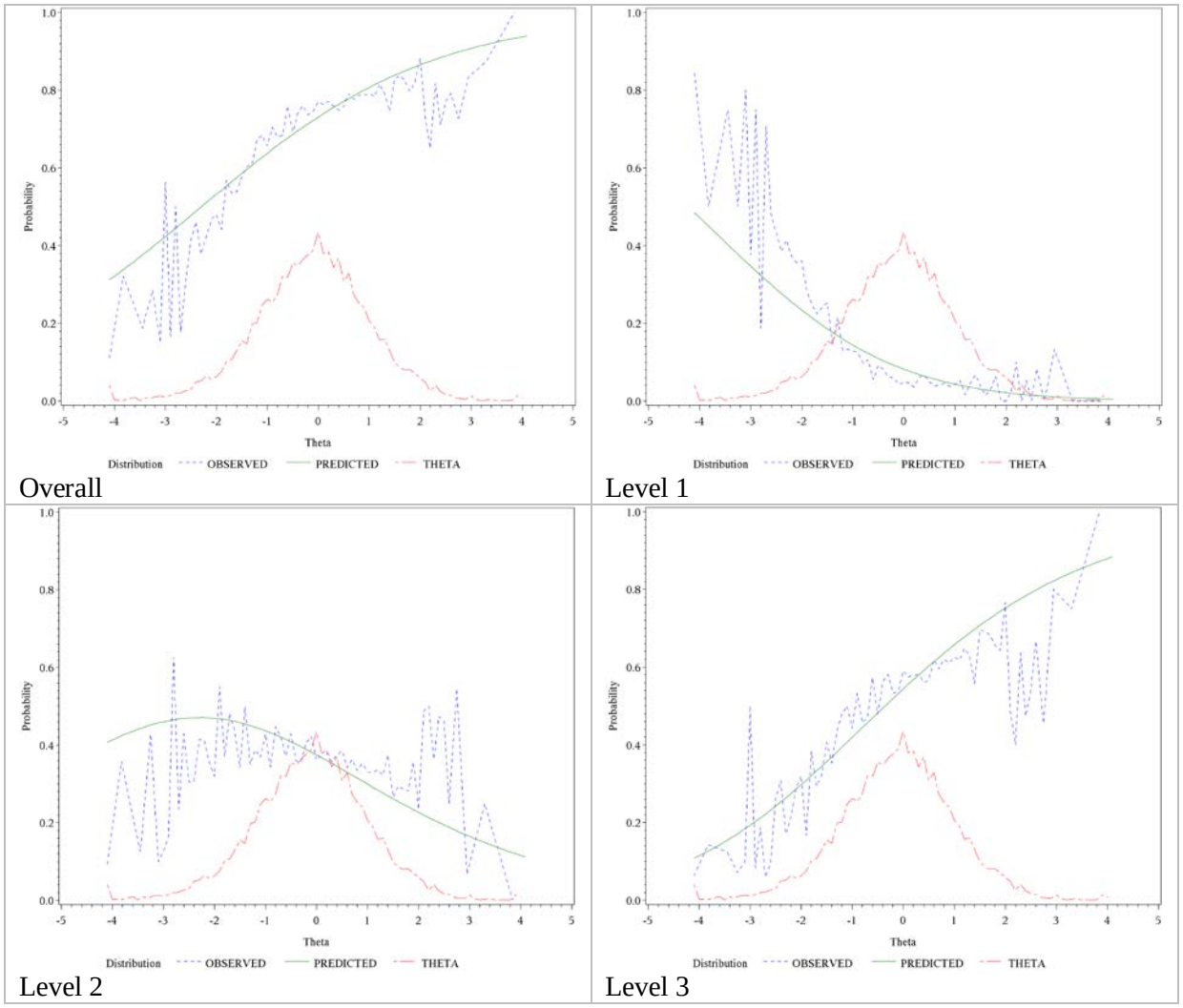
Grade	Science	
	LOSS	HOSS
4	400	780
8	483	850
11	501	900

Table 7.3: Number and Percent of Students at the HOSS and Penultimate HOSS, by Grade for Science

				# Students at the		% Students at the	
Grade	N	HOSS	Penultimate HOSS	HOSS	Penultimate HOSS	HOSS	Penultimate HOSS
4	8564	780	741	7	40	0.08	0.47
8	7937	850	806	3	16	0.04	0.20
11	7463	900	842	11	22	0.15	0.29

Copyright © 2014 by North Dakota Department of Public Instruction

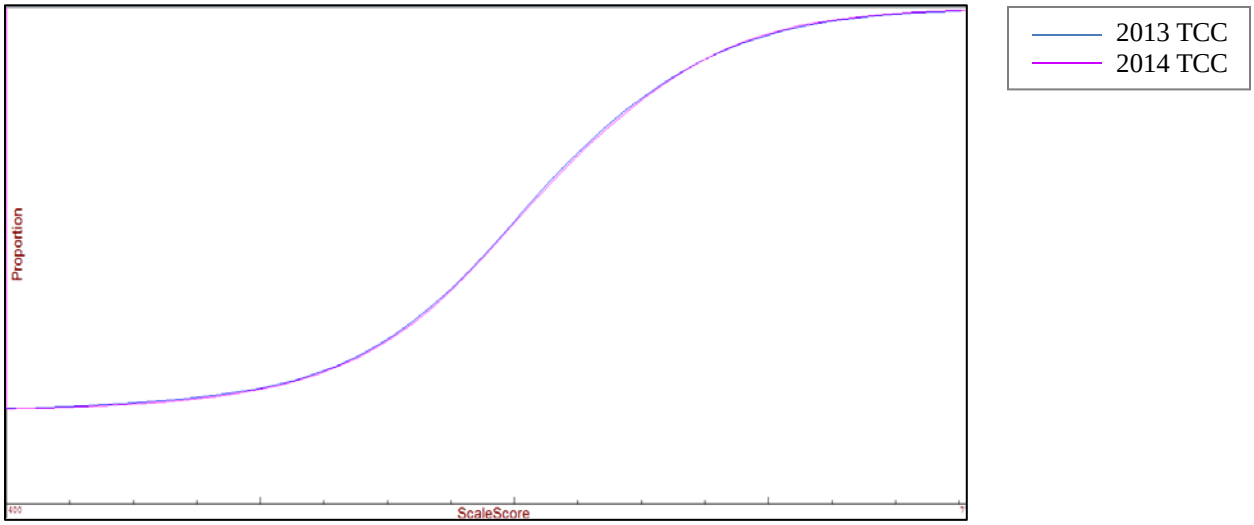
Figure 7.1: Science, Grade 8, Item #34, CR Item (3 Levels)



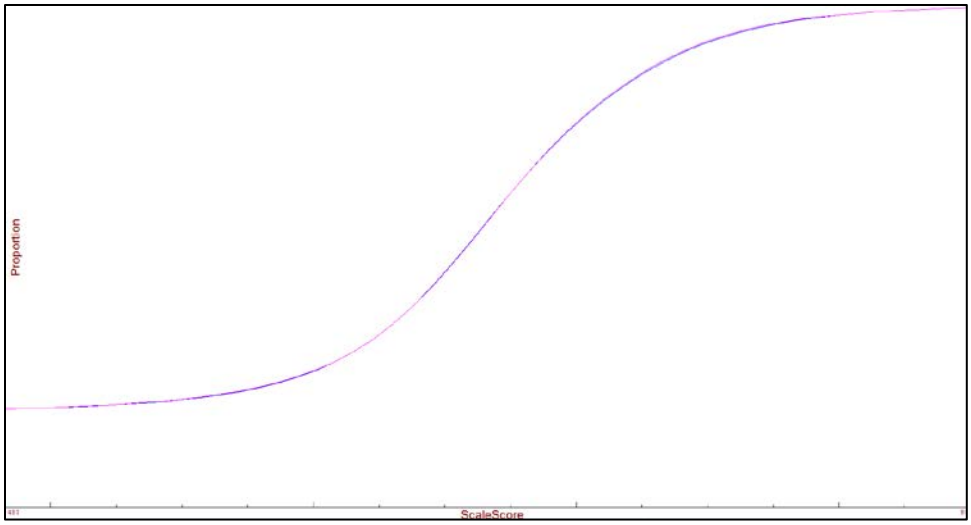
Copyright © 2014 by North Dakota Department of Public Instruction

Figure 7.2: Comparison of Test Characteristics Curves (TCCs) Between 2013 and 2014 Science NDSAs

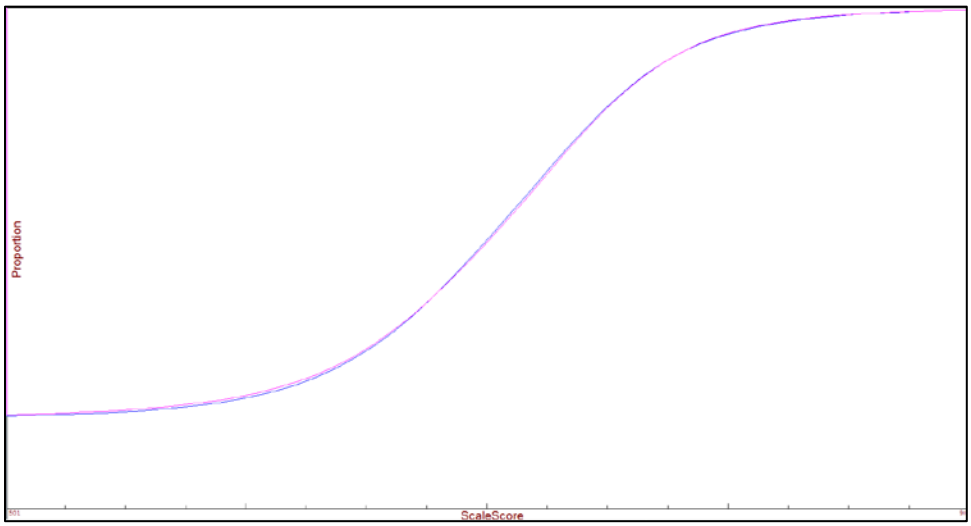
Grade 4



Grade 8



Grade 11



Copyright © 2014 by North Dakota Department of Public Instruction

Part 8: Fairness

Minimizing Bias through Careful Test Development

The development of a test that is fair for all examinees begins in the early stages of planning and development. The item and test development processes that were used to minimize bias are summarized below.

First, careful attention was paid to content validity during the item development and item selection process. Bias can occur if the test is measuring different things for different groups. By eliminating irrelevant skills or knowledge from the items, the possibility of bias is reduced.

Second, item writers and test developers followed several published guidelines for reducing or eliminating bias. These included *Guidelines for Bias-Free Publishing* (McGraw-Hill, 1983) and *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993). Test developers reviewed the items and other testing materials with these guidelines in mind. Internal editorial reviews were conducted by at least three different people: a content editor who directly supervised the item writers, a style editor, and a content supervisor. The final test was again reviewed by at least these same people and was also subjected to an independent review by a quality assurance editor.

Third, careful attention was given to item statistics throughout the test development process. As part of the test assembly process, attempts were made to avoid using or reusing items with poor statistical fit or distractors with positive point-biserial correlations because these statistics may be indicators that an item is tapping an ability that is irrelevant to the construct being measured. Differential item functioning (DIF) statistics were also examined during test construction. Items that exhibited significant DIF against females or males were removed from further consideration, unless it was essential to include them in order to meet content specifications.

Additional steps to reduce bias, including the use of bias review committees composed of North Dakota participants, are described in more detail in Part 2 of this report.

Evaluating Bias through Differential Item Functioning Statistics

After administering the test, an empirical approach known as DIF was used to examine the items. DIF statistics indicate the degree to which members of a particular subgroup perform better or worse than expected on each item compared to the reference group. The DIF procedures used and the results of these analyses are detailed in this section.

The position of CTB/McGraw-Hill concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of

Copyright © 2014 by North Dakota Department of Public Instruction

difference should be considered large and to determine how these differences will affect the outcome of a particular test.

Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements in the performances on which the measurement is based. If these tests require that students have cultural-specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. These tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

In order to lessen this bias, CTB/McGraw-Hill strives to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above and in Part 2 of this report, careful attention is given during the test development and test construction processes to lessen the influence of these elements for large numbers of students (including the use of bias review committees). Unfortunately, in some cases these elements may continue to play a substantial role. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted after each operational test administration.

An IRT-based DIF procedure (Linn & Harnisch, 1981) was used to assess DIF for the NDSAs. The Linn-Harnisch method compares the observed performance of the focal group on a particular item with the expected performance of the total group on the same item that is predicted from the IRT parameter estimate. If the focal group's observed performance is statistically better than what is expected on a given item, the item will be flagged for DIF "in favor" of the focal group. If the focal group's observed performance is statistically worse than its predicted performance, the item will be flagged for DIF "against" the focal group.

The criterion used to determine whether the expected-observed difference is statistically large uses two DIF indices: a standardized difference (Z) and an overall difference (D). The performance difference between the focal and total groups needs to be sufficiently large in both indices for an item to be flagged for DIF. For example, even if the statistical measure (Z) indicates that an item has a significant amount of DIF, perhaps due to a large sample size, the item will not be flagged unless the overall expected-observed difference (D) indicates that the difference is more or less consistent throughout a range of abilities.

More specifically, in the Linn-Harnisch DIF procedure for MC items, students in a given gender group are divided into ten roughly equivalent groups (deciles) based on their ability estimates. For each ability decile (g), the difference between the expected and observed proportion correct scores is computed. D is the mean score difference averaged over the ten ability groups. Consider two situations: (a) when the differences are consistently positive for most of the ability groups, the average difference across all of the ability groups will likely be positive and relatively large and (b) when the differences are positive for about half of the ability groups and

Copyright © 2014 by North Dakota Department of Public Instruction

negative for the remaining ability groups, the average difference across all of the ability groups will likely be relatively small. In other words, the *D* (in absolute value) is expected to be above a criterion value when the predicted-observed performance differences are relatively large in the same direction (negative or positive) consistently across the ability groups.

The standardized difference score for ability group *g* is computed as follows:

$$Z_{ig} = \frac{1}{n_g} \sum_{j \in g} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \right],$$

where $U_{ij} = 1$ when person *j* answers item *i* correctly and $U_{ij} = 0$ otherwise. The standardized difference over all the ability groups is

$$Z_{i.} = \frac{\sum_g n_g Z_{ig}}{\sqrt{\sum_g n_g^2}}.$$

A generalization of this procedure was used to measure DIF for CR items. An item is flagged for negative DIF when $D < -0.10$ and $Z < -2.58$ and for positive DIF when $D > 0.10$ and $Z > 2.58$. As an aid for interpretation in the case of a CR item, a *D* of +0.10 means that the group would have a mean score of 2.6 on the item when their expected mean score is 2.5.

Items that exhibit DIF should not be interpreted as biased for or against a certain group. If there is a difference between the two groups, this may be cause to investigate an item further; however, it is not significant justification to consider an item biased.

Table 8.1 shows the total number of items flagged for DIF in each grade for all subgroups (gender, ethnicity, LEP, NSLP, 504 Plan, and IEP).

If an item is flagged for DIF, then this flag will follow the item through future item selections. Flagged items are further evaluated by CTB content editors for evidence of bias. If an item is flagged for DIF but content experts cannot find a content-related reason for that bias, then the item is deemed appropriate for future form selections. The flagged items were reviewed for content by CTB and DPI staff prior to inclusion in future operational assessments and the content was deemed appropriate for all demographic groups of students.

Detailed listings of items flagged for DIF, including grade, focal group, and DIF statistics, are presented in Table 8.2. Only DIF-flagged items where the focal group size was greater than 100 were reported in these tables. Note that a single item can be flagged for multiple groups.

Copyright © 2014 by North Dakota Department of Public Instruction

Table 8.1: Summary of the Number of Items Flagged for DIF in Science NDSA

	Total Number of Flagged Items
Grade	**Science**
4	0
8	9
11	13

Copyright © 2014 by North Dakota Department of Public Instruction

Table 8.2: 2013 NDSA Linn-Harnisch Statistics: Flagged Items, Science

Grade	Item Number	Item Type	Direction	Focal Group	N	Z	D+	D-	Obs.	Pred.	Obs.-Pred.
8	5	MC	Against	Asian	103	-2.67	0.04	-0.14	0.65	0.76	-0.11
8	10	MC	Against	LEP	215	-3.37	0.03	-0.12	0.54	0.65	-0.11
8	19	MC	Against	LEP	215	-3.55	0.07	-0.15	0.28	0.40	-0.11
8	20	MC	Against	Asian	103	-2.69	0.02	-0.18	0.70	0.81	-0.11
8	20	MC	Against	LEP	215	-3.97	0.10	-0.14	0.48	0.60	-0.12
8	29	MC	In Favor	LEP	215	4.61	0.12	0.00	0.88	0.76	0.12
8	31	MC	Against	Asian	103	-2.62	0.12	-0.21	0.45	0.55	-0.10
8	32	MC	Against	LEP	215	-3.31	0.09	-0.16	0.46	0.56	-0.11
8	34	CR	Against	Asian	103	-3.10	0.08	-0.26	2.18	2.37	-0.19
8	55	MC	In Favor	Asian	103	2.99	0.18	-0.04	0.52	0.39	0.13
11	4	MC	Against	LEP	114	-2.83	0.09	-0.16	0.50	0.63	-0.13
11	17	MC	Against	Asian	124	-4.08	0.10	-0.18	0.54	0.70	-0.16
11	17	MC	Against	LEP	114	-3.70	0.00	-0.17	0.31	0.48	-0.17
11	18	MC	Against	IEP	650	-9.04	0.00	-0.12	0.58	0.70	-0.12
11	20	MC	In Favor	Asian	124	2.95	0.13	-0.04	0.81	0.69	0.12
11	29	MC	In Favor	Asian	124	3.75	0.17	-0.03	0.65	0.49	0.15
11	34	CR	In Favor	Asian	124	3.00	0.22	-0.02	1.35	1.22	0.12
11	35	CR	Against	Male	3668	-8.92	0.00	-0.13	2.26	2.39	-0.13
11	35	CR	In Favor	Female	3486	6.90	0.12	-0.02	2.32	2.21	0.11
11	35	CR	Against	American Indian	552	-2.98	0.08	-0.17	1.82	1.93	-0.11
11	35	CR	Against	IEP	650	-4.23	0.01	-0.17	1.63	1.77	-0.14
11	45	MC	In Favor	LEP	114	2.99	0.16	-0.13	0.45	0.32	0.13
11	46	MC	In Favor	Black	218	3.36	0.13	-0.10	0.62	0.51	0.11
11	50	MC	In Favor	LEP	114	3.31	0.14	-0.02	0.34	0.21	0.13
11	52	MC	Against	Female	3486	-14.93	0.00	-0.11	0.40	0.51	-0.11
11	53	MC	Against	Asian	124	-4.02	0.04	-0.23	0.44	0.61	-0.17
11	53	MC	Against	Black	218	-3.85	0.05	-0.14	0.45	0.57	-0.13
11	53	MC	Against	LEP	114	-2.66	0.06	-0.14	0.28	0.40	-0.12
11	60	MC	Against	LEP	114	-2.96	0.08	-0.18	0.24	0.36	-0.13

Copyright © 2014 by North Dakota Department of Public Instruction

Part 9: Summary of Operational Test Results

This section summarizes the test results for the total population of North Dakota students and for selected demographic subgroups. Longitudinal test results are also presented.

Scale Score Distribution Summary

Table 9.1 provides the mean and standard deviation of the scale scores for all students and for students disaggregated by subgroup (gender, ethnicity, LEP, NSLP, 504 Plan, and IEP). The overall mean scale scores increase across the grade levels. In general, males slightly outperformed females in Science. Students of White outperformed students of other ethnicities in all three grades.

Table 9.2 shows the mean and standard deviation of scale scores for all students for the past nine testing administrations (2006 through 2014). As shown in Table 9.2, similar mean scale scores with a slight increase were observed for Science grades 4 and 8, and a slight decrease occurred for Science grade 11 between the 2013 and 2014 test administrations.

Performance Level Distribution Summary

Table 9.3 presents the percentage of students in each performance level for all students, as well as by gender, ethnicity, LEP, NSLP, 504 Plan, and IEP in Science. With the scale score mean pattern, it can be observed that a higher percentage of male students were classified in the performance level *Proficient* and above for Science. Also, more White and Asian students were classified in the performance level *Proficient* and above than any other ethnic groups.

Table 9.4 shows the percentage of students in each performance level as well as the percentage of students classified as performance level *Proficient* and above for the past nine administrations (2006 through 2014) of the Science assessment. The percentage of students classified in the performance level *Proficient* and above showed slight increase or similar results across the three grades between the 2013 and 2014 administrations.

Copyright © 2014 by North Dakota Department of Public Instruction

Table 9.1: Mean and Standard Deviation of Scale Score for All Students, Students Disaggregated by Subgroup, Science

Subgroup	Category		Grade		
			4	8	11
	All	Mean	638.57	692.47	726.78
		SD	31.62	30.90	33.66
Gender	Male	Mean	640.76	694.19	731.15
		SD	32.19	32.28	35.21
	Female	Mean	636.23	690.81	722.25
		SD	30.69	29.27	31.30
Ethnicity	White	Mean	643.57	697.24	731.43
		SD	29.42	27.80	31.53
	Asian	Mean	629.72	683.71	708.15
		SD	42.19	43.53	43.45
	Am Indian	Mean	619.33	670.43	706.57
		SD	29.22	32.97	31.59
	Hispanic	Mean	625.29	679.70	713.27
		SD	29.67	27.80	32.35
	Black	Mean	619.80	674.95	698.08
		SD	34.27	37.68	39.39
LEP	Yes	Mean	596.37	647.20	663.63
		SD	32.82	33.84	31.39
NSLP	Yes	Mean	627.14	678.99	712.70
		SD	31.54	31.76	35.51
504 Plan	Yes	Mean	639.39	694.28	722.44
		SD	30.80	26.05	30.85
IEP	Yes	Mean	621.08	667.50	695.63
		SD	32.91	32.60	34.23

Copyright © 2014 by North Dakota Department of Public Instruction

Table 9.2: Mean and Standard Deviation of Scale Scores for 2006–2014 by Grade—Census (Does Not Include Home-Education Students)

Category		Grade		
		4	8	11
2006	Mean	636.02	694.18	724.97
	SD	30.73	32.26	32.91
2007	Mean	636.97	693.10	722.71
	SD	31.04	32.63	36.08
2008	Mean	638.73	695.43	723.24
	SD	29.74	30.77	34.75
2009	Mean	644.48	694.50	725.06
	SD	30.75	31.19	34.12
2010	Mean	638.02	695.67	725.75
	SD	30.63	33.85	33.77
2011	Mean	635.14	693.73	726.16
	SD	29.99	31.72	35.61
2012	Mean	639.55	692.31	729.22
	SD	30.31	32.03	35.45
2013	Mean	637.92	692.62	727.05
	SD	32.92	31.97	35.49
2014	Mean	638.57	692.47	726.78
	SD	31.62	30.9	33.66

Copyright © 2014 by North Dakota Department of Public Instruction

Table 9.3: Percent of Students in Each Performance Level for All Students, Students Disaggregated by Subgroup, Science

Grade	Subgroup	Category	N	Performance Levels				
				Novice	Partially Proficient	Proficient	Advanced	Proficient & Above
4		All	8553	9.52	22.81	49.33	18.34	67.67
	Gender	Male	4492	8.88	21.48	48.42	21.22	69.64
		Female	4047	10.16	24.29	50.33	15.22	65.55
	Ethnicity	White	6170	5.66	20.06	52.87	21.41	74.28
		Asian	116	18.97	25.00	37.07	18.97	56.04
		Am Indian	891	24.80	33.45	35.47	6.29	41.76
		Hispanic	323	17.34	32.82	43.34	6.50	49.84
		Black	317	21.77	35.33	35.65	7.26	42.91
	LEP	Yes	202	49.50	32.18	16.83	1.49	18.32
	NSLP	Yes	2689	17.11	30.27	42.17	10.45	52.62
	504 Plan	Yes	145	5.52	26.90	50.34	17.24	67.58
	IEP	Yes	991	23.71	34.51	32.59	9.18	41.77
8		All	7928	9.84	30.34	44.40	15.43	59.83
	Gender	Male	4040	9.68	28.27	44.11	17.95	62.06
		Female	3866	9.88	32.41	44.83	12.88	57.71
	Ethnicity	White	5902	5.51	28.01	48.98	17.50	66.48
		Asian	102	23.53	25.49	34.31	16.67	50.98
		Am Indian	758	28.89	42.88	23.22	5.01	28.23
		Hispanic	284	17.25	44.37	30.63	7.75	38.38
		Black	260	26.54	33.08	33.08	7.31	40.39
	LEP	Yes	215	57.21	36.28	6.51	0.00	6.51
	NSLP	Yes	2327	20.24	39.28	32.92	7.56	40.48
	504 Plan	Yes	185	6.49	33.51	44.86	15.14	60.00
	IEP	Yes	923	31.31	43.99	20.59	4.12	24.71
11		All	7460	11.13	25.39	47.94	15.55	63.49
	Gender	Male	3824	9.94	21.42	49.08	19.56	68.64
		Female	3623	12.37	29.48	46.78	11.37	58.15
	Ethnicity	White	5662	7.72	22.82	51.93	17.54	69.47
		Asian	117	36.75	17.09	32.48	13.68	46.16
		Am Indian	498	24.50	37.95	33.13	4.42	37.55
		Hispanic	172	19.19	38.95	33.72	8.14	41.86
		Black	202	30.20	38.12	27.72	3.96	31.68
	LEP	Yes	114	78.95	16.67	4.39	0.00	4.39
	NSLP	Yes	1513	22.14	31.39	37.61	8.86	46.47
	504 Plan	Yes	185	12.97	30.27	43.78	12.97	56.75
	IEP	Yes	653	38.90	34.61	24.50	1.99	26.49

Copyright © 2014 by North Dakota Department of Public Instruction

Table 9.4: Percent of Students in Each Performance Level for 2006–2014 by Grade: Science—State

Grade	Year	N Count	Performance Levels				
			Novice	Partially Proficient	Proficient	Advanced	Proficient & Above
4	2006	7392	11.15	25.70	46.05	17.10	63.15
	2007	7332	9.71	25.63	45.54	19.12	64.66
	2008	7131	8.11	22.66	49.53	19.70	69.23
	2009	6922	6.20	19.08	47.02	27.69	74.72
	2010	7372	9.50	22.54	49.99	17.97	67.96
	2011	7504	11.35	24.72	48.29	15.63	63.92
	2012	7841	9.85	22.20	46.28	21.67	67.95
	2013	8314	11.03	22.00	48.62	18.35	66.97
	2014	8553	9.52	22.81	49.33	18.34	67.67
8	2006	8158	10.44	28.92	45.29	15.35	60.64
	2007	7726	9.82	27.23	47.20	15.74	62.94
	2008	7540	8.34	29.14	45.50	17.02	62.52
	2009	7459	9.75	28.46	45.10	16.69	61.79
	2010	7445	10.57	25.12	47.70	16.62	64.32
	2011	7567	9.61	28.40	47.03	14.96	61.99
	2012	7638	11.29	29.55	43.49	15.67	59.16
	2013	7504	10.14	29.94	43.99	15.92	59.91
	2014	7928	9.84	30.34	44.40	15.43	59.83
11	2006	7719	11.48	26.64	45.55	16.34	61.89
	2007	7599	13.83	27.15	42.81	16.21	59.02
	2008	7426	14.06	25.15	45.02	15.77	60.79
	2009	7410	12.08	28.62	43.43	15.87	59.30
	2010	7224	11.49	26.15	45.16	17.21	62.37
	2011	7209	12.03	26.16	44.61	17.20	61.81
	2012	7317	10.80	24.59	44.85	19.76	64.61
	2013	7271	11.70	22.91	48.93	16.45	65.38
	2014	7460	11.13	25.39	47.94	15.55	63.49

Copyright © 2014 by North Dakota Department of Public Instruction

References

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.

Angoff, W.H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins University Press.

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

Bock, R., Muraki., E., & Pfeiffenberger, W. (1998). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

Burket, G. R. (1995). PARDUX [Computer program]. Monterey, CA: CTB/McGraw-Hill.

Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items.* Sage Publications.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253–260.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

CTB/McGraw-Hill (1996). *TerraNova* Assessment Series (1st Ed.). Monterey, CA: Author.

CTB/McGraw-Hill (2000). *TerraNova* Assessment Series (2nd Ed.). Monterey, CA: Author.

CTB/McGraw-Hill (2006). *TerraNova* Assessment Series (3rd Ed.). Monterey, CA: Author.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations, *Applied Measurement in Education*, *17*(3), 265–300.

Copyright © 2014 by North Dakota Department of Public Instruction

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Glas, G. A. W. (2000). Item calibration and parameter drift. In W.J. van der Linden & C.AW. Glas (eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 83–199). Netherlands: Kluwer.

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

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Li, X. & Reckase, M. (2009). *Multidimensionality and item parameter drift: An investigation of linking items in a large-scale certification test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement 32*, 179–197.

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

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

Macmillan/McGraw-Hill. (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals.* New York: Author.

McGraw-Hill. (1983). *Guidelines for bias-free publishing.* New York: Author.

Miller, G.E., & Fitzpatrick, S. J. (2009*).* Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educational and Psychological Measurement, 69*(3), 357–368.

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

Patz, R. J., & Hanson, B. (2002). *Psychometric issues in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New

Copyright © 2014 by North Dakota Department of Public Instruction

Orleans, LA.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics 4*, 207–230.

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

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

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

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

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

Copyright © 2014 by North Dakota Department of Public Instruction

Appendix A: North Dakota 2014 Scoring Tables

Tables 1 through 3 in Appendix A contain the North Dakota 2014 raw score to scale score to standard error of measurement scoring tables for Science.

Copyright © 2014 by North Dakota Department of Public Instruction

Table 1. Science, Grade 4 Scoring Table

Raw Score	Scale Score	Standard Error of Measurement	Raw Score	Scale Score	Standard Error of Measurement
0	400	154	46	647	12
1	400	154	47	652	12
2	400	154	48	657	13
3	400	154	49	662	13
4	400	154	50	668	14
5	400	154	51	675	15
6	400	154	52	682	16
7	400	154	53	691	17
8	400	154	54	702	20
9	400	154	55	716	25
10	400	154	56	741	36
11	402	152	57	780	65
12	468	86			
13	494	59			
14	511	43			
15	522	33			
16	531	27			
17	539	23			
18	545	20			
19	550	18			
20	556	17			
21	560	16			
22	565	15			
23	569	14			
24	572	13			
25	576	13			
26	580	12			
27	583	12			
28	586	11			
29	589	11			
30	592	11			
31	596	11			
32	599	10			
33	602	10			
34	605	10			
35	608	10			
36	611	10			
37	614	10			
38	617	10			
39	620	11			
40	624	11			
41	627	11			
42	631	11			
43	635	11			
44	639	12			
45	643	12			

Copyright © 2014 by North Dakota Department of Public Instruction

Table 2. Science, Grade 8 Scoring Table

Raw Score	Scale Score	Standard Error of Measurement	Raw Score	Scale Score	Standard Error of Measurement
0	483	143	46	710	11
1	483	143	47	714	12
2	483	143	48	719	12
3	483	143	49	724	13
4	483	143	50	730	14
5	483	143	51	736	15
6	483	143	52	744	16
7	483	143	53	753	18
8	483	143	54	764	21
9	483	143	55	779	26
10	483	143	56	806	39
11	483	143	57	850	75
12	542	84			
13	569	57			
14	585	41			
15	596	31			
16	605	25			
17	611	21			
18	617	18			
19	622	16			
20	627	15			
21	631	14			
22	635	13			
23	638	12			
24	642	11			
25	645	11			
26	648	11			
27	651	10			
28	654	10			
29	657	10			
30	659	10			
31	662	9			
32	665	9			
33	668	9			
34	670	9			
35	673	9			
36	676	9			
37	679	9			
38	682	9			
39	685	10			
40	688	10			
41	691	10			
42	695	10			
43	698	10			
44	702	11			
45	706	11			

Copyright © 2014 by North Dakota Department of Public Instruction

Table 3. Science, Grade 11 Scoring Table

Raw Score	Scale Score	Standard Error of Measurement	Raw Score	Scale Score	Standard Error of Measurement
0	501	153	46	748	11
1	501	153	47	751	11
2	501	153	48	755	11
3	501	153	49	760	12
4	501	153	50	764	12
5	501	153	51	769	13
6	501	153	52	775	14
7	501	153	53	782	15
8	501	153	54	790	17
9	501	153	55	800	20
10	501	153	56	816	26
11	535	119	57	842	39
12	582	72	58	900	86
13	603	51			
14	617	37			
15	627	29			
16	636	25			
17	643	21			
18	649	19			
19	654	17			
20	659	16			
21	663	15			
22	668	14			
23	672	14			
24	676	13			
25	679	13			
26	683	12			
27	686	12			
28	690	12			
29	693	12			
30	696	11			
31	700	11			
32	703	11			
33	706	11			
34	709	11			
35	712	11			
36	715	10			
37	719	10			
38	722	10			
39	725	10			
40	728	10			
41	731	10			
42	734	10			
43	737	10			
44	741	10			
45	744	10			

Copyright © 2014 by North Dakota Department of Public Instruction