South Dakota State Test of Educational Progress

Dakota STEP

Technical Report: 2008 Spring Administration

Pearson

San Antonio, TX

Copyright © 2008 by NCS Pearson, Inc. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the copyright owner. *Pearson* and the *Pearson* Logo are trademarks, in the U.S. and/or other countries, of Pearson Education Inc. or its affiliates. Portions of this work were previously published. Printed in the United States of America.

Table of Contents

CHAPTER 1: BACKGROUND 5
1.1 Statewide Testing and Accountability 5
1.2 South Dakota Content Standards 6
1.3 South Dakota Achievement Descriptors 8
1.4 Organizations and Groups Involved 9
1.5 Plans for Future Assessments 10

CHAPTER 2: ITEM DEVELOPMENT AND TEST CONSTRUCTION 11
2.1 Test Design 11
2.2 Transition to Embedded Field-Testing 14
2.3 Test Construction Specifications 14
2.4 Item Specifications 16
2.5 Passage Specifications 17
2.6 Item Development 21
2.7 South Dakota Teacher Committees 22

CHAPTER 3: TEST ADMINISTRATION 27
3.1 Test Security Guidelines 27
3.2 Test Administration 28
3.3 Assessment of Students with Disabilities 29
3.4 Test Accommodations 30
3.5 Assessment of LEP Students 32

CHAPTER 4: ACHIEVEMENT LEVELS 34
4.1 Achievement Level Descriptors 34
4.2 Achievement Level Cut Points 34
4.3 Achievement Level Raw Scores (Scale Scores) 34

CHAPTER 5: REPORTS 37
5.1 Scores Reported 37
5.2 Appropriate Uses of Scores 38

CHAPTER 6: CALIBRATION, LINKING (EQUATING), AND SCALING 39
6.1 Calibration 39
6.2 Linking Procedure 40
6.3 Scale Score Creation 46

CHAPTER 7: PSYCHOMETRIC CHARACTERISTICS: CLASSIC ITEM STATISTICS, RELIABILITY, AND DIF 49
7.1 Reliability and Standard Error of Measurement of Raw Scores 54
7.2 Consistency and Accuracy of Achievement Level 55

CHAPTER 8: PSYCHOMETRIC CHARACTERISTIC: VALIDITY 58
8.1 Content Related Evidence of Validity 58

8.2 Construct Related Evidence of Validity 61

CHAPTER 9: QUALITY CONTROL PROCEDURES..... 69
9.1 Quality Control for Item Development and Test Construction 69
9.2 Quality Control for the Receipt of Non-Scannable Documents 69
9.3 Quality Control for Data Preparation..... 69
9.4 Quality Control in Document Tracking 70
9.5 Quality Control in Editing and Data Input..... 70
9.6 Quality Control in Scoring and Reporting..... 71

GLOSSARY OF TERMS 73

REFERENCES 78

APPENDIX A: Test Blueprints 80
APPENDIX B: Test Security Documents..... 103
APPENDIX C: Sample Score Reports 107
APPENDIX D: Rasch (IRT) Item Statistics 110
APPENDIX E: Test Characteristic, Test Information, and
Conditional Standard Error Curves..... 144
APPENDIX F: Raw Score to Scale Score Conversion Tables 162
APPENDIX G: Raw Score to Scale Score Frequency Tables..... 196
APPENDIX H: Scaled Rasch Item Difficulty Parameter Summary 230
APPENDIX I: Item Statistics..... 232
APPENDIX J: Differential Item Functioning (DIF) Classification..... 247
APPENDIX K: Summary of Differential Item Functioning (DIF) Results..... 292
APPENDIX L: Subgroup Raw Score Analysis Summary..... 296
APPENDIX M: Reporting Category Raw Score Analysis Summary 307
APPENDIX N: Alignment Analysis Summary 314
APPENDIX O: Reading Alignment Report 329
APPENDIX P: Mathematics Alignment Report..... 509
APPENDIX Q: Science Alignment Report 756

Tables and Figures

Table 2.1: Dakota STEP Subjects and Grades Tested 11
Table 2.2: Grade 3 Reading Test Blueprint 13
Table 2.3: Dakota STEP Item and Test Form Counts by Content Domain 14
Table 2.4: Approximate Percentages for Passage Type 17
Table 2.5: Passage Word Count Ranges 17
Table 2.6: 2008 Content/Bias Review Results 23
Table 2.7: 2008 Data Review Results 25
Table 3.1: Key Events and Dates for the 2008 Dakota STEP Administration 27
Table 3.2: Approximate Testing Times for the Dakota STEP 29
Table 3.3: Dakota STEP Assessment Accommodations 31
Table 4.1: Raw Score (Scale Score) Ranges for Dakota STEP Reading 35
Table 4.2: Raw Score (Scale Score) Ranges for Dakota STEP Mathematics 36
Table 4.3: Raw Score (Scale Score) Ranges for Dakota STEP Science 36
Table 6.1: Mathematics Linking Item Statistics from 2007 Administration 43
Table 6.2: Science Linking Item Statistics from 2007 Administration 44
Table 6.3: Mathematics Linking Item Statistics from 2008 Administration 45
Table 6.4: Science Linking Item Statistics from 2008 Administration 46
Table 6.5: Scale Score Summary for Dakota STEP Reading 47
Table 6.6: Scale Score Summary for Dakota STEP Mathematics 47
Table 6.7: Scale Score Summary for Dakota STEP Science 47
Table 6.8: SE of SS Cut Points by Achievement Level 48
Table 7.1: Reading, Mathematics, and Science Item P-Value Summary 50
Table 7.2: Reading, Mathematics, and Science Item P-Value Classification 51
Table 7.3: Reading, Mathematics, and Science Item Point-Biserial Summary 52
Table 7.4: Reading, Mathematics, and Science Item Point-Biserial Classification 53
Table 7.5: Reading Assessment Reliability and CTT SEM 56
Table 7.6: Mathematics Assessment Reliability and CTT SEM 56
Table 7.7: Science Assessment Reliability and CTT SEM 56
Table 7.8: Percentage of Students at Each Achievement Level 57
Table 7.9: Estimated Consistency and Accuracy for Proficient Decisions (Percentage) 57
Table 8.1: Correlation Among Dakota STEP Assessments 62
Table 8.2: First Ten Eigenvalues Obtained from the PCA 63
Table 8.3: Variance Explained by the First Three Principal Components 65

Figure 8.1: Reading Assessment PCA Scree Plot 66
Figure 8.2: Mathematics Assessment PCA Scree Plot 67
Figure 8.3: Science Assessment PCA Scree Plot 68

CHAPTER 1: BACKGROUND

This technical report provides information on the development, administration, scoring, and reporting of the 2008 South Dakota State Test of Educational Progress (*Dakota STEP*) as well as psychometric data used to score the test and ensure its quality. It is intended for those responsible for reviewing the effectiveness of the *Dakota STEP* as well as anyone seeking a deeper understanding of the methods used to develop the test and its results.

The chapters in this report cover the major processes and methods used to develop, administer, score, and report the results of the *Dakota STEP*. Data from the test results and its items, including indicators used for quality control, are presented separately in the appendices. This report also includes a glossary of key terms.

1.1 Statewide Testing and Accountability

The *Dakota STEP* is South Dakota's annual statewide assessment of student progress administered to students in grades 3 through 8 and 11 each spring. The *Dakota STEP* fulfills the requirements for the statewide assessment contained in the federal *No Child Left Behind Act of 2001* (NCLB, 2002), the latest reauthorization of the *Elementary and Secondary Education Act of 1965* (ESEA). NCLB requires each state to adopt challenging academic content and student academic achievement standards for all public-school students and to implement a set of high-quality assessments to measure *adequate yearly progress* (AYP) toward meeting these standards. Moreover, the assessments must report student results using at least three *performance* or *achievement levels*: basic, proficient, and advanced (NCLB, §6311).

All South Dakota public-school students in grades 3 through 8 and 11 are required to take the *Dakota STEP* assessment. Private and alternative-site schools that are classified as accredited or approved by the South Dakota Department of Education (SD DOE) also administer the *Dakota STEP*. Students who are state placed or district placed must be included in the State of South Dakota's Accountability Plan for *No Child Left Behind*. Bureau of Indian Education (BIE) schools participate in the *Dakota STEP* and are expected to follow the same guidelines as public schools.

The annual administration of the *Dakota STEP* provides critical feedback to educators, students, and parents regarding students' academic achievement and mastery of South Dakota's Content Standards. The state is required to notify the public of the performance of its schools and school districts utilizing the NCLB Report Card which evaluates attendance, graduation rates, and student academic achievement. The NCLB Report Card reports the percentage of students scoring at each achievement level for each school, district, and state. Schools and districts that fail to meet AYP for a single year are put on the *alert status*. If schools and districts fail to meet AYP for two consecutive years, those schools and districts are identified for school improvement and sanctions required under NCLB are applied. The *Dakota STEP* is not "high stakes" for students in that the results are not used in making grade promotion decisions or high school graduation decisions.

1.2 South Dakota Content Standards

In 1997, the South Dakota State Legislature passed SB170 that amended South Dakota Codified Law 13-3-48 to address the issue of having challenging state content standards. The adopted amendment reads as follows: "The Secretary of the Department of Education and Cultural Affairs shall prepare and submit for approval of the South Dakota Board of Education academic content standards in Language Arts, Mathematics, Social Studies, and Science for kindergarten through grade twelve. Each school district shall adopt and implement clearly defined and measurable course guidelines so as to meet the state academic content standards."

The South Dakota Content Standards provide a common set of goals and expectations for all students in all schools and are designed to guide the planning of curriculum and to anchor the assessment of learning from kindergarten through twelfth grade. As students move from kindergarten through grade 12, levels of cognitive demand and complexity of content, skills, and processes increase. New skills emerge, and basic skills are subsumed within more advanced skills as students progress through the grades. Grade-level standards specify what students should know and have mastered at the proficient level by the time they are assessed for these skills at the end of each grade level.

Dakota STEP test content is specified by the South Dakota Academic Content Standards. The following content standards documents were used to develop items for the operational 2008 assessment and are available online at http://doe.sd.gov/contentstandards.

South Dakota Reading Content Standards, approved June 2004
South Dakota Mathematics Content Standards, approved June 2004
South Dakota Science Content Standards, approved June 2005

Content standards for each domain are hierarchically organized by content strands, indicators, and grade-specific standards, and achievement descriptors that define the standards-based objectives for South Dakota students. Specifically:

- **Goals, strands, and indicators** represent expected outcomes for all students graduating from South Dakota schools.
- **Grade-specific content standards** represent expected outcomes for students at each grade level.

Reading

The Reading Content Standards consist of grade-specific standards within each of the following indicators. The Reading Content Standards can be found at http://doe.sd.gov/contentstandards/languagearts/reading/old.asp.

Reading Goal: *Students are able read at increasing levels of complexity for a variety of reasons.*

Indicator 1: *Students are able to apply various reading strategies to comprehend and interpret text.*
Indicator 2: *Students are able to evaluate text structures, literary elements, and literary devices within various genres to develop interpretations and form responses.*
Indicator 3: *Students are able to interpret and respond to diverse works from various cultures and time periods.*

Indicator 4: *Students are able to retrieve, analyze, synthesize, and evaluate a variety of informational texts.*

Mathematics

The Mathematics Content Standards consist of grade-specific standards within Algebra, Geometry, Measurement, Number Sense, and Statistics and Probability that represent expected outcomes for all students preparing to graduate from South Dakota schools. The Mathematics Content Standards can be found at http://doe.sd.gov/contentstandards/math.

Algebra (Strand 1): *Students will use the language of algebra to explore, describe, represent, and analyze number expressions and relations that represent variable quantities*.

Indicator 1: *Use procedures to transform algebraic expressions.*
Indicator 2: *Use a variety of algebraic concepts and methods to solve equations and inequalities.*
Indicator 3: *Interpret and develop mathematical models.*
Indicator 4: *Describe and use properties and behaviors of relations, functions, and inverses.*

Geometry (Strand 2): *Students will use the language of geometry to discover, analyze, and communicate geometric concepts, properties, and relationships*.

Indicator 1: *Use deductive and inductive reasoning to recognize and apply properties of geometric figures.*
Indicator 2: *Use properties of geometric figures to solve problems from a variety of perspectives.*

Measurement (Strand 3): *Students will apply systems of measurement and use appropriate measurement tools to describe and analyze the world around them*.

Indicator 1: *Apply measurement concepts in practical applications.*

Number Sense (Strand 4): *Students will develop and use number sense to investigate the characteristics of numbers in a variety of forms and modes of operation*.

Indicator 1: *Analyze the structural characteristics of the real number system and its various subsystems. Analyze the concepts of value, magnitude, and relative magnitude of real numbers.*
Indicator 2: *Apply number operations with real numbers and other number systems.*
Indicator 3: *Develop conjectures, predictions, or estimations to solve problems and verify or justify the results.*

Statistics and Probability (Strand 5): *Students will apply statistical methods to analyze data and explore probability for making decisions and predictions*.

Indicator 1: *Use statistical models to gather, analyze, and display data to draw conclusions.*
Indicator 2: *Apply the concepts of probability to predict events/outcomes and solve problems.*

Science

The Science Content Standards consist of grade-specific standards within Nature of Science, Physical, Life, and Earth/Space Science, and Science, Technology, Environment, and Society that represent expected outcomes for all students preparing to graduate from South Dakota schools. The Science Content Standards can be found at http://doe.sd.gov/contentstandards/science.

Nature of Science (Strand 1): *Students will explore, evaluate, and communicate personal and scientific investigations to understand the nature of science* (grades 8 and 11)**.**

Indicator 1: *Understand the nature and origin of scientific knowledge.*
Indicator 2: *Apply the skills necessary to conduct scientific investigations.*

Physical Science (Strand 2): *Students will use appropriate scientific models to describe and quantify the nature and interactions of matter and energy* (grades 5, 8, and 11)**.**

Indicator 1: *Describe structures and properties of and changes in matter.*
Indicator 2: *Analyze forces, their forms, and their effects on motions.*
Indicator 3: *Analyze interactions of energy and matter.*

Life Science (Strand 3): *Students will describe structures and attributes of living things, processes of life, and interaction with each other and the environment* (grades 5 and 11)**.**

Indicator 1: *Understand the fundamental structures, functions, classifications, and mechanisms found in living things.*
Indicator 2: *Analyze various patterns and products of natural and induced biological change.*
Indicator 3: *Analyze how organisms are linked to one another and the environment.*

Earth/Space Science (Strand 4): *Students will analyze the composition, formative processes, and history of the universe, solar system, and Earth* (grades 5, 8, and 11)**.**

Indicator 1: *Analyze the various structures and processes of the Earth system.*
Indicator 2: *Analyze essential principles and ideas about the composition and structure of the universe.*

Science, Technology, Environment, and Society (Strand 5): *Students will identify and evaluate the relationships and ethical implications of science upon technology, environment, and society* (grades 5, 8, and 11)**.**

Indicator 1: *Analyze various implications/effects of scientific advancement within the environment and society*.
Indicator 2: *Analyze the relationships/interactions among science, technology, environment, and society*.

1.3 South Dakota Achievement Descriptors

Achievement (performance) descriptors provide information to all stakeholders regarding student progress toward mastery of the content standards. Achievement descriptors within the South Dakota Content Standards provide descriptions of three achievement levels: *Basic*, *Proficient*, and *Advanced*. The general achievement descriptors describe how a student at each achievement level would be expected to perform the grade-specific standards.

Advanced: *A student performing at the advanced level exceeds expectations for that grade level. The student is able to perform the content standards for the grade at a high level of difficulty, complexity, or fluency.*

Proficient: *A student performing at the proficient level meets expectations for that grade level. The student is able to perform the content standards for the grade at the level of difficulty, complexity, or fluency specified by the standards.*

Basic: *A student performing at the basic level performs below expectations for that grade level. The student is able to perform some of the content standards for the grade below the level of difficulty, complexity, or fluency specified by the grade-level standards.*

A student performing below the basic level is unable to perform the content standards, therefore; no description is provided for this achievement level. Grade-specific achievement descriptors are documented within the South Dakota Content Standards and are available online at http://doe.sd.gov/contentstandards.

1.4 Organizations and Groups Involved

The South Dakota Department of Education Office of Curriculum, Technology, and Assessment oversees the development, administration, and reporting of the *Dakota STEP* utilizing the experience, expertise, and capacity of the following external contractors.

Pearson supports the SD DOE by developing standards-based items for the *Dakota STEP*. In this capacity, Pearson conducts activities related to item development and convenes item bias and content review sessions with South Dakota teachers. It publishes and distributes the *Dakota STEP* integrated test materials and provides support in training school and district staff to administer the test. Pearson scores the tests and provides the state with results in the form of a scored data file.

The Buros Institute for Assessment Consultation and Outreach (BIACO) of the University of Nebraska conducts workshops to align the pool of items developed for the *Dakota STEP* with the South Dakota Content Standards for all content areas–Reading (grades 3–8 and 11), Mathematics (grades 3–8 and 11), and Science (grades 5, 8, and 11). BIACO also conducts standard setting workshops for the *Dakota STEP* following the operational administration of a new subtest assessing revised content standards.

Norman L. Webb conducts alignment studies of the *Dakota STEP* operational assessments (published forms) with the South Dakota Content Standards for all content areas–Reading (grades 3–8 and 11), Mathematics (grades 3–8 and 11), and Science (grades 5, 8, and 11).

South Dakota educators are involved throughout the item development process participating in content and bias review, alignment workshops, data review as well as standard-setting panels. Many district and school personnel serve as test coordinators and/or building coordinators. The educators and other personnel provide feedback, both formal and informal, to the SD DOE regarding all aspects of test administration.

1.5 Plans for Future Assessments

Revised South Dakota Content Standards for Reading were approved by the Board of Education (BOE) in early spring 2007, aligned with curriculum in the 2007/2008 school year, and applied in curriculum in the 2008/2009 school year. Item development to assess the new standards began in early spring of 2007 following an alignment analysis of the existing items with the new content standards (BIACO, Dec. 2007). Newly developed items were field tested in spring 2008 and the *Dakota STEP* assessment in spring 2009 will fully assess the depth and breadth of the revised Reading Content Standards.

Revised South Dakota Content Standards for Mathematics, planned for BOE approval in early 2011, will be aligned with curriculum in the 2011/2012 school year, and applied in curriculum in the 2012/2013 school year. Item development to assess the new mathematics standards will begin in early spring of 2011 following an alignment analysis of the existing items with the new content standards. Newly developed items will be field tested in spring 2012 and the *Dakota STEP* assessment in spring 2013 will fully assess the depth and breadth of the revised Mathematics Content Standards.

Revised South Dakota Content Standards for Science, planned for BOE approval in early spring 2012, will be aligned with curriculum in the 2012/2013 school year, and applied in curriculum in the 2013/2014 school year. Item development to assess the new standards will begin in early spring of 2012 following an alignment analysis of the existing items with the new content standards. Newly developed items will be field tested in spring 2013 and the *Dakota STEP* assessment in spring 2014 will fully assess the depth and breadth of the revised Science Content Standards.

CHAPTER 2: ITEM DEVELOPMENT AND TEST CONSTRUCTION

This chapter describes the process of developing the standards-based items and tests for the *Dakota STEP* assessment. Virtually all steps in the development process contribute toward the goal of creating a test that shows both *reliability* and *validity* evidence.

> **Reliability** refers to *the consistency of test results* across multiple administrations of the same or alternate forms of the test. Simply, if an examinee were to take an exam multiple times, how consistent would his or her scores be? Reliability is inversely related to measurement error—as error increases, reliability decreases. Measurement error is due to such factors as ". . . guessing, distractions in the testing situation, administration errors, content sampling, scoring errors, and fluctuations in the individual examinee's state" (Crocker & Algina, 1986, p. 106).
>
> **Validity** is *the degree of the adequacy and appropriateness of the interpretations made from assessment results with respect to a particular use.* It is concerned with the general question, "To what extent will this assessment's information or test score help me make appropriate decisions?"

The alignment of assessment to curriculum enhances consistency and strengthens the influence of academic content standards and assessment on improving teaching and learning.

2.1 Test Design

The *Dakota STEP* test assesses students' achievement of the South Dakota Content Standards. The following Content Standards documents used for the development of the *Dakota STEP* assessment are available online at http://doe.sd.gov/contentstandards/:

- *South Dakota Reading Content Standards*, approved June 2004, statewide assessments operational in spring 2005
- *South Dakota Mathematics Content Standards*, approved June 2004, statewide assessments operational in spring 2006
- *South Dakota Science Content Standards*, approved June 2005, statewide assessments operational in spring 2007

The assessment is administered in combined-content test booklets consisting of grade-specific subtests as shown in Table 2.1.

Table 2.1: Dakota STEP Subjects and Grades Tested

Test	Grade						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Reading	✓	✓	✓	✓	✓	✓	✓
Mathematics	✓	✓	✓	✓	✓	✓	✓
Science	—	—	✓	—	—	✓	✓

The *Dakota STEP* assessment is composed of multiple-choice core (operational) items for each content domain and grade. All students are assessed with the same core items for each content

domain. All multiple-choice items are worth one raw score point and are the basis of student scores. Linking items (anchor items) are core items used to link the current assessment to the previous year's score scale, and are included in the count of core items.

Newly developed items are embedded within the core items to create unique test forms for each content domain and grade level. These field-test items do not count toward students' scores but are evaluated for statistical quality for potential use as core items in future test administrations by South Dakota educators and the SD DOE. Although the number of field-test items is the same across all grade levels within subjects, the distribution of items across standards will differ from form to form.

Universal Test Design

Universal Test Design is an approach to item development and test design that calls for minimizing the need for accommodations in test administration by making items accessible to the greatest number of students, including those with disabilities and non-native speakers of English. In applying these principles, item writers attempt to maximize readability, legibility, and compatibility with accommodations by avoiding wordiness and ambiguity, selecting reader-friendly constructions and terminology, and applying concept names and graphic conventions consistently. Universal test design principles are also used to make decisions about test layout and design, including, but not limited to, type size, line length, spacing, and graphics. All items, whether core (operational) or field test, are intended to have a uniform appearance and are developed and composed using the same style guidelines for items in all subjects.

Test Blueprints

The first step in the creation (or revision) of a standards-based assessment is the development of a *test blueprint*, which specifies the content standards to be assessed and the number of items to assess each standard. The underlying principle guiding the creation of test blueprints is establishing a clear relationship between the assessment and the standards. The blueprint then serves as the foundation for the entire item development and test construction process.

The number of items for each reported category should be consistent across content areas and grade levels in order to establish a rationale and coherent assessment system. The reported category can be a content standard, indicator, or goal while the number of items contributing to a student's score influences the reliability of the score. All things being equal, the fewer the number of items, the lower the reliability and stability of that score.

Test blueprints, which specify the content standards to be assessed and the number of items needed to assess each standard, were developed and approved by the SD DOE for each content domain and grade level. To provide reliability evidence of scores, a sufficient number of items addressing each content standard or indicator for which a score is to be reported ("reporting category") must be maintained. Results for reading are reported at the standard level while results for mathematics and science content domains are reported at the indicator level. The item counts within the test blueprints represent the number of core (operational) items on each subtest. The SD DOE requires a minimum of 7 items per reporting category as illustrated in the grade 3 Reading test blueprint (Table 2.2). Test blueprints for all content domains and grades are included in Appendix A.

In Table 2.2, the grade 3 Reading blueprint shows the target item counts for each of the content standards. In addition, targets have been established to include a range of items at various cognitive complexity levels with the goal of at least 50% or more of the items assessing each content standard to be at or above the cognitive complexity of that standard.

Table 2.2: Grade 3 Reading Test Blueprint

Content Standards	Bloom's Taxonomy Designation	TEST BLUEPRINT			
		Basic	**Proficient**	**Advanced**	**Total**
3.R.1.1 Decoding and word recognition skills	Application	2	3	2	**7**
3.R.1.2 Comprehension strategies	Analysis	2	3	2	**7**
3.R.1.3 Identify text organizational features and their purpose	Knowledge	2	3	2	**7**
3.R.2.1 Locate, describe, and use text structures	Application	2	3	2	**7**
3.R.2.2 Distinguish differences among various literary elements and devices	Analysis	2	3	2	**7**
3.R.3.1 Respond to ideas and attitudes expressed in literature by making personal connections	Application	2	3	2	**7**
3. R.4.1 Gather information to research a topic	Application	2	3	2	**7**
3. R.4.2 Utilize a set of directions, a model, or diagram in order to complete a project	Application	2	3	2	**7**
TOTAL		**16**	**24**	**16**	**56**

Cognitive Complexity

Cognitive complexity can be described in several different ways. The South Dakota content standards use Bloom's Taxonomy to describe the cognitive complexity for each standard. The cognitive complexity levels in Bloom's Taxonomy include: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. For the purpose of alignment studies, cognitive complexity levels are defined as follows:

Low level (L) – (Bloom's Taxonomy level: **Knowledge**)
This level requires mainly recall, remembering factual information or definitions of terms, or the display of fairly routine skills. This level tends to deal with a single idea or procedure, require a display of concrete understanding, or ask for a demonstration of something learned directly from instruction.

Moderate level (M) – (Bloom's Taxonomy levels: **Comprehension** and **Application**)

This level requires more intellectual skill than those characterized as "Low", but may seem like it is something less than "High." This level may require the application of rules that are practiced extensively in the classroom, but are now applied to a new situation.

High level (H) – (Bloom's Taxonomy levels: **Analysis**, **Synthesis,** and **Evaluation**) This level involves the application of ideas and procedures to solve problems or create new understandings. The situations are not habitual or routine; they are novel for most learners. Often multiple ideas are drawn upon or a high level of abstraction needs to be dealt with.

2.2 Transition to Embedded Field-Testing

Prior to the 2008 administration of the *Dakota STEP*, new items were field tested in separate, stand-alone administrations. Beginning with the 2008 administration (and in subsequent years), field-test items were embedded within the operational items creating eight test forms for each content domain. All students are assessed with the same operational or "core" items on the test forms for each content domain, and field-test items were embedded within the core items. Field-test items do not count toward a student's score but are evaluated for statistical quality for potential use as core items in future test administrations.

Table 2.3 details the number of core and field-test items in the 2008 test forms for each content domain. The number of field-test items was the same across all grade levels within subjects although the distribution of items across standards differed based on item development targets.

Table 2.3: Dakota STEP Item and Test Form Counts by Content Domain

Content Domain		Core Items	Field Test Items	Total Items	Total Forms
Reading					
	Per form	56	20	76	1
Grades 3-8	Per grade	56	160	216	8
	Per form	40	20	60	1
Grade 11	Per grade	40	160	200	8
	Total	**376**	**840**	**1216**	**56**
Mathematics					
	Per form	84	8	92	1
Grades 3-8, 11	Per grade	84	64	148	8
	Total	**588**	**448**	**1036**	**56**
Science					
	Per form	70	7	77	1
Grades 5, 8	Per grade	70	56	126	8
	Per form	84	9	93	1
Grade 11	Per grade	84	72	156	8
	Total	**224**	**184**	**408**	**24**

2.3 Test Construction Specifications

Each *Dakota STEP* assessment is designed to ensure that the South Dakota Content Standards are validly and fairly assessed. The selection of items and assembly of each test form is guided by a set of specifications developed to ensure that test content corresponds to the purposes, objectives, and skills defined in the content standards.

Pearson assessment specialists constructed the 2008 operational *Dakota STEP* assessment following a thorough analysis of the item data as well as alignment study findings.

- The 2008 *Dakota STEP* Mathematics assessment was constructed by replacing existing standards-based items with new operational items that were field-tested in fall 2006 in order to improve the overall alignment of the assessment. New items to assess the content standards were field tested within eight test forms per grade.
- The 2008 *Dakota STEP* Science assessment was constructed by replacing existing standards-based items with new operational items that were field-tested in fall 2006 in order to improve the overall alignment of the assessment. New items to assess the content standards were field tested within eight test forms per grade.
- The 2008 *Dakota STEP* Reading assessment was an exact duplication of the 2007 operational assessment (except in grades 4, 5, and 6, one item was replaced, respectively). New items to assess the revised content standards were field tested within eight test forms per grade.

Selection of Operational Items

The following guidelines were used to select the operational items for the 2008 tests. In all cases, the statistics examined were generated after field testing.

Item Discrimination

The point-biserial correlation is the correlation between item score and total test score. Items selected for operational use discriminate between high and low-performing students. The target point-biserial correlation for the 2008 operational items was >= 0.30. Items with point-biseral values less than 0.30 were only selected for operational use if they were needed to meet the requirements of the test blueprint and approved by the SD DOE.

Differential Item Functioning (DIF)

DIF statistics examine whether an item functions differently for examinees with similar abilities across subgroups. Mantel-Haenszel DIF statistics are calculated for all items, using subgroups that have a minimal sample size of 300. In statistical analysis, items are categorized by their degree of DIF using a system developed by the Educational Testing Service (ETS): Negligible DIF (A), moderate DIF (B), and large DIF (C). It is desirable not to use items classified in category B or C unless content experts determine that flagged items are not biased and are needed for content coverage. Items that were flagged at C category were only selected for 2008 operational use if they were needed to meet the requirements of the test blueprint and approved by the SD DOE.

Reading

The standards-based core items from the 2007 assessment were re-administered in 2008 in order to field test a substantial number of new passages and items developed to assess the revised Reading Content Standards in spring 2009. Administering the same core form in 2008 as in 2007 meant that the raw score-to-scale score tables created in 2007 were also used for the 2008 Reading assessment (except grades 4, 5, and 6).

Mathematics

In order to equate the 2008 test with the 2007 test, a linking set was constructed by selecting items from each of the mathematics content indicators (reporting categories). The items selected resemble the difficulty range of the test.

Science

In order to equate the 2008 test with the 2007 test, a linking set was constructed by selecting items from each of the science content indicators (reporting categories). The items selected resemble the difficulty range of the test.

Placement of Operational Items

Keyed Response

In any particular form, keys of multiple-choice items should have similar percentages across options (ideally, 25% of 4-option items). Items with the same key should never appear more than three times consecutively. Because test-wise examinees may be aware that *C* tends to be the correct answer for difficult items, it is essential to check keys of difficult items to ensure that the percentage of *C* being the correct answer is similar to the percentage of easy items.

Items Containing Graphics

Items containing graphics should be balanced with those without graphics so that items with graphics are evenly distributed across test forms.

Selection and Placement of Field-test Items

Item Selection

Items approved at Content and Bias Review meetings were selected for field testing based on (1) the need for standards coverage and (2) the need for format variety.

Item Location

SD DOE has requested the random placement of field test items throughout the operational forms.

2.4 Item Specifications

The *Dakota STEP* subtests are constructed with multiple-choice items that require students to choose the correct answer and mark their response by filling in a bubble on their answer document. The *Dakota STEP* Item Specifications require that:

- Items appropriately measure and align with the grade-specific content standards. Items focus on what students should know and be able to do at their grade level.
- Items developed for the *Dakota STEP* are reviewed and approved by the SD Content and Bias Review committees and SD DOE prior to field testing.
- Items are drawn from the passage or stimulus in such a way that a student's previous knowledge does not impact his or her response. Items focus on skills and comprehension strategies, avoiding measurement of a student's feelings or values.
- Items are worded precisely and clearly.
- Correct answers are approximately equally distributed among A's, B's, C's, and D's.
- Response choices are approximately the same length, have the same format, and are syntactically and parallel. Students should not be able to rule out a wrong answer or identify a correct response simply by virtue of its looking or sounding different.
- Distractors are plausible although incorrect in the context of the stimulus. Distractors adopt the language and sense of the material in the stimuli so that students must think their way to the correct answer.
- The order of presentation is dictated by logic (chronologically, spatially, etc.)
- A balance of gender and active/passive roles by gender is maintained.
- To the greatest extent possible, no item or response choice clues the answer to any other item.

2.5 Passage Specifications

Pearson commissions quality reading passages from experienced writers. Passages must be written to be age-appropriate, relevant, and engaging. They must also be free of bias and sensitivity concerns, tightly focused, and well organized. In order to fulfill blueprint requirements and to maintain a grade-appropriate emphasis on various passage types, passage development reflects the approximate proportions as shown in Tables 2.4 and 2.5 below.

Table 2.4: Approximate Percentages for Passage Type

Grade	Passage Type		
	Literary	**Informational**	**Functional**
3-5	50	35	15
6-8, 11	40	35	25

Table 2.5: Passage Word Count Ranges

Grade	Informational and Literary	Functional	Poems (Maximum)	Paired Passages (Maximum)
3	300-700	300-500	150	700
4	300-700	300-500	150	700
5	400-800	300-600	200	900
6	500-900	400-700	250	900
7	500-900	400-700	250	1000
8	600-1000	500-800	400	1100
11	700-1200	600-1100	500	1200

Informational Passages

General non-fiction articles and biographies: These selections emphasize documented fact, with the primary purpose being to inform the reader. Topics include noteworthy people who are no longer living, historical and current events, and expository passages in content areas (e.g., art, music, science, history, government). Informational passages should be written so that the reader may be asked to:

- Identify major points and distinguish them from supporting details.
- Extract main ideas and indicate relevance of supporting details.
- Locate information using graphics such as charts, graphs, maps, and timelines.
- Analyze and use organizational features of text such as headings, bold print, etc.
- Analyze how text is organized (sequence, cause/effect, chronology, or categorization).
- Explain how the organization of the text supports the writer's major concepts and purposes.
- Connect concepts, issues, and information presented in text with commonly understood concepts about the world.

Paired Informational

These passages consist of a pairing of informational passages that exhibit the qualities described above. Additionally, paired passages lend themselves to items that involve comparison and/or contrast of details, ideas, issues, organization, scope, purpose, etc. Paired passages usually focus on similar events or people, or they focus on the same event or person but through a different perspective.

Informational Narrative

These passages are related from first-person point of view and involve a sequence of events. They focus on a topic and provide information, and they exhibit some of the characteristics of an informational passage, but they do so through the context of a meaningful experience that the speaker shares. For example, a speaker may share an experience about planting a garden with his or her grandparent and explain about different types of seeds and plants.

Functional Passages

These selections convey practical, everyday information. Examples include Rules, guidelines, and entry forms; Safety, assembly, or operating instructions; Memos, manuals, and handbooks; Applications, resumes, and letters, both business and personal; Print ads, including posters and flyers; Newspaper articles and letters to the editor. Functional passages should be written so that the reader may be asked to:

- Demonstrate an accurate understanding of the information.
- Explain how key aspects of the content are relevant to a reader's informational needs.
- Select information appropriate to the task and apply it logically and validly to complete the task.
- Select information and details from the text to complete the task accurately, adequately, and appropriately.
- Cite an adequate amount of information to support conclusions about the relevance and importance of the information read.
- Identify and evaluate persuasive techniques.

Paired Functional

These passages consist of a paring of functional passages that exhibit the qualities described above. Usually, these pairings consists of two advertisements for similar products, two fliers for similar activities, or a newspaper article and letter to the editor about a relevant issue. These selections require students to process and discriminate information found in "real-world" materials.

Literary Passages

Fictional Narrative

Short stories must contain a clear progression of events, believable characters, a well-developed plot and/or theme, a clear and deliberate point of view, an identifiable conflict, a climax and/or turning point, and a resolution. Fiction genres include contemporary fiction, mystery, historical fiction, science fiction, fantasy, legends, tall tales, and folk tales. Fictional narratives must be written so that the reader may be asked to:

- Understand the development of characters using details from that story to accurately report what the characters do, say, and think; what other characters say about them and what the writer says directly.
- Understand how the depiction of major and minor characters is important to a story's plot and theme, or simply what about the description of characters makes them worth remembering.
- Use an adequate amount of details from the story to describe and discuss its characters and their roles.
- Understand the author's use of symbolism and identify details that establish the symbolism.
- Describe the importance of symbolism to the story's plot and theme.
- Recognize a number of techniques and details the author has used to establish symbolism in a story.
- Identify the moods of a story by locating details the author has used to establish the mood.
- Describe the importance of the mood to the story's plot and theme.
- Recognize a number of techniques and details the author has used to establish one or more moods throughout a story.
- Identify initiating action, or conflict, in a story's plot in relation to its highest point, the climax.
- Understand how high points in a story's action builds to its climax and/or can identify the presentation of a problem and its solution or resolution as basic to the story's plot structure.
- Describe all of the significant events in a story and the sequence in which those events occur.
- Articulate a reasonable story theme.
- Explain how various elements of the story contribute to the development of its theme.
- Provide enough information to support his or her interpretation of the story's theme.
- Relate details from the story to social, cultural, or historical contexts.
- Recognize and explain the use of figurative language, imagery, etc.

Poetry

Poems must (unless the form dictates otherwise) use literary techniques such as meter, rhyme, sound and figurative language to convey emotion, experience, or ideas. Most importantly,

however, poems must have depth. A poem, for example, that just focuses on a scene in nature lacks the depth needed for assessment purposes. Instead, the same poem might focus on a bee's journey through nature in search of its hive. Every poem must have a universal theme. Examples of poetic forms include lyric, ballad, haiku, narrative, free verse, and sonnet.

Cultural Diversity

Multicultural literature refers to texts that feature storylines that are about members of racial, religious, or language micro-cultures other than Euro-American. The five most populous micro-cultures in the United States are:

- African American;
- Asian American (including Chinese, Japanese, Korean, and Vietnamese descent);
- Latino American (including Cuban Americans, Mexican Americans, Puerto Ricans, and others of Spanish descent);
- Native American (a general term referring to the many tribes of American Indians); and
- Jewish.

Culturally diverse texts

- Children who see people like themselves represented in texts derive self-esteem and pride in their own heritage.
- Children learn about people and their cultures.
- Each culture has much to teach to other cultures.
- Engagement with a certain culture can reduce students' prejudices toward micro-cultures.
- Reading about people of one's own culture facing issues/problems may help children of that same and other cultures cope with the same problems themselves.

Features of culturally diverse texts:

- Avoids racial and cultural stereotyping. Instead multi-faceted, well-rounded characters are featured. The nature of stereotyping is that it unfairly assigns a fixed image or fixed characteristics to everyone within a group, thereby denying everyone with the group the right to any individuality or choice.
- Cultural details are reasonably accurate. Regional differences make it impossible for portrayals of events, celebrations, holidays, and so forth, to be declared as "the way" something is done. Authors should be careful to avoid sweeping generalizations.

Taboo Topics

Pearson will consider accepting passages with contexts that address most topics as long as they are tastefully handled, well written, and present an interesting, real-world situation. Writers are advised, however, to avoid subject matter that would cause a selection to be deemed unacceptable by SD DOE and SD educators for any of the following reasons:

1. The topic could evoke unpleasant emotions in the test takers that are likely to hamper their ability to take the remainder of the test in the optimal frame of mind.
2. The topic is controversial among the adult population and might not be acceptable in a classroom setting.

3. The topic has been over used in tests or textbooks and is thus overly familiar and/or boring to students.
4. The topic will appear biased against (or toward) some group of people.

2.6 Item Development

It is essential that the *Dakota STEP* assessment measures the depth, breadth, and intent of the South Dakota State Content Standards for each grade level. Important information regarding student performance must be derived from the standards-based test. The items developed must adequately measure and align with the South Dakota Content Standards in order to provide a valid assessment.

The item development process begins with a thorough review of the content standards, the test specifications and blueprints, and the pool of items available for operational use by Pearson's assessment specialists. The result is a clear focus on areas in need of additional development. The "gaps," or content standards with insufficient numbers of items within the existing item pool, provide a plan for item development. A sufficient number of new items must be written to allow for attrition throughout the review and field testing process.

Once the item development plan has been reviewed and approved by the South Dakota Department of Education, item development assignments and support materials are provided to Pearson's cadre of trained, professional item writers. The item writers submit their items to Pearson's assessment specialists who conduct a thorough quality review of items using criteria such as:

- Does the item measure the standard it intends to measure?
- Does the item conform to best practices in item design?
- Is the content of the item accurate?
- Is the item unique?
- Is the language of the item clear and concise?
- Are the distractors plausible?
- Does the item have only one correct answer?
- Is the language of the item simple and concise?

The initial review is followed by an additional review in which a senior assessment specialist analyzes the item for accurate content and best testing practices. Thereafter, the items go through an editorial review to make sure that proper vocabulary, spelling, and grammar are used. Editorial specialists also correct items that do not conform to the current style for item formatting. New items are developed to stylistically match the existing assessment. There is an intentional parallelism among all items, art, and passages to include phraseology, sentence structure, and stimulus attributes. Edited items are then passed back to the lead assessment specialist, who conducts an additional review.

Items that are not acceptable are returned to the item writer with feedback and suggestions for rewriting the item. Feedback about accepted items is also frequently supplied to writers. Edits are explained and suggestions are given on how to improve future work. This feedback process allows our item writers to create higher-quality items with each new submission.

All items go through an internal review process at Pearson before they are deemed suitable for external committee reviews. This internal review phase allows the items to be refined and aligned with content standards before they are presented to South Dakota educators and the Department of Education. The additional series of reviews continues to add value to the items as they are scrutinized for alignment to the content standards and the absence of bias and stereotyping.

2.7 South Dakota Teacher Committees

One of the most important sources of evidence of validity for an assessment is the test content. How well the test items represent the emphasis of subject matter and learning processes taught and assessed in classrooms is a vital element in judging the validity of test results. For this reason, South Dakota teachers are actively involved in the item development process as well as alignment and standard setting workshops. SD DOE recruits teachers for various committee meetings based on a demographic sampling plan and nomination process. The SD DOE communicates with teachers/administrators (i.e., superintendents, principals, and curriculum directors) describing the work that will take place, including dates/times of meetings and rate of pay. Teachers/administrators then submit nomination forms for participants to the SD DOE.

The SD DOE recruits and selects committee participants based on the number of teachers needed for each of the content and grade-specific panels as well as the following criteria: geographic region of the state, district size, and teachers who have worked with various student groups including special education and limited English proficient students. Selected participants are assigned to specific grade or grade span groups based on professional experience and knowledge. Meetings are conducted under the guidance of SD DOE staff, and facilitated by subject matter experts from Pearson (e.g., content bias review, data review) or experts from the Buros Institute for Assessment Consultation and Outreach (e.g., standard setting, alignment study).

Content and Bias Review

Content and Bias review meetings are facilitated annually by Pearson assessment specialists in order to review newly developed items and reading passages prior to field testing. Participants receive training in a general session to cover the objectives and process of the review session. All items are reviewed to ensure that they are appropriate and engaging for the proposed grade level, measure the desired content standards, are clearly worded, have only one correct answer, and are of appropriate difficulty. Items and passages are also reviewed to ensure that they are free of bias and sensitivity concerns that could unfairly advantage or disadvantage (unrelated to mastery of the content) a student or group with certain personal characteristics (gender, race, ethnicity, religion, socioeconomic status, disability, or geographic region). Reviewers are asked to consider whether the subject matter and language of items or reading passages are acceptable to students, their parents, and other members of South Dakota communities.

Once general training has been completed, participants separate into their grade/grade span specific groups to begin reviewing items. Committee review books containing the items to be reviewed are distributed as secure documents to each participant. The process begins with each group participant conducting an individual review of a designated number of items (usually all items developed to assess a specific content standard or strand). Once individual reviews have been completed, the Pearson facilitator leads the group in discussion of each item and in

carefully reviewing the individual comments until the group reaches a consensus to accept, accept with edits, or reject each item. This process continues until all items have been reviewed and committee consensus has been documented and reconciled. Participants review the items in order to make recommendations as to whether each item is suitable for field testing. SD DOE then reviews the committee's recommendations for final approval of the items.

Table 2.6: 2008 Content/Bias Review Results

	Grade	Reviewed	Accepted	Accepted with Revisions	Rejected
Reading Passages	3	16	6	10	0
	4	16	5	10	1
	5	16	2	12	2
	6	16	3	13	0
	7	17	1	14	2
	8	18	5	11	2
	11	15	1	11	3
Reading Items	3	203	150	50	3
	4	213	164	47	2
	5	201	151	50	0
	6	204	163	37	4
	7	210	133	67	10
	8	200	116	76	8
	11	201	148	46	7
Math Items	3	80	24	54	2
	4	74	18	56	0
	5	80	38	42	0
	6	79	57	22	0
	7	78	44	34	0
	8	80	66	13	1
	11	86	46	38	2
Science Items	5	57	14	38	5
	8	59	36	14	9
	11	75	29	40	6

Alignment Workshops

South Dakota conducts two types of alignment studies annually. BIACO facilitates “pre-test” alignment workshops to align newly developed items to the content standards and designate the cognitive complexity levels (low, medium, and high) of items prior to field testing. The assigned content standard and cognitive complexity of each item is then captured for reference if and when the items are used on operational forms in order to fulfill the test blueprint requirements.

The findings of these "pre-test" alignment studies are also utilized to identify the strengths and weaknesses of the item pool in order to target item development efforts.

Participants receive training by BIACO consultants in a general session to cover the objectives and process of the alignment workshop. Pearson provides an overview of the item and test development process and provides content support for the grade-specific workgroups. Following the general training session, the panelists are separated into content and grade-specific groups, and are provided with the relevant core content standards; item review books, and rating forms to record their individual findings.

The panelists review a portion of the items individually, and then pull together as a group to discuss their individual judgments. The groups are tasked with reaching consensus on the content standard that each item is aligned with as well as the cognitive complexity of each item. BIACO specialists have the content standard and cognitive complexity of the items as determined by the Content and Bias Review committee. If the alignment group determines that an item is aligned to a different standard, the South Dakota Curriculum Specialist and the BIACO facilitator are available to offer explanation and guidance. The group then records their final independent decision on the group rating form.

Following the operational administration of the test, Norman Webb facilitates a "post-test" alignment study of the intact operational forms. This study intends to provide content validity evidence on the overall alignment of the operational assessment. Additional information regarding the overall alignment of the spring 2008 test forms is available in section 8.4 and Appendices O, P, and Q of this technical report.

Data Review

Data Review meetings serve as the final review of newly developed items after field testing and before use as operational items. These meetings are conducted under the guidance of SD DOE staff, and facilitated by psychometric staff and subject matter experts from Pearson. During the meeting, participants review the items based on their field test statistics in order to make recommendations as to whether the items reviewed are eligible for inclusion in future revisions of the operational test.

Following a general training session explaining the item statistics that participants will review and their relevance to the process, participants are separated into content and grade-specific groups, and the actual review of the items takes place. For each item, the committee will review the following item statistics:

- p-value (item difficulty)
- point-biserial correlation (item discrimination)
- Mantel-Haenszel differential item functioning statistics (item bias whereas minimum n-counts, say 200, per demographic group can be obtained)
- Overall student response broken down by student scores

The committee reviews the items in the light of the values of these statistics, concentrating on potential defects in items that may result in:

- inappropriate difficulty (too hard or too easy)

- non-discriminating characteristics
- differentially difficult for some demographic group(s)

Care must be taken to ensure that only items with specific and identifiable defects are rejected. Items that address appropriate content at the appropriate grade level can occasionally have item statistics that are not optimal for reasons other than defects intrinsic to the item. Such items can provide valuable information to teachers, parents, and educational administrators and so should not be rejected simply on the basis of item statistics.

Items accepted by SD DOE based on the Data Review Committee recommendations are added to the pool of items available for future operational use. Rejected items are normally discarded, however; if a rejected item is revised, it must be treated as a new item, field tested, and reviewed by a subsequent Data Review committee in order to be eligible for operational use.

Table 2.7: 2008 Data Review Results

	Grade	Items Reviewed	Accepted	Rejected	Revise & Re-Field Test
Reading	3	160	152	8	0
	4	160	157	3	0
	5	160	154	6	0
	6	160	153	7	0
	7	160	148	12	0
	8	160	145	15	0
	11	160	142	18	0
Math	3	64	63	1	0
	4	64	57	4	3
	5	64	61	0	3
	6	64	60	1	3
	7	64	48	7	9
	8	64	53	4	7
	11	64	62	2	0
Science	5	56	45	11	0
	8	56	48	8	0
	11	72	62	10	0

Standard Setting

Standard setting is conducted following the administration of a new assessment. After the operational test has been scored, statistical data from student performance are reviewed to ensure that both the items and the test as a whole meet the established test design and psychometric criteria. Standards setting committees then establish achievement level “cut score” recommendations after reviewing the operational assessment and evaluating what students must

know in order to answer each item. The cut scores represent the threshold between each level of performance—Basic, Proficient, and Advanced.

Standard setting workshops are facilitated as needed by the BIACO. The SD DOE recruits educators from the targeted subject area. Participants are presented an orientation, followed by an overview of the method to be used in the standard setting process. Following this session, participants separate into grade-level groups. Participants then review the performance level descriptors (Basic, Proficient, and Advanced) for their respective grade levels and experienced the standard-setting method using a practice test composed of relevant content and grade-level material.

Participants in the grade-specific groups then begin their Round 1 Review of the assessment. They rate each item on the test as Below Basic, Just Basic, Just Proficient, Just Advanced, and Above Advanced using specially designed forms. When all of the participants within a group complete their first round of ratings, they are provided with information about how the respective grade-level students in South Dakota actually performed on the test questions during the administration. They are also shown a cumulative percent distribution using their Round 1 cut score results that indicates the percentage of South Dakota students at that grade level who would be classified as Basic, Proficient, and Advanced. Participants are then asked to reconsider their Round 1 ratings during Round 2.

The analysis that follows the Round 2 ratings involves the computation of cut scores for each grade-specific assessment for Basic, Proficient, and Advanced achievement level descriptors. These cut points, and their respective standard errors of estimate, are used as boundary values for the policy decision for establishing cut scores for the *Dakota STEP* tests. The SD DOE then reviews and presents the recommended cut scores to the SD BOE for approval. The existing cut points for the *Dakota STEP* Reading, Mathematics, and Science assessments were established in spring 2005, 2006, and 2007 respectively. Additional information regarding translation of raw scores to scale scores is provided in Chapter 4 of this technical report.

CHAPTER 3: TEST ADMINISTRATION

The administration of the *Dakota STEP* is a shared and coordinated effort involving the SD DOE, its testing contractor, Pearson, and district and school personnel. Pearson is responsible for printing and shipping test materials to districts as well as receiving and scoring materials from districts following each test administration. Table 3.1 lists the key events and dates for the 2008 *Dakota STEP* administration.

Table 3.1: Key Events and Dates for the 2008 Dakota STEP Administration

February 11– February 14, 2008	Pretest workshops
March 5–11, 2008	Delivery of *Dakota STEP* materials
March 12–21, 2008	Deadline for requesting additional materials
March 31–April 18, 2008	*Dakota STEP* test administration
April 25, 2008	Deadline for receipt of materials from districts at the Pearson Scoring Center
June 6, 2008	Scored data file to SD DOE

The SD DOE and Pearson staff conducted four training sessions for test coordinators at locations throughout the state (Rapid City, Pierre, Aberdeen, and Sioux Falls), covering the test administration process, test security measures, and the roles/responsibilities of testing coordinators. Approximately 75% of the state's district test coordinators attended these workshops. One training session was recorded and placed on the State website for coordinators unable to attend a session in person (http://doe.sd.gov/octa/assessment/dakSTEP/index.asp). After the initial training by SD DOE and Pearson the district test coordinators provided school level training within their districts.

District-level test coordinators serve as the point of contact among the districts, schools, the DOE, and Pearson. They are responsible for receiving, verifying, and distributing all test materials to schools, as well as returning materials to Pearson for scoring. In larger districts, test coordinators designate building coordinators to oversee some of these responsibilities. They also provide training to building coordinators and teachers/examiners regarding test administration procedures and are available during testing to answer questions. Teachers/examiners administer the *Dakota STEP* to students and supervise testing sessions.

3.1 Test Security Guidelines

The *Dakota STEP* test administration must be completed in a timely manner and conducted in such a way to ensure appropriate and consistent testing conditions, as well as the secure handling of all test materials. Test security guidelines described below prohibit activities that could result in an unfair advantage to some students, misrepresentation of results, or exposure of secure test items.

Teachers/examiners administering assessments are not to provide answers to students, copy any portion of the test, or in any way contribute to test security practices that violate the Code of Professional Ethics set for South Dakota educators. The South Dakota Professional Teachers Practices and Standards Commission Administrative Rule Codes 24:08:03:01 (Obligations to Students), 24:08:03:02 (Obligations to the Public), and 24:08:03:03 (Obligations to the Profession) provide direction for educators. The Code of Ethics for professional administrators is cited in Administrative Rule 24:11:03:01:

Unethical test practices include, but are not limited to:

1. Encouraging students to be absent the day of testing.
2. Encouraging students not to do their best.
3. Not testing all students.
4. Reclassifying students solely for the purpose of avoiding state testing.
5. Failing to provide necessary accommodations during testing (if applicable).
6. Interpreting, explaining, or paraphrasing a test item.
7. Copying of the test in any way.
8. Changing or altering student responses at any time.
9. Using secure test items for instruction.

The SD DOE and Pearson require any person who handles test materials (including test coordinators, building coordinators, and teachers/examiners) to sign a test security agreement prior to the test administration stating that they have been made aware of these regulations and procedures and agree to follow them. The *Dakota STEP Test Security Agreement* and *Test Security Affidavit* are included in Appendix B.

Dakota STEP test materials must be kept in a locked secure location before, between, and after testing sessions. Upon completion of testing, all scorable documents and non-scorable test booklets/materials must be collected and returned to Pearson. Any missing documents or other potential breaches of security are to be reported to the test coordinator. If the documents are not found or if a security breach is suspected, the SD DOE must be notified utilizing the *Report of Test Irregularity Form* (Appendix B).

3.2 Test Administration

With respect to the 2008 administration, test coordinators working with building coordinators and test administrators scheduled *Dakota STEP* assessment sessions within the statewide administration period of March 31–April 18, 2008 (including any make-up sessions). Schedules had to allow time for district test coordinators to assemble and ship all scorable documents to arrive at Pearson on or before April 25, 2008.

With respect to its annual statewide assessment, the SD DOE strongly recommends that each content area of the assessment be administered on different days with stretch breaks provided to students between test sessions as noted in the *Directions for Administering* and students' test materials. The Reading, Mathematics, and Science subtests are further divided into multiple test sessions as shown in Table 3.2.

The *Dakota STEP* subtests are not timed. Approximate testing times are provided to districts and schools for planning purposes only. While the time allotted for the administration of the subtests

is adequate for most students, additional time is allotted for students to complete the subtests provided they are actively engaged in testing. Table 3.2 shows the approximate testing times for each of the *Dakota STEP* subtests.

Table 3.2: Approximate Testing Times for the Dakota STEP

Subtests and Sessions	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
Reading Session 1	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	60 minutes
Reading Session 2	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	60 minutes
Reading Session 3	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	N/A
Total Reading	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours**
Mathematics Session 1	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes
Mathematics Session 2	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes
Mathematics Session 3	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes	45 minutes
Total Mathematics	**2 hours 25 min.**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**	**2 hours 25 min**
Science Session 1			45 minutes			45 minutes	45 minutes
Science Session 2			45 minutes			45 minutes	45 minutes
Science Session 3			45 minutes			45 minutes	45 minutes
Total Science			2 hours 25 min			2 hours 25 min	2 hours 25 min
Total Testing Time	**4 hours 30 min**	**4 hours 30 min**	**6 hours 45 min.**	**4 hours 30 min**	**4 hours 30 min**	**6 hours 45 min**	**6 hours 30 min**

3.3 Assessment of Students with Disabilities

The *Individuals with Disabilities Education Act of 2004* (IDEA) requires the development of policies and procedures for the inclusion of students with disabilities in statewide assessment and, when necessary to ensure a student's full participation, the provision of testing accommodations. Individualized Education Program (IEP) must include accommodations that may be necessary in order for the student to participate in assessment. A student who is perceived to have a disability based on Section 504 of the *Rehabilitation Act of 1973* (Section 504) must also be afforded accommodations if those accommodations are part of the services provided in the student's Section 504 Plan.

State-approved assessment accommodations for students with disabilities include accommodations in presentation, response, setting, timing, and scheduling as shown in Table 3.3. Each accommodation provided during testing must be documented in the space provided on the student's answer document. Students must have ample time to learn to use instructional and assessment accommodations before the *Dakota STEP* administration. Accommodations must be documented within the student' IEP, used in day-to-day instruction, and should be in place at least 5 weeks prior to test administration.

Decisions about whether a student participates in the *Dakota STEP* assessment under the standardized conditions or participates with state-approved accommodations must be made on an individual basis by the student's IEP or Section 504 team. If it is determined that a student with a disability can participate in the *Dakota STEP* assessment with approved accommodations, the team making that decision must specifically indicate the type and extent of accommodations that will be provided. This information must be included in the student's IEP or addressed in the student's Section 504 Plan. The parent/guardian of the student must be made aware of the decision during the development of the student's IEP or Section 504 Plan. If it is determined that a student with a disability cannot participate in the *Dakota STEP* assessment with approved accommodations and the student meets the significant cognitive disability criteria, the student must be assessed utilizing the *Dakota STEP-A* (Alternate Assessment). The IEP team must document their decision in the student's IEP.

3.4 Test Accommodations

Standardization is an essential feature of educational assessments and is necessary to produce valid and reliable information about student learning. Strict adherence to guidelines detailing instructions and procedures for the administration of accommodations is necessary to ensure that the *Dakota STEP* assessment results reflect actual student learning. Once decisions have been made about providing accommodations to meet individual student needs, the logistics of providing the actual accommodations during the *Dakota STEP* administration must be planned.

It is not uncommon for members of the IEP team, most often special education teachers, to be given the responsibility for determining, arranging, coordinating, and providing assessment accommodations for all students who may need them. Thus, it is essential for all IEP team members to know and understand the requirements and consequences of statewide assessment, including the use of state-approved accommodations when necessary for a student's meaningful participation. It is important to engage the appropriate personnel to plan the logistics and provisions of assessment accommodations prior to testing. Detailed information regarding accommodations can be found in the South Dakota Accommodations Manual available at: http://doe.sd.gov/oess/specialed/forms/pdf/SDAccommodationsManual.pdf .

Table 3.3: Dakota STEP Assessment Accommodations

Reading		Mathematics		Science		Accommodation Codes	Accommodations for Special Populations
IEP/ 504	LEP	IEP/ 504	LEP	IEP/ 504	LEP		**Presentation Accommodations**
✔		✔		✔		1	Large-print test booklets and/or answer documents
✔		✔		✔		2	Visual aids (magnifiers, templates)
✔		✔		✔		3	Sign language (ASL, cued speech)
		✔				4	Abacus for visually impaired VI students
✔		✔		✔		5	Braille test booklets
✔	✔	✔	✔	✔	✔	6	Test Items read aloud, including audiotape/cd and assistive technology (does NOT include *Reading Comprehension* passages)
✔	✔	✔	✔	✔	✔	7	Repeating and/or simplifying directions
✔		✔		✔		8	Amplification equipment (hearing aid, auditory trainer)
						9	N/A - DO NOT GRID ON THE DEMOGRAPHIC INFORMATION PAGE
		✔		✔		10	Talking calculators, provided the student is tested individually or with the use of headphones
							Response Accommodations
✔		✔		✔		11	Responses marked directly in test booklet or on large-print answer document, also includes oral, sign language, Braille, and recorded responses to test items (answers must be transcribed to a scorable answer document by scribe)
✔		✔		✔		12	Large-diameter pencil, pencil grip, special pencil or pen
	✔		✔		✔	13	Use of dictionaries, word lists, glossaries
✔	✔	✔	✔	✔	✔	14	Visual organizers (i.e. templates, masks, markers, graph paper, rulers)
							Setting, Timing, and Scheduling Accommodations
✔	✔	✔	✔	✔	✔	15	Environmental modifications (i.e. special lighting, adaptive or special furniture, location with minimal distractions, noise buffers, carrels, special seating)
✔	✔	✔	✔	✔	✔	16	Small-group administration
✔	✔	✔	✔	✔	✔	17	Individual administration
✔		✔		✔		18	Home/Hospital Setting
✔	✔	✔	✔	✔	✔	19	Flexible schedules (i.e. time of day, multiple breaks, etc.)
						20	**Other (with prior approval from SD DOE)**

State-approved accommodations for the *Dakota STEP* test administration are categorized as follows and detailed in Table 3.3.

> **Presentation accommodations** allow students to access information in ways that do not require them to visually read standard print. These alternate modes of access are auditory, multi-sensory, tactile, and visual.
>
> **Response accommodations** allow students to complete activities, assignments, and assessments in different ways or to solve or organize problems using some type of assistive device or organizer.
>
> **Setting, timing, and scheduling accommodations** change the location or the conditions of the assessment setting, increase the allowable length of time to complete an assessment, and/or change the way the time is organized.

When identifying accommodations for the state assessment, it is important to review state policies and procedures in order to determine whether the use of a particular accommodation may result in the invalidation of the test (e.g., inappropriately lowering or not counting a student's score). Appropriate accommodations do not reduce learning expectations; they provide access to an assessment. However, inappropriate modifications or alterations may change, lower, or reduce learning expectations. Providing a student with such a modification during a state accountability assessment may constitute a test irregularity and may result in an investigation into the school or district's testing practices.

Some accommodations are standard for both students with disabilities and limited English proficient (LEP) students, while other accommodations are only permissible for one student classification and/or subject area. Each accommodation, numbered 1 through 20, corresponds to the "Accommodations" grid on the student demographic information page. Following the test administration, test administrators or case managers must grid the corresponding circle(s) for each accommodation provided for each content area.

3.5 Assessment of LEP Students

All identified limited English proficient (LEP) students in grades 3 through 8 and 11 must participate in the *Dakota STEP* assessment (Title I, Part A, Section 1111 of the *No Child Left Behind Act of 2001*). Academic assessment of eligible LEP students must be accomplished in English for those LEP students who have attended school in the United States for 3 or more consecutive school years. If a student has fewer than 3 consecutive school years in the United States, the student should take the assessment in the language and form most likely to yield accurate data. In South Dakota, the *Dakota STEP* is only provided in English; therefore, LEP students are provided state-approved accommodations based on their IEP Plan.

LEP students in their first year of enrollment in a school in the United States are not required to take the *Dakota STEP* Reading subtest if these students have participated in the annual *Dakota English Language Proficiency (Dakota ELP)* assessment. Participation in the *Dakota ELP* assessment will constitute participation in *Dakota STEP* Reading for purposes of determining AYP. Students who enroll for the first time in a school in the United States after the testing window for *Dakota ELP* annual progress has ended will meet participation requirements for Reading through the completion of the LEP eligibility assessment.

LEP students in their first year of enrollment in a school in the United States are required to take the *Dakota STEP* Mathematics and Science assessments, indicating participation for AYP determination. The results of the Mathematics and Science assessments for LEP students in their first year of enrollment in a United States school will not be included in the determination of AYP for the school, district, or state, even if the student meets the requirements of attendance for a full academic year.

State-approved assessment accommodations for LEP students include: Test items read aloud (in English), repeating and/or simplifying directions, word lists/glossaries, visual organizers, environmental modifications, small group/individual administrations, and flexible testing schedules as shown in Table 3.3. Each accommodation provided during testing must be documented in the space provided on the student's answer document. Test coordinators are urged to plan ahead when providing accommodations to LEP students. Since every student is different and language abilities and needs vary widely, any testing accommodation made available to one student will not necessarily be applied universally.

Prior to the administration of the *Dakota STEP* assessment, districts should determine each LEP student's need for accommodations that are most likely to yield accurate and reliable information on what the student knows and can do in the subjects tested. Such determinations should be made by LEP teams comprised of teachers, counselors, and administrators with specific knowledge of the student involved on an individual basis.

The following are general guidelines for LEP teams to consider when determining accommodations for individual students:

- The student's ability to comprehend and follow standard instructions delivered in English (whether oral or written) as compared to another language.
- The student's ability to comprehend and appropriately respond to standard test items written in English.
- The language that will best allow the student to demonstrate his or her proficiency in the skill(s) being tested.
- Timing or pacing variations that may assist in English comprehension.
- Responsive variations that may minimize English-language limitations.
- Encoding or decoding assistance, including interpreters or translators.

If it is determined that an LEP student will participate in the *Dakota STEP* assessment with accommodations, the team making that decision must specifically indicate the type and extent of accommodations that will be provided. A student's LEP team must consider the individual needs of the student in daily instructional settings as well as the additional needs that arise in a secure testing environment. Decisions should be documented in writing and maintained in the student's permanent file.

CHAPTER 4: ACHIEVEMENT LEVELS

4.1 Achievement Level Descriptors

The South Dakota Content Standards provide achievement descriptors for three levels–Basic, Proficient, and Advanced. The following general achievement descriptors are written at a broad level and are applicable to all content domains. Grade-specific achievement descriptors are documented within the South Dakota Academic Content Standards and are available online at http://doe.sd.gov/contentstandards.

Basic: A student performing at the basic level performs below expectations for that grade level. The student is able to perform some of the content standards for the grade below the level of difficulty, complexity, or fluency specified by the grade-level standards.

Proficient: A student performing at the proficient level meets expectations for that grade level. The student is able to perform the content standards for the grade at the level of difficulty, complexity, or fluency specified by the standards.

Advanced: A student performing at the advanced level exceeds expectations for that grade level. The student is able to perform the content standards for the grade at a high level of difficulty, complexity, or fluency.

A student performing at the Below Basic level is unable to perform the content standards for the grade; therefore, no description is provided for this achievement level.

4.2 Achievement Level Cut Points

The achievement level descriptors are translated into *Dakota STEP* cut points by Standard-Setting Committees. The panelists' judgments serve as empirical evidence to inform the policymaking process. Decisions regarding the final cut points are a policy decision made by the SD DOE informed by the panel's recommendations. The final cut points are then submitted to the South Dakota Board of Education for approval. The existing cut points for the *Dakota STEP* Reading, Mathematics, and Science assessments were established in spring 2005, 2006, and 2007 respectively.

4.3 Achievement Level Raw Scores (Scale Scores)

Tables 4.1, 4.2, and 4.3 provide the raw scores and corresponding scale scores for each achievement level as recommended by the Standard-Setting Committee, finalized by the SD DOE, and approved by the South Dakota Board of Education. Note that the raw score to scale score conversions were obtained based on the ability estimates, and the calibration was done using all of the students with valid test scores excluding extreme scores (i.e., zero or perfect scores).

Table 4.1 provides the raw scores and corresponding scale scores assigned to each achievement level for the spring 2008 *Dakota STEP* Reading assessment.

Table 4.1: Raw Score (Scale Score) Ranges for Dakota STEP Reading

Grade	Below Basic Raw score (Scale score)	Basic Raw score (Scale score)	Proficient Raw score (Scale score)	Advanced Raw score (Scale score)
3	0 – 3 (401 – 497)	4 – 23 (498 – 594)	24 – 44 (595 – 663)	45 – 56 (664 – 805)
4	0 – 5 (418 – 524)	6 – 23 (525 – 606)	24 – 43 (607 – 667)	44 – 56 (668 – 813)
5	0 – 5 (438 – 543)	6 – 24 (544 – 620)	25 – 43 (621 – 677)	44 – 56 (678 – 822)
6	0 – 6 (444 – 555)	7 – 26 (556 – 635)	27 – 43 (636 – 691)	44 – 56 (692 – 842)
7	0 – 4 (463 – 564)	5 – 25 (565 – 650)	26 – 44 (651 – 708)	45 – 56 (709 – 849)
8	0 – 4 (475 – 573)	5 – 26 (574 – 667)	27 – 45 (668 – 727)	46 – 56 (728 – 864)
11	0 – 5 (513 – 624)	6 – 24 (625 – 718)	25 – 35 (719 – 779)	36 – 40 (780 – 885)

Data file 5/29/07 for grades 3, 7, 8, and 11 raw score to sale score conversion.
Data file 5/19/08 for grades 4, 5, and 6 raw score to sale score conversion.

Table 4.2 provides the raw scores and corresponding scale scores assigned to each achievement level for the spring 2008 *Dakota STEP* Mathematics assessment.

Table 4.2: Raw Score (Scale Score) Ranges for Dakota STEP Mathematics

Grade	Below Basic Raw score (Scale score)	Basic Raw score (Scale score)	Proficient Raw score (Scale score)	Advanced Raw score (Scale score)
3	0 – 18 (371 – 529)	19 – 50 (530 – 599)	51 – 75 (600 – 667)	76 – 84 (668 – 793)
4	0 – 12 (406-547)	13 – 43 (548 – 617)	44 – 71 (618 – 683)	72 – 84 (684 – 824)
5	0 – 18 (422 – 581)	19 – 43 (582 – 636)	44 – 70 (637 – 698)	71 – 84 (699 – 842)
6	0 – 19 (439 – 600)	20 – 46 (601 – 660)	47 – 72 (661 – 727)	73 – 84 (728 – 864)
7	0 – 20 (455 – 618)	21 – 48 (619 – 677)	49 – 73 (678 – 741)	74 – 84 (742 – 875)
8	0 – 14 (477 – 624)	15 – 41 (625 – 686)	42 – 70 (687 – 753)	71 – 84 (754 – 897)
11	0 – 15 492 – 644)	16 – 46 (645 – 714)	47 – 72 (715 – 780)	73 – 84 (781 – 916)

Data file 5/19/08 for raw score to sale score conversion.

Table 4.3 presents the raw scores and corresponding scale scores assigned to each achievement level for the spring 2008 *Dakota STEP* Science assessment.

Table 4.3: Raw Score (Scale Score) Ranges for Dakota STEP Science

Grade	Below Basic Raw score (Scale score)	Basic Raw score (Scale score)	Proficient Raw score (Scale score)	Advanced Raw score (Scale score)
5	0 – 23 (399 – 573)	24 – 36 (574 – 604)	37 – 59 (605 – 669)	60 – 70 (670 – 802)
8	0 – 21 (398 – 568)	22 – 38 (569 – 610)	39 – 60 (611 – 674)	61 – 70 (675 – 803)
11	0 – 23 (394 – 566)	24 – 45 (567 – 609)	46 – 71 (610 – 668)	72 – 84 (669 – 808)

Data file 5/19/08 for raw score to sale score conversion.

CHAPTER 5: REPORTS

The *Dakota STEP* student answer documents are scored by Pearson immediately after the receipt of test materials from South Dakota test coordinators. The multiple-choice test questions are scored by machine.

5.1 Scores Reported

Standards-Based Scores

The final student data file was provided to the SD DOE and uploaded to a secure website. District and school personnel could create reports, graphs, and/or external data files with tools for querying, computation, and disaggregation. Examples of the standards-based Student and Roster reports are included in Appendix C.

The Student report provides a summary of individual student results by content area and reporting category.

- Raw Score (total number correct) and Max Score (total points possible) for each content area and its respective reporting categories
- Percent Correct for each content area and its respective reporting categories
- Achievement Level for each content area

The Roster report provides a list of student results for a selected group of students. The following results are presented by student for each content area:

- Total Raw Score
- Scale Score
- Achievement Level
- Number of items correct by content reporting category

The ***Scale Score*** allows for comparison of test scores across years. Whenever a new version of a test that is built to the same test blueprint is introduced, there will likely be some variations in the difficulty of items from previous versions. Scale scores control for these small variations in difficulty and allow comparisons of the achievement levels of students in different test years. Because test items change, raw scores (i.e., number or percent correct scores) will not always have the same meaning or represent the same level of proficiency. Without equating, each administration of a test with different items would lead to a new reporting scale independent of that used previously. It would still be possible to measure relative achievement, but it would not be possible to indicate growth across years for schools, districts, or the state. Chapter 6 presents more detailed information on the equating process and calculation of scale scores for the *Dakota STEP*.

The ***Performance (Achievement) Level*** is one of the four levels (Below Basic, Basic, Proficient, Advanced) described in the content standards, bound by the cut points recommended by the standard-setting committees and finalized by the SD DOE. Chapter 4 provides information regarding the establishment of cut scores and achievement levels.

5.2 Appropriate Uses of Scores

Cautions in Interpreting Scores

There are many factors to be considered when looking at test performance. First, it is possible that the student has not yet been exposed to all of the content covered by the test. More broadly, however, it is important to remember that achievement in school and on the test may be affected by one or more of these non-curricular factors:

Student/Home Factors:

The student's general health and physical condition.
The stability of the student's home environment.
The support and/or help the student is able to receive at home.
The student's age relative to other students in the grade.
The ability of the student to get along with others.
The student's school attendance record.
The student's interest in school.
The student's study and work habits.

School Factors:

The appropriateness of the level of instruction for the student.
The amount of time available for instruction.
The expectations set for the student.
The appropriateness of school settings/groupings for the student.
The appropriateness of instructional materials and methods for the student.

None of these factors invalidate the test results; the *Dakota STEP* measures achievement, regardless of the reasons and circumstances behind it. Consideration of these factors is important, however, when using test results to make judgments about the quality of a student's classroom instruction, the student's subject mastery, or when considering measures to improve a student's achievement.

The individual student report provides a snapshot of student performance. This one-page report, however, only yields achievement levels. To appropriately use the standards-based scores, the Standards-Based Report should be the driving mechanism to evaluate curriculum and individual student achievement.

CHAPTER 6: CALIBRATION, LINKING (EQUATING), AND SCALING

This chapter describes the three-step process of *calibration*, *linking*, and *scaling* used for the *Dakota STEP* assessment to transform performance on each test into a scale score. This process was carried out in 2008 for the *Dakota STEP* Reading, Mathematics, and Science assessments.

The process of *calibration*, *linking*, and *scaling* utilizes *item response theory* (IRT), which is based on the idea that the characteristics of individual items can be used in conjunction with student item responses to produce estimates of students' levels of achievement. The particular model used in calibration and linking for the *Dakota STEP* assessment is the Rasch model (Rasch, 1960).

6.1 Calibration

In calibration, students' responses to each item are analyzed to derive estimates of item difficulty and student achievement. The Rasch IRT model describes, for each item, the probability of a person with achievement level θ ($P(\theta)$) answering the item correctly. $P(\theta)$ is a function based on a student's achievement level (θ) and the difficulty of the item (δ):

$$P(\theta) = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}},$$

e is a constant equal to 2.718.

Item difficulty and student achievement are estimated using the joint maximum likelihood (JML) method by entering student responses into the WINSTEPS computer program (Linacre, 2003). The WINSTEPS program seeks the best-fitting Rasch model estimates of item difficulty and student achievement level simultaneously.

WINSTEPS provides two kinds of fit statistics, the outfit and infit statistics, which indicate how accurately or predictably the data fit the model.

Infit: an information weighted fit statistic that focuses on the overall performance of an item or person, i.e., the information-weighted average of the squared standardized deviation of observed performance from expected performance. The statistic plotted and tabled by Rasch is this mean square normalized (http://www.rasch.org/rmt/glossary.htm).

Outfit: an outlier sensitive fit statistic that picks up rare events that have occurred in an unexpected way. It is the average of the squared standardized deviations of the observed performance from the expected performance. Rasch plots and tables use the normalized unweighted mean squares so that the graphs are symmetrically centered on zero (http://www.rasch.org/rmt/glossary.htm).

The expected value of the fit statistics is 1.0. Values that are less than 1.0 indicate that observations are too predictable (i.e., data overfit the model), while values greater than 1.0 indicate unpredictability (i.e., data underfit the model).

Appendix D shows the Rasch item statistics, namely Rasch item difficulty, standard error (SE) of difficulty parameter, and infit and outfit statistics. The WINSTEPS software (Linacre, 2003) was used for the *Dakota STEP* assessment's item and ability calibrations. All students with valid scores were included in the calibration (note the WINSTEPS software excludes extreme scores in the calibration). The reason that the sample sizes in Appendix D and Tables 6.1–6.3 are different is because the actual data analyses for this technical report only included public school students with valid test scores, while all of the students with valid scores excluding extreme scores were included in the calibration.

In IRT, a test characteristic curve (TCC) represents the sum of the item characteristic curve (ICC) of each item on the test. The ICCs are calculated based on the probability of item responses for the scored items on the scale score continuum. The TCCs for the *Dakota STEP* assessment was computed this way by grade.

Test reliability in IRT can be represented by the test information function (TIF) and conditional standard error. The TIF is the sum of the item information based on all the items in the test. Higher values of test information indicate better reliability. The TIF was calculated for the *Dakota STEP* assessment for each grade. Conditional standard error, on the other hand, is the inverse of the test information. In the Rasch model (Rasch, 1960), item information is:

$$I_i(\theta) = P_i(\theta)(1 - P_i(\theta)).$$

Test information function and standard error can be written as, respectively,

$$TIF(\theta) = \sum_{i=1}^{I} I_i(\theta), \text{ and}$$

$$SE(\theta) = \frac{1}{\sqrt{TIF(\theta)}}.$$

TCC, TIF, and conditional standard error were graphed for each grade and content area. In the graphs, three lines were used to show the Basic, Proficient, and Advanced achievement levels. Appendix E displays the TCC, TIF, and conditional standard error charts for the *Dakota STEP* assessment for each of the grades.

6.2 Linking Procedure

Since all students take the same operational items within a given year, there is no need for equating within the same test year. For equating from year to year, the linking items (anchor items, equating items) are selected from the operational items used the year before. The operational items are calibrated based on the Rasch IRT model (Rasch, 1960) using the WINSTEPS computer program (Linacre, 2003). Item parameters of linking items are fixed to the values of the existing parameters (e.g., the values obtained in spring 07). The operational items that are not selected as the linking items are freed and thus are calibrated based on the data obtained in that year (e.g., spring 08).

Linking items are selected from operational items administered in the previous year that meet the following guidelines (to the greatest extent possible):

1. The number of linking items should be at least 30 percent of the total number of items in each content area (e.g., Reading, Mathematics, and Science).
2. The linking items should appear in approximately the same item sequence position (within one third of the test length) as they appeared in the previous-year's test form.
3. The linking items should represent a miniature version of the test in terms of content and statistical properties.
    - The percentage distribution of the points from linking items across different reporting categories is similar to that of the whole test.
    - Average item difficulty is similar to that of the whole test. The range of item difficulty may be narrower than the whole test because extremely difficult or extremely easy items may not be selected as linking items.
4. The linking items should have strong statistical properties.
    - Linking items should have good discrimination (e.g., point-biserial correlation > .3).
    - Infit and Outfit statistics should be close to 1.0 (e.g., within the range of 0.7 to 1.3).
5. There should not be any change in the linking item from one administration to the other in terms of wording, formatting, or any other characteristics.
6. Items that have only been field tested should not be used as linking items; for an item to be included in the linking set it must have been an operational item in the previous year. This means that all linking items have been previously calibrated as operational items.
7. Items with less desirable statistics may be used as linking items only when the requirement of the test blueprint cannot be fulfilled without these items.

In addition to meeting the above criteria, the linking items are screened for performance between years based on the visual inspection of the estimated Rasch item difficulty plots (e.g., comparison between the Rasch difficulty estimates from 2007 and 2008), and displacement values obtained from the WINSTEP software (Linacre, 2003). WINSTEPS computes the displacement value in each calibration run when an item is anchored or fixed to a specified value. Displacement is the approximation of the difference between the estimate and the statistically better estimate that would result from the best fit of data to the model (www.winsteps.com, 2005). For the *Dakota STEP* assessment, items with a displacement value larger than 0.30 are deleted from the linking item set and no more than 20% of the linking items can be dropped from the linking set.

Within this linking item selection system, each content area (i.e., Reading, Mathematics, and Science) presents some unique issues as described below.

Reading

The standards-based items from 2007 were re-administered in 2008. Administering the same core form in 2008 as in 2007 means that the raw score-to-scale score tables created in 2007 are still valid for 2008 use. However, one item was replaced because the item's statistics were not optimal in grades 4, 5, and 6, respectively. Therefore, for grades 4, 5, and 6, all the items that were the same as in 2007 were used as the linking items (no more items were deleted following the linking item screening criterion based on the WINSTEP displacement values) to calibrate the 2008 Reading test. New raw score-to-scale score tables were created for those three grades.

Grades 3, 7, 8, and 11 used the same item and ability parameter estimates from 2007; and thus the same raw score-to-scale score conversions were used for those grades.

Mathematics

In order to equate the 2008 test to the 2007 test, a linking set was constructed by selecting items from each of the mathematics content indicators (reporting categories). Twenty-eight items were selected as the linking items for the 2008 test based on the item statistics obtained in 2007 administration. Following the linking screening criteria, 2, 3, 5, 1, 0, 2, and 0 linking items were unanchored for grades 3–8, and 11, respectively.

Science

In order to equate the 2008 test to the 2007 test, a linking set was constructed by selecting items from each of the science content indicators (reporting categories). The Science assessment had 24, 24, and 28 linking items for grades 5, 8, and 11, respectively. Following the linking item screening criteria, 1, 2, and 4 linking items were not anchored for grades 5, 8, and 11, respectively.

Tables 6.1–6.2 contain summary statistics for the items that were initially selected as the linking items based on the 2007 data for the Mathematics and Science tests. Tables 6.3–6.4 present summary statistics for the items that were actually used as the final linking items to link the 2008 Mathematics and Science tests following the screening criteria. The summary item statistics listed in Tables 6.3–6.4 were from the 2008 data.

Characteristics of *Dakota STEP* Item Parameters

The *Dakota STEP* Reading and Mathematics (prior to 2008) assessments were augmented tests that included the Stanford 10 items. The Stanford 10 tests are on a vertical scale, thus the *Dakota STEP* Reading and Mathematics items were calibrated onto this vertical scale with a grade/level constant adjusted to item parameters. Science, on the other hand, is not an augmented test and is not on a vertical scale. The Stanford 10 vertical scale was originally created by administering the on-grade level and one grade level lower test to the students in Kindergarten through grade 11 (Pearson, 2002). The level constants were therefore obtained. Although from 2008, the *Dakota STEP* Mathematics became a totally customer designed test, the item difficulty parameters of the link items came from 2007 in which the Stanford 10 test level constants were applied. The reason that the item difficulty parameters in Appendix D do not center on zero is due to this added level constant for the Reading and Mathematics assessments.

Table 6.1: Mathematics Linking Item Statistics from 2007 Administration

Grade	Item Statistics	Number of Items	Mean	Standard Deviation	Minimum	Maximum
3	P-value	28	0.72	0.12	0.48	0.90
	Point-biserial	28	0.40	0.06	0.30	0.52
	Rasch difficulty	28	-0.37	0.74	-1.75	1.00
	Infit	28	1.01	0.07	0.89	1.13
	Outfit	28	1.02	0.15	0.72	1.27
4	P-value	28	0.62	0.08	0.46	0.74
	Point-biserial	28	0.42	0.06	0.31	0.55
	Rasch difficulty	28	0.66	0.40	0.00	1.46
	Infit	28	0.99	0.06	0.86	1.11
	Outfit	28	0.98	0.09	0.77	1.17
5	P-value	28	0.63	0.12	0.43	0.89
	Point-biserial	28	0.44	0.06	0.32	0.57
	Rasch difficulty	28	1.10	0.69	-0.68	2.12
	Infit	28	0.95	0.06	0.83	1.07
	Outfit	28	0.93	0.10	0.74	1.14
6	P-value	28	0.63	0.13	0.45	0.90
	Point-biserial	28	0.43	0.07	0.32	0.56
	Rasch difficulty	28	1.66	0.79	-0.21	2.64
	Infit	28	0.99	0.08	0.85	1.14
	Outfit	28	0.98	0.12	0.73	1.20
7	P-value	28	0.67	0.12	0.49	0.88
	Point-biserial	28	0.47	0.07	0.34	0.62
	Rasch difficulty	28	1.78	0.76	0.34	2.83
	Infit	28	0.95	0.09	0.77	1.09
	Outfit	28	0.91	0.15	0.64	1.19
8	P-value	28	0.58	0.12	0.36	0.84
	Point-biserial	28	0.43	0.07	0.32	0.58
	Rasch difficulty	28	2.63	0.67	1.08	3.72
	Infit	28	0.98	0.07	0.83	1.11
	Outfit	28	0.97	0.11	0.76	1.16
11	P-value	28	0.62	0.13	0.40	0.86
	Point-biserial	28	0.47	0.08	0.28	0.58
	Rasch difficulty	28	2.97	0.71	1.47	4.13
	Infit	28	0.94	0.09	0.84	1.15
	Outfit	28	0.91	0.14	0.69	1.27

Table 6.2: Science Linking Item Statistics from 2007 Administration

Grade	Item Statistics	Number of Items	Mean	Standard Deviation	Minimum	Maximum
5	P-value	24	0.65	0.14	0.35	0.81
	Point-biserial	24	0.39	0.10	0.10	0.49
	Rasch difficulty	24	0.04	0.70	-0.88	1.52
	Infit	24	0.98	0.08	0.88	1.25
	Outfit	24	0.97	0.14	0.77	1.40
8	P-value	24	0.63	0.11	0.43	0.77
	Point-biserial	24	0.41	0.07	0.30	0.51
	Rasch difficulty	24	0.08	0.57	-0.71	1.06
	Infit	24	0.99	0.08	0.88	1.11
	Outfit	24	0.98	0.12	0.77	1.16
11	P-value	28	0.64	0.12	0.42	0.85
	Point-biserial	28	0.41	0.10	0.21	0.58
	Rasch difficulty	28	-0.03	0.63	-1.30	1.07
	Infit	28	0.98	0.10	0.82	1.20
	Outfit	28	0.98	0.18	0.68	1.42

Table 6.3: Mathematics Linking Item Statistics from 2008 Administration

Grade	Item Statistics	Number of Items	Mean	Standard Deviation	Minimum	Maximum
3	P-value	26	0.71	0.11	0.50	0.86
	Point-biserial	26	0.41	0.06	0.33	0.52
	Rasch difficulty	26	-0.28	0.69	-1.52	1.00
	Infit	26	1.03	0.10	0.82	1.21
	Outfit	26	1.07	0.19	0.65	1.51
4	P-value	25	0.65	0.07	0.49	0.78
	Point-biserial	25	0.43	0.06	0.34	0.55
	Rasch difficulty	25	0.59	0.37	0.00	1.46
	Infit	25	1.01	0.08	0.86	1.12
	Outfit	25	1.00	0.11	0.81	1.15
5	P-value	23	0.63	0.13	0.42	0.88
	Point-biserial	23	0.45	0.06	0.29	0.58
	Rasch difficulty	23	1.08	0.73	-0.68	2.12
	Infit	23	0.95	0.08	0.82	1.12
	Outfit	23	0.92	0.12	0.72	1.21
6	P-value	27	0.65	0.13	0.46	0.89
	Point-biserial	27	0.43	0.07	0.31	0.56
	Rasch difficulty	27	1.64	0.79	-0.22	2.64
	Infit	27	1.00	0.09	0.86	1.16
	Outfit	27	0.99	0.14	0.72	1.28
7	P-value	28	0.70	0.13	0.51	0.91
	Point-biserial	28	0.47	0.08	0.33	0.63
	Rasch difficulty	28	1.78	0.76	0.34	2.83
	Infit	28	0.95	0.11	0.74	1.20
	Outfit	28	0.91	0.18	0.56	1.32
8	P-value	26	0.59	0.12	0.40	0.88
	Point-biserial	26	0.44	0.07	0.30	0.59
	Rasch difficulty	26	2.68	0.61	1.09	3.72
	Infit	26	1.00	0.10	0.78	1.19
	Outfit	26	1.01	0.16	0.66	1.26
11	P-value	28	0.63	0.13	0.38	0.88
	Point-biserial	28	0.47	0.09	0.30	0.60
	Rasch difficulty	28	2.97	0.71	1.48	4.13
	Infit	28	0.96	0.11	0.82	1.19
	Outfit	28	0.94	0.17	0.74	1.33

Table 6.4: Science Linking Item Statistics from 2008 Administration

Grade	Item Statistics	Number of Items	Mean	Standard Deviation	Minimum	Maximum
5	P-value	23	0.65	0.15	0.34	0.85
	Point-biserial	23	0.39	0.09	0.12	0.51
	Rasch difficulty	23	0.06	0.70	-0.88	1.52
	Infit	23	0.96	0.11	0.78	1.24
	Outfit	23	0.95	0.16	0.70	1.37
8	P-value	22	0.64	0.12	0.42	0.83
	Point-biserial	22	0.41	0.07	0.30	0.52
	Rasch difficulty	22	0.10	0.57	-0.71	1.06
	Infit	22	0.99	0.09	0.77	1.11
	Outfit	22	0.98	0.12	0.68	1.18
11	P-value	24	0.64	0.13	0.40	0.88
	Point-biserial	24	0.41	0.09	0.26	0.56
	Rasch difficulty	24	0.00	0.65	-1.30	1.07
	Infit	24	0.97	0.09	0.81	1.15
	Outfit	24	0.96	0.16	0.65	1.26

6.3 Scale Score Creation

After calibration and linking, students' ability estimates were used as scores on the θ scale. Estimates of student ability were placed on the scale score (SS) metric using the following linear transformation for each of the grades of the *Dakota STEP* Reading, Mathematics, and Science assessments, respectively.

$$SS = 35*\theta + 600$$

Raw Score (RS) to Scale Score (SS) conversion tables and IRT standard error (SE) for the 2008 *Dakota STEP* Reading, Mathematics, and Science assessments are provided in Appendix F. This IRT SE is different from the classical test theory (CTT) SEM shown in Chapter 7. The IRT SE varies depending on the individual's scale score, while the CTT SEM, based on the raw score total, remains a constant. Appendix F also shows the upper (SS+1SE) and the lower (SS-1SE) bounds for an approximate 68% confidence interval for each SS and the achievement level.

Tables 6.5–6.7 provide scale score summaries for the 2008 *Dakota STEP* Reading, Mathematics, and Science assessments. As indicated earlier, the Reading and Mathematics tests were linked onto a vertical scale which is why the mean of the scale score increased from grade to grade, while the Science assessment is not on a vertical scale, and therefore the mean scale score did not increase across grades. Table 6.8 provides the standard error at each achievement classification cut point in the scale score metric. The raw score and scale score frequency distributions for all grades and subjects are provided in Appendix G.

The Rasch item difficulty was scaled using the same linear transformation as listed above to obtain scaled item difficulty. The scaled item difficulty was then compared to the corresponding

cut points to assign items to different achievement levels. Appendix H shows scaled Rasch item difficulty parameter summary.

Table 6.5: Scale Score Summary for Dakota STEP Reading

Grade	N-count	Mean	Standard Deviation
3	8944	636.45	34.62
4	8882	649.22	34.46
5	9004	655.90	32.31
6	9035	673.10	35.69
7	9141	683.77	33.32
8	9305	693.70	29.84
11	8313	733.75	37.31

Data file 07/30/08 public schools only.

Table 6.6: Scale Score Summary for Dakota STEP Mathematics

Grade	N-count	Mean	Standard Deviation
3	8963	634.60	42.65
4	8896	650.01	38.41
5	9019	665.46	37.66
6	9043	689.58	38.99
7	9161	704.93	41.67
8	9325	715.13	40.72
11	8309	732.89	40.50

Data file 07/30/08 public schools only.

Table 6.7: Scale Score Summary for Dakota STEP Science

Grade	N-count	Mean	Standard Deviation
5	9003	630.88	31.90
8	9306	631.69	34.71
11	8310	627.03	32.96

Data file 07/30/08 public schools only.

Table 6.8: SE of SS Cut Points by Achievement Level

Grade	Achievement Level	Reading Minimum SS cut	Reading SS SE	Mathematics Minimum SS cut	Mathematics SS SE	Science Minimum SS cut	Science SS SE
3	Basic	498	19	530	10	- -	- -
	Proficient	595	11	600	9	- -	- -
	Advanced	664	12	668	14	- -	- -
4	Basic	525	16	548	11	- -	- -
	Proficient	607	10	618	8	- -	- -
	Advanced	668	12	684	12	- -	- -
5	Basic	544	16	582	10	574	9
	Proficient	621	10	637	8	605	9
	Advanced	678	12	699	10	670	12
6	Basic	556	15	601	10	- -	- -
	Proficient	636	10	661	8	- -	- -
	Advanced	692	12	728	12	- -	- -
7	Basic	565	17	619	10	- -	- -
	Proficient	651	10	678	8	- -	- -
	Advanced	709	12	742	12	- -	- -
8	Basic	574	17	625	11	569	10
	Proficient	668	10	687	8	611	9
	Advanced	728	13	754	10	675	13
11	Basic	625	16	645	11	567	9
	Proficient	719	12	715	8	610	8
	Advanced	780	19	781	11	669	11

Reading cut points were obtained in 2005, Mathematics in 2006, and Science in 2007.

CHAPTER 7: PSYCHOMETRIC CHARACTERISTICS: CLASSIC ITEM STATISTICS, RELIABILITY, AND DIF

The reliability of a test reflects the degree to which test scores are free from errors of measurement that arise from various sources. Test reliability indicates the extent to which differences in test scores reflect real differences in the construct being measured across some variation in one or more factors, such as time or specific test items used. Different coefficients can be distinguished accordingly. This chapter presents measures of reliability for the *Dakota STEP* Reading, Mathematics, and Science tests as well as measures of consistency and accuracy for classification of students into achievement levels.

The classical test theory (CTT) item statistic tables of *p*-value and point-biserial correlation for all the grades and subjects for the 2008 *Dakota STEP* are provided separately in Appendix I. Tables 7.1–7.2 presents summaries of the *p*-values, and the number and percentage of items that are classified to different ranges according to each item's *p*-value. Similarly, Tables 7.3–7.4 contains the point-biserials, and the number and percentage of items that fall into different ranges based on the point-biserial correlation coefficients. These CTT item statistics are included to document the properties of the test items.

The *p-value* is the proportion of students that answer an item correctly. A high *p*-value means an item is easy; and a low *p*-value means an item is hard. Table 7.1 shows that the average *p*-value ranges from 0.62 (grade 8) to 0.68 (grade 11) for the Reading test; from 0.64 (grades 8 and 11) to 0.75 (grade 3) for the Mathematics assessment; and from 0.63 (grade 11) to 0.66 (grade 5) for the Science test. The item *p*-value of the Reading test ranges from 0.16 to 0.96. It is from 0.25 to 0.95 for the Mathematics test; and ranges from 0.14 to 0.96 for the Science assessment. Table 7.2 indicates that majority of the items (over 90%) have the *p*-value in the range of 0.2 to 0.9 while a few items have high (>= 0.9) or low (< 0.1) item difficulty.

The *point-biserial correlation* is an index of the association between item score and total test score. It shows the ability of an item to discriminate between low- and high-performing students. An item with a high point-biserial correlation discriminates the low- and the high- performing students better than an item with a low point-biserial correlation. Table 7.3 contains the summary statistics of the point-biserial correlation coefficients, while Table 7.4 shows the number and percentage of items falling into different ranges. The average point-biserial correlation coefficient ranges from 0.36 to 0.40, from 0.40 to 0.43, and from 0.36 to 0.39 for the Reading, Mathematics, and Science assessments, respectively. The item point-biserial correlation ranges from 0.06 to 0.60, 0.06 to 0.62, and 0.10 to 0.55 for the Reading, Mathematics, and Science tests, respectively. Table 7.4 shows that there is one item with the point-biserial value less than 0.10 for each of the subjects. Fourteen percent to 29% of the items have point-biserial values in the 0.1 to 0.3 range for the Reading test. The percentage of the items with point-biserial values in the 0.1 to 0.3 range for the Mathematics and Science tests ranges from 5% to 13%, and 17% to 26%, respectively. The majority of the items yield a point-biserial value larger than 0.3. Specifically, these items range from 71% to 88%, 87% to 95%, and 73% to 83% for the Reading, Mathematics, and Science assessments, respectively.

Table 7.1: Reading, Mathematics, and Science Item P-Value Summary

Test	Grade	Mean	Standard Deviation	Minimum	Q1[1]	Median	Q3[2]	Maximum
Reading	**3**	0.65	0.16	0.29	0.52	0.65	0.77	0.94
	4	0.66	0.15	0.29	0.55	0.66	0.77	0.96
	5	0.64	0.13	0.25	0.60	0.64	0.75	0.92
	6	0.66	0.18	0.16	0.51	0.71	0.82	0.92
	7	0.64	0.14	0.36	0.54	0.62	0.75	0.91
	8	0.62	0.14	0.29	0.52	0.65	0.73	0.96
	11	0.68	0.17	0.26	0.57	0.71	0.83	0.90
Mathematics	**3**	0.75	0.13	0.42	0.64	0.77	0.86	0.95
	4	0.67	0.14	0.36	0.58	0.69	0.78	0.94
	5	0.66	0.15	0.34	0.55	0.65	0.78	0.95
	6	0.68	0.16	0.28	0.58	0.70	0.83	0.94
	7	0.69	0.14	0.36	0.59	0.70	0.79	0.92
	8	0.64	0.16	0.29	0.54	0.65	0.75	0.91
	11	0.64	0.16	0.25	0.50	0.64	0.76	0.95
Science	**5**	0.66	0.15	0.31	0.56	0.71	0.78	0.91
	8	0.65	0.15	0.36	0.53	0.66	0.78	0.94
	11	0.63	0.14	0.35	0.52	0.65	0.75	0.96

[1] Q1 is the first quartile.
[2] Q3 is the third quartile.
Data file 07/30/08 public schools only.

Table 7.2: Reading, Mathematics, and Science Item P-Value Classification

Reading		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
***P*-value < 0.1**	No. of Items	0	0	0	0	0	0	0
	%	0	0	0	0	0	0	0
0.1 <= P-value < 0.2	No. of Items	0	0	0	1	0	0	0
	%	0	0	0	2	0	0	0
0.2 <= *P*-value < 0.5	No. of Items	12	8	7	13	11	11	6
	%	21	14	13	23	20	20	15
0.5 <= *P*-value < 0.7	No. of Items	23	25	32	12	25	28	14
	%	41	45	57	21	45	50	35
0.7 <= *P*-value < 0.9	No. of Items	17	20	16	29	18	16	17
	%	30	36	29	52	32	29	43
***P*-value >= 0.9**	No. of Items	4	3	1	1	2	1	3
	%	7	5	2	2	4	3	8
Mathematics		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
***P*-value < 0.1**	No. of Items	0	0	0	0	0	0	0
	%	0	0	0	0	0	0	0
0.1 <= *P*-value < 0.2	No. of Items	0	0	0	0	0	0	0
	%	0	0	0	0	0	0	0
0.2 <= *P*-value < 0.5	No. of Items	3	14	14	13	10	16	20
	%	4	17	17	15	12	19	24
0.5 <= *P*-value < 0.7	No. of Items	24	29	35	29	32	37	33
	%	29	35	42	35	38	44	39
0.7 <= *P*-value < 0.9	No. of Items	50	40	32	39	37	30	28
	%	60	48	38	46	44	36	33
***P*-value >= 0.9**	No. of Items	7	1	3	3	5	1	3
	%	8	1	4	4	6	1	4
Science		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
***P*-value < 0.1**	No. of Items	- -	- -	0	- -	- -	0	0
	%	- -	- -	0	- -	- -	0	0
0.1 <= *P*-value < 0.2	No. of Items	- -	- -	0	- -	- -	0	0
	%	- -	- -	0	- -	- -	0	0
0.2 <= *P*-value < 0.5	No. of Items	- -	- -	10	- -	- -	14	19
	%	- -	- -	14	- -	- -	20	23
0.5 <= *P*-value < 0.7	No. of Items	- -	- -	24	- -	- -	28	38
	%	- -	- -	34	- -	- -	40	45
0.7 <= *P*-value < 0.9	No. of Items	- -	- -	33	- -	- -	25	26
	%	- -	- -	47	- -	- -	36	31
***P*-value >= 0.9**	No. of Items	- -	- -	3	- -	- -	3	1
	%	- -	- -	4	- -	- -	4	1

Data file 07/30/08 public schools only.

Table 7.3: Reading, Mathematics, and Science Item Point-Biserial Summary

Test	Grade	Mean	Standard Deviation	Minimum	Q1[1]	Median	Q3[2]	Maximum
Reading	3	0.38	0.09	0.17	0.32	0.38	0.45	0.55
	4	0.38	0.09	0.19	0.32	0.40	0.46	0.53
	5	0.37	0.08	0.15	0.33	0.38	0.43	0.53
	6	0.40	0.10	0.06	0.35	0.41	0.48	0.60
	7	0.38	0.08	0.22	0.32	0.38	0.44	0.53
	8	0.36	0.11	0.10	0.29	0.37	0.44	0.54
	11	0.40	0.08	0.23	0.34	0.40	0.46	0.52
Mathematics	3	0.42	0.07	0.23	0.37	0.42	0.47	0.54
	4	0.41	0.08	0.20	0.36	0.41	0.47	0.57
	5	0.40	0.08	0.16	0.34	0.42	0.46	0.57
	6	0.41	0.08	0.06	0.35	0.42	0.47	0.56
	7	0.43	0.09	0.20	0.38	0.43	0.48	0.61
	8	0.42	0.09	0.12	0.37	0.42	0.50	0.59
	11	0.42	0.09	0.13	0.36	0.42	0.49	0.62
Science	5	0.36	0.09	0.10	0.30	0.37	0.43	0.55
	8	0.39	0.07	0.24	0.33	0.39	0.44	0.54
	11	0.38	0.09	0.11	0.32	0.38	0.45	0.55

[1] Q1 is the first quartile.
[2] Q3 is the third quartile.
Data file 07/30/08 public schools only.

Table 7.4: Reading, Mathematics, and Science Item Point-Biserial Classification

Reading		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Point-Biserial < 0.1	No. of Items	0	0	0	1	0	0	0
	%	0	0	0	2	0	0	0
0.1 <= Point-Biserial < 0.3	No. of Items	10	13	9	9	8	16	5
	%	18	23	16	16	14	29	12
Point-Biserial >= 0.3	No. of Items	46	43	47	46	48	40	35
	%	82	77	84	82	86	71	88
Mathematics		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Point-Biserial < 0.1	No. of Items	0	0	0	1	0	0	0
	%	0	0	0	1	0	0	0
0.1 <= Point-Biserial < 0.3	No. of Items	4	7	11	6	8	5	6
	%	5	8	13	7	10	6	7
Point-Biserial >= 0.3	No. of Items	80	77	73	77	76	79	78
	%	95	92	87	92	90	94	93
Science		**Grade**						
		3	**4**	**5**	**6**	**7**	**8**	**11**
Point-Biserial < 0.1	No. of Items	- -	- -	1	- -	- -	0	0
	%	- -	- -	1	- -	- -	0	0
0.1 <= Point-Biserial < 0.3	No. of Items	- -	- -	18	- -	- -	12	17
	%	- -	- -	26	- -	- -	17	20
Point-Biserial >= 0.3	No. of Items	- -	- -	51	- -	- -	58	67
	%	- -	- -	73	- -	- -	83	80

Data file 07/30/08 public schools only.

Differential item functioning (DIF) refers to the difference in difficulty of an item between focal and reference groups after matching on total test score. For *Dakota STEP*, the Mantel-Haenszel approach for DIF was used and the evaluation of DIF severity follows the classification rules commonly referred to as the "ETS DIF Classification Rules" (Zieky, 1993). DIF classifications include A, B, and C, where A represents negligible DIF, B represents moderate DIF and C represents large DIF.

For multiple-choice items, the Mantel-Haenszel Delta DIF (MH D-DIF) statistic is used to identify items that exhibit varying degrees of DIF at a significance level of 0.05.

- Items are classified in category A if MH D-DIF is not statistically different from zero or the magnitude of the MH D-DIF values is less than one delta unit in absolute value.
- Items are classified in category C if the MH D-DIF absolute value exceeds 1.5 and is statistically significantly greater than 1.0.
- All other items are classified in category B.

The DIF analysis for the *Dakota STEP* was conducted for Male versus Female, White versus American Indian, Non-IEP versus IEP, and not eligible for Free or Reduced Priced Lunch versus Free or eligible for Reduced Priced Lunch group comparisons. The DIF analyses for individual items within each domain are presented in Appendix J. Summary results of the DIF analyses for each content domain are presented in Appendix K. Due to an insufficient number of examinees in certain subgroups (i.e., Black, Hispanic, Asian, LEP, and Migrant Status), these subgroups were not included in DIF analysis. In general, most of the items in the 2008 *Dakota STEP* tests had DIF flag of A, while a few items were classified into the B DIF category and occasionally one or two items had a C flag (see Appendix K). The B or C DIF flagged items were reviewed by content experts and no apparent reasons of flagging were discovered. Those items were used in the 2008 operational test to fulfill the test blueprint requirements.

Test-level CTT-based statistics (raw score means and standard deviations, KR20, and standard error), item counts, and examinee counts are provided in Appendix L for both the entire population of public school students and subgroups by:

1. Gender
2. Economic Status
3. IEP Status
4. LEP Status
5. Migrant Status
6. Race (Asian, Black, Hispanic, Native American, Pacific Islander, and White).

7.1 Reliability and Standard Error of Measurement of Raw Scores

Reliability indices for the 2008 *Dakota STEP* assessment were computed using KR20:

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_i (pq)_i}{\hat{\sigma}_X^2}\right)$$

k is the number of items in the test, pq is the variance of the ith item, and $\hat{\sigma}_X^2$ is the total test variance.

The overall standard error of measurement (SEM) for raw scores was computed using the formula:

$$SEM = SD\sqrt{1 - reliability}$$

SD stands for standard deviation of the total test.

The CTT SEM represents the uncertainty related to the raw score (RS), which is a constant across all raw scores. It is the standard deviation of the error score contained in the RS. Under the CTT assumptions, RS ± 1.96*SEM can be used as a 95% confidence interval for the true score for a specific examinee.

Tables 7.5–7.7 provide reliability estimates and the CTT SEM information for the *Dakota STEP* Reading, Mathematics, and Science tests. These estimates, based on raw scores, were calculated for the total set of items. The range of the raw score CTT reliabilities of the *Dakota STEP*

Reading assessment is from 0.86 to 0.90. For the Mathematics assessment, the CTT reliabilities range from 0.94 to 0.95. The Science CTT reliabilities range from 0.90 to 0.93. Mathematics has higher reliabilities than Reading across all grades.

The raw score analysis summary by subgroups is also provided in Appendix L. Note that the subgroups are presented by gender, IEP, LEP, Eligibility for Free or Reduced Priced Lunch, Migrant status, and race/ethnicity (Asian, Black, Hispanic, American Indian, Pacific Islander, and White). The statistics in the summary include: Number of students tested, number of students with valid scores, number of items, raw score maximum (max), raw score minimum (min), raw score mean, raw score standard deviation (SD), KR20 reliability, and classical test theory standard error of measurement (SEM). Appendix M contains the raw score analysis by score reporting category.

7.2 Consistency and Accuracy of Achievement Level

Using the scale score achievement level cut points, students were classified into one of four achievement levels—Below Basic, Basic, Proficient, or Advanced. Table 7.8 shows the percentage of students at different achievement levels for 2008 *Dakota STEP* Reading, Mathematics, and Science assessments.

Livingston and Lewis's method (1995) was applied to derive measures of the decision accuracy and consistency of the classifications. According to Livingston and Lewis, accuracy is "the extent to which the actual classifications of the test-takers agree with those that would be made on the basis of their true score, if their true scores could somehow be known." Consistency is "the agreement between classifications based on two non-overlapping, equally difficult forms of the test" (1995, p.180).

In addition, Livingston and Lewis point out that accuracy estimates depend on the level of "agreement between classifications based on an observable variable (scores on . . . a test) and classifications based on an unobservable variable (the test-takers' true scores)" (1995, p.189). Since these true scores are not available, Livingston and Lewis provide a method for estimating the true score distribution using a four-parameter beta function. Consistency is estimated using actual test response data to create two (artificial) parallel half forms of the test and comparing classifications on those alternate forms.

The estimates of the decision accuracy and decision consistency of the 2008 *Dakota STEP* assessment with respect to the cut points for Proficient or above versus Basic or below are provided in Table 7.9 for both Reading, Mathematics, and Science.

Table 7.5: Reading Assessment Reliability and CTT SEM

Grade	Number of Items	N-count[1]	Raw Score Mean	Raw Score SD[2]	Raw Score Min[3]	Raw Score Max[4]	KR20 Reliability	CTT SEM[5]
3	56	8944	36.32	9.47	3	56	0.89	3.15
4	56	8882	36.86	9.55	0	56	0.89	3.15
5	56	9004	36.11	9.45	0	56	0.88	3.23
6	56	9035	37.14	9.63	5	56	0.90	3.05
7	56	9141	35.90	9.71	1	56	0.89	3.21
8	56	9305	34.89	9.17	0	54	0.87	3.28
11	40	8313	27.12	6.78	1	40	0.86	2.58

[1] N-count is the number of students with valid scores.
[2] SD is the standard deviation.
[3] Min represents minimum.
[4] Max represents maximum.
[5] CTT SEM represents classical test theory standard error of measurement.
Data file 07/30/08 public schools only.

Table 7.6: Mathematics Assessment Reliability and CTT SEM

Grade	Number of Items	N-count[1]	Raw Score Mean	Raw Score SD[2]	Raw Score Min[3]	Raw Score Max[4]	KR20 Reliability	CTT SEM[5]
3	84	8963	62.58	14.42	0	84	0.94	3.49
4	84	8896	56.62	15.21	0	84	0.94	3.76
5	84	9019	55.37	14.95	0	84	0.94	3.81
6	84	9043	57.42	14.71	3	84	0.94	3.69
7	84	9161	58.20	15.71	10	84	0.95	3.66
8	84	9325	53.48	16.02	4	84	0.94	3.81
11	84	8309	53.46	16.02	0	84	0.94	3.79

[1] N-count is the number of students with valid scores.
[2] SD is the standard deviation.
[3] Min represents minimum.
[4] Max represents maximum.
[5] CTT SEM represents classical test theory standard error of measurement.
Data file 07/30/08 public schools only.

Table 7.7: Science Assessment Reliability and CTT SEM

Grade	Number of Items	N-count[1]	Raw Score Mean	Raw Score SD[2]	Raw Score Min[3]	Raw Score Max[4]	KR20 Reliability	CTT SEM[5]
5	70	9003	46.39	11.20	8	70	0.90	3.53
8	70	9306	45.84	12.04	1	70	0.92	3.52
11	84	8310	53.00	14.55	1	84	0.93	3.95

[1] N-count is the number of students with valid scores.
[2] SD is the standard deviation.
[3] Min represents minimum.
[4] Max represents maximum.
[5] CTT SEM represents classical test theory standard error of measurement.
Data file 07/30/08 public schools only.

Table 7.8: Percentage of Students at Each Achievement Level

			Percentage			
Test	**Grade**	**N-Count**	**Below Basic**	**Basic**	**Proficient**	**Advanced**
Reading	**3**	8944	0	11	66	22
	4	8882	0	11	61	29
	5	9004	0	14	62	25
	6	9035	0	16	54	30
	7	9141	0	17	62	21
	8	9305	0	20	68	12
	11	8313	0	32	59	9
Mathematics	**3**	8963	0	20	61	19
	4	8896	0	21	61	18
	5	9019	0	23	59	18
	6	9043	0	23	60	16
	7	9161	1	25	56	18
	8	9325	0	25	59	17
	11	8309	0	34	53	13
Science	**5**	9003	3	17	69	11
	8	9306	3	24	62	10
	11	8310	2	28	61	9

Data file 07/30/08 public schools only.

Table 7.9: Estimated Consistency and Accuracy for Proficient Decisions (Percentage)

	Reading		Mathematics		Science	
Grade	**Accuracy**	**Consistency**	**Accuracy**	**Consistency**	**Accuracy**	**Consistency**
3	95	93	95	93	- -	- -
4	95	93	95	93	- -	- -
5	94	92	94	92	93	91
6	95	93	94	92	- -	- -
7	94	91	95	93	- -	- -
8	93	90	94	92	93	90
11	90	85	94	91	93	90

Data file 07/30/08 public schools only.

For Reading, the accuracy of the decision to classify the students into Proficient or above versus Basic or below ranges from 90% to 95% across all grades; the consistency of the decision ranges from 85% to 93%. For Mathematics, the accuracy of the decision to classify the students into Proficient or above versus Basic or below ranges from 94% to 95%; the consistency of the decision ranges from 91% to 93%. For Science, the accuracy of the decision to classify the students into Proficient or above versus Basic or below is about 93%; the consistency of the decision ranges from 90% to 91%.

CHAPTER 8: PSYCHOMETRIC CHARACTERISTIC: VALIDITY

Validity refers to the adequacy and appropriateness of the interpretations made from assessments with regard to test scores. Of all the essential characteristics of a good test, none surpasses validity. If a test score is not valid for the purpose used, it has little or no value. Validity is specific. That is, a test score may be valid for one purpose and not others.

Some key aspects of validity are as follows:

- Validity is concerned with the general question, "To what extent will this assessment information or test score help me make appropriate decisions?"
- Validity refers to the decisions that are made from assessment information, not the assessment approach or test itself. It is not appropriate to say, "This assessment information is valid" unless you also say for what decisions or groups it is valid. Keep in mind that assessment information validity for one decision or group of students is not necessarily valid for others.
- Validity is a matter of degree; it does not exist on an all-or-nothing basis. Think of assessment validity in terms of categories: Highly valid, moderately valid, and invalid.
- Validity involves an overall evaluative judgment. It requires an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence. Educators also must consider assessment results in terms of the consequences of those interpretations and uses.

Test score validity refers to the degree that a test measures what it is intended to measure. Evidence that supports a test score's validity is gathered from different aspects and through different methods. The most recognized aspects are content and construct validity evidence. Content related evidence of validity refers to how well a test covers the content of interest. The process does not involve any statistical computation. Instead, it examines the correspondence between test blueprints that describe the intended content and test items. Construct related evidence of validity is comprised of an integrated argument based on theory and empirical data. Empirical data often include analyses of the internal relationships of test components to one another and analyses of the extent to which examinees engage in response processes in solving items that are related to the target domain and not to irrelevant factors.

8.1 Content Related Evidence of Validity

Evidence of content validity for the *Dakota STEP* assessment has been provided in Chapter 2, which contains descriptions of the item development and test construction process. In addition, alignment analyses are conducted to verify that the standards-based items of the assessment are aligned with the content standards for each corresponding subject and grade level.

BIACO conducts "pre-test" alignment workshops annually to review the alignment of new items with the content standards and cognitive complexity designations. The standard and cognitive complexity of each item is then captured for reference if and when the items are used on the operational forms. The criteria for alignment include:

- *Balance of Representation*: The alignment criterion for Balance of Representation is that items should be distributed among all of the standards at least to some degree.

- *Range of Knowledge*: The alignment criterion for Range of Knowledge is that 50% of the standards have to have at least one related assessment item.
- *Depth of Knowledge*: The alignment criterion for Depth of Knowledge is that at least 50% of the items corresponding to a standard have to be at or above the cognitive complexity of the standard. .

The findings of the alignment studies are utilized to identify the strengths and weaknesses of the South Dakota pool of items in order to focus item development efforts and construct operational test forms that are aligned with the content standards and cognitive complexity designations. Item Bank Alignment information is presented in Appendix N.

"Post-alignment" workshops and/or analysis of items selected for operational use provide an alignment profile of the intact core forms as they are administered. An alignment process, the web-based software known as the Web Alignment Tool (WAT) developed by the contractor, Norman L. Webb, is utilized for this analysis (http://www.wcer.wisc.edu/wat/index.aspx). The WAT was developed under the direction of Norman Webb at the Wisconsin Center for Education Research in cooperation with the Council of Chief School Officers (CCSSO) and funded by the National Science Foundation and the United States Department of Education (USDE). This process has been used to analyze curriculum standards and assessments in over 20 states to meet or to prepare to meet Title I compliance as required by the USDE. The complete alignment reports for the 2008 *Dakota STEP* assessment are provided in Appendices O-Q. These reports describe the overall alignment of each assessment form with the South Dakota Content Standards.

Reading Alignment

The 2008 *Dakota STEP* Reading assessment was the same as the 2007 tests in grades 3, 7, 8, and 11 while the other grades had one item replaced from the 2007 tests. The objective of the alignment analysis work completed by BIACO was to match items evaluated in previous alignment studies to the operational items selected for the 2008 *Dakota STEP* Reading assessment and evaluate how well these items matched the state standards in terms of content and cognitive complexity. Specifically, the 2008 *Dakota STEP* Reading assessment was evaluated in order to determine if it contained sufficient items to support decisions made at each reporting category.

Overall, the results of the BIACO alignment analyses suggest that the 2008 *Dakota STEP* Reading assessment aligned with South Dakota's content standards in terms of Balance of Representation, Range of Knowledge, Depth of Knowledge, and Categorical Concurrence. Due to the revision of the Reading Content Standards, the 2007 operational assessment was repeated in 2008. New operational forms to assess the new standards will be administered in 2009. Items for the new forms were field tested in spring 2008. Prior to the construction of the forms, these items were reviewed for alignment and cognitive complexity as well as statistical characteristics in order to select operational items that satisfy revised test blueprint requirements.

Norman Webb conducted an alignment study of the spring 2008 intact Reading test utilizing the WAT. The complete alignment report is included in Appendix O of this technical report. Results of the study suggest that the alignment for each of the assessments with the South Dakota Reading Standards for grades 3-8 and 11 was either acceptable or in need of slight improvement

to attain full alignment. The alignment was acceptable for grades 4, 5, 7, and 8 and in need of slight improvement for grades 3, 6, and 11. The main alignment issue was that the assessments did not have a sufficient number of items (N=6) that mapped to each of the four indicators to meet the criteria for categorical concurrence. This was particularly true for Indicator R.3 (Respond to Diverse Works) in grades 3-6 and 8. Reviewers did find an increased number of items that mapped to Indicator R.3 with an increase in grade. Indicator R.4 (Retrieve, Analyze and Synthesize) in grade 11 also had fewer than six items. All of the grades 7 Indicators had at least six items.

The other alignment issue for a few indicators was that the items did not match the complexity of the standards. One indicator for five of the seven assessments did not have at least 50% of the items at or above the DOK level of the corresponding standards. In particular, items corresponding to Indicator R.2 and R.4 on the grade 11 assessment were low. The range and balance were met for each indicator across all grades.

Overall, from one to seven items on the assessments would need to be added or replaced to have full alignment. Reviewers did find items in need of more work that were indicated by source-of-challenge comments or in their notes. As noted previously, items to assess the revised Reading content standards were field tested in spring 2008 and will be utilized to construct new operational forms in 2009.

Mathematics Alignment

The 2008 *Dakota STEP* Mathematics assessment was constructed utilizing items reviewed by South Dakota educators participating in alignment studies facilitated by BIACO that were conducted in the spring of 2005, spring of 2006, and the fall of 2006. The objective of the alignment analysis work completed by BIACO was to match items evaluated in previous alignment studies to the operational items selected for the 2008 *Dakota STEP* Mathematics assessment and evaluate how well these items matched the state standards in terms of content and cognitive complexity.

The results of the BIACO alignment analyses suggest that the 2008 *Dakota STEP* Mathematics assessment aligned with South Dakota's content standards in terms of Balance of Representation, Range of Knowledge, Depth of Knowledge, and Categorical Concurrence. Across grade levels, the alignment criteria were fully or partially met. For cases where the requirements were not fully met, values tended to be just below the decision threshold. The results of the alignment study will be utilized to target new item development for the spring 2009 operational assessment.

Norman Webb conducted an alignment study of the spring 2008 intact Mathematics test utilizing the WAT. The complete alignment report is included in Appendix P of this technical report. Results of the study suggest that the alignment between the mathematics assessment and standards was found to be at least acceptable for all seven grades. The grade 8 assessment and standards were fully aligned. The seven assessments, each with 84 items, had over six items for each of the five mathematics strands, enough to have an acceptable level for the Categorical Concurrence criterion for each strand for each grade. Range also was acceptable for all strands for all seven grades. For most assessments reviewers found at least one item for each of the approximately 20 underlying standards. The main alignment issue was a weak Depth-of-

Knowledge Consistency criterion for one or two strands for a grade. On six of the seven assessments fewer than 50% of the items had a DOK level that was the same or higher than the DOK level of the assigned standard for one or two strands. From one to three items would need to be replaced to remove this weakness in DOK. There were also balance weaknesses, but these were associated with a weakness in DOK or where the overemphasis of a standard was considered more a matter of preference than a major alignment issue.

Overall, only one to three items would need to be replaced on six of the assessments to attain full alignment between the assessments and the standards (see the summary table below). As such, the alignment was considered acceptable for all grades. Reviewers did make a number of comments on how individual items could be improved. Reviewers felt the assessments were too long and the emphasis for some grades was missed place while some important topics were not assessed.

Science Alignment

In October 2006, BIACO conducted an alignment study for the *Dakota STEP* Science assessment, reviewing all new items that were field tested in fall 2006. The primary task for the alignment study participants was to evaluate how well those items matched the *Dakota STEP* Science Content Standards in terms of content and cognitive complexity. The results of the fall 2006 alignment study as well as empirical data from field testing were utilized to construct the spring 2008 operational assessment.

Norman Webb conducted an alignment study of the spring 2008 intact Science test utilizing the WAT. The complete alignment report is included in Appendix Q of this technical report. Results of the study suggest that the alignment between the South Dakota science standards and assessments varied by grade. The grade 5 science standards and the assessment were acceptably aligned, requiring the replacement of only two items to attain full alignment. The main alignment issue for grade 5 was an absence of assessment items that targeted Strand 5.1 (Nature of Science). The alignment of the grade 8 science standards and the assessment required major improvement, with 13 items needing replacement to attain full alignment. Similarly, the grade 11 science standards and the assessment needed major improvement with 14 items requiring replacement for full alignment. For both grades 8 and 11, the assessment items were not as complex as the expectations in the standards. Although the proportion of the standards judged to have a DOK level 3 increased across the grades, the level of complexity of the assessment items remained fairly constant, with essentially all of the items given a DOK level 1 or 2. Aside from the DOK inconsistency, the tests for grades 8 and 11 were judged to have items that covered the content expected by the standards, with a good range of standards covered, and a good balance in the representation of those standards.

8.2 Construct Related Evidence of Validity

Internal Structure

The purpose of studying the internal structure of a test is to evaluate the extent to which test components, including subtests and items, relate to one another in theoretically or logically meaningful ways. Methods that are used to provide evidence of the internal structure of a test are usually associated with correlations.

Table 8.1 reports the correlation matrices among the *Dakota STEP* Reading, Mathematics, and Science assessments. The correlations between Reading and Mathematics ranges from .73 in grade 11 to .81 in grade 3, while the correlation between Reading and Science ranges from .80 in grade 11 to .81 in grades 5 and 8, and the correlation of Mathematics and Science is from .77 in grade 5 to .80 in grade 11.

Table 8.1: Correlation Among Dakota STEP Assessments

Grade	Test	Reading	Mathematics	Science
3	**Reading**	1.00	.81	- -
	Mathematics		1.00	- -
4	**Reading**	1.00	.79	- -
	Mathematics		1.00	- -
5	**Reading**	1.00	.77	.81
	Mathematics		1.00	.77
	Science			1.00
6	**Reading**	1.00	.78	- -
	Mathematics		1.00	- -
7	**Reading**	1.00	.76	- -
	Mathematics		1.00	- -
8	**Reading**	1.00	.75	.81
	Mathematics		1.00	.79
	Science			1.00
11	**Reading**	1.00	.73	.80
	Mathematics		1.00	.80
	Science			1.00

Data file 07/30/08 public schools only.

Dimensionality is a unique aspect of construct validity. Investigation is necessary when item response theory (IRT) is used because IRT models assume that a test measures predominantly one dimension. Although it is generally agreed that unidimensionality is a matter of degree rather than an absolute situation, there is no consensus on what defines dimensionality or on how to evaluate it. Approaches that evaluate dimensionality can be categorized into answer patterns, reliability, components and factor analysis, and latent traits. The approach consisting of component or factor analysis is the most popular method for evaluation of dimensionality (Hattie, 1985; Abedi, 1997).

Principal component analysis (PCA) was used with the *Dakota STEP* Reading, Mathematics, and Science assessments for each grade to discover how many principal components were included in each of the assessments. PCA is a data reduction method. This reduction is achieved by finding the combinations of the items to produce principal components (i.e., eigenvectors). The PCA was used to explore the dimensionality of each of the *Dakota STEP* assessment for each grade.

Lord (1980) stated that if the ratio of the first to the second eigenvalue is large and the second eigenvalue is close to other eigenvalues, the test is essentially unidimensional. Divgi (1980)

expanded Lord's idea and created an index by considering the pattern of the first three eigenvalues. The Divgi Index examines the ratio of the difference of the first and second eigenvalues over the difference of the second and third eigenvalues. A large ratio indicates a greater difference between the first and second eigenvalues, thus creating an essentially unidimensional tendency.

The results of the principal component analysis shown in Tables 8.2–8.3 and Figures 8.1–8.3 indicate that the *Dakota STEP* Reading, Mathematics, and Science assessments are each essentially unidimensional**.**

Table 8.2: First Ten Eigenvalues Obtained from the PCA

	Reading		Mathematics		Science	
Grade	**Number**	**Eigenvalue**	**Number**	**Eigenvalue**	**Number**	**Eigenvalue**
3	1	14.372	1	27.390	- -	- -
	2	1.677	2	2.223	- -	- -
	3	1.375	3	2.053	- -	- -
	4	1.274	4	1.639	- -	- -
	5	1.230	5	1.484	- -	- -
	6	1.155	6	1.437	- -	- -
	7	1.142	7	1.272	- -	- -
	8	1.073	8	1.187	- -	- -
	9	1.030	9	1.147	- -	- -
	10	1.016	10	1.056	- -	- -
4	1	14.709	1	24.679	- -	- -
	2	1.627	2	2.171	- -	- -
	3	1.313	3	1.657	- -	- -
	4	1.145	4	1.517	- -	- -
	5	1.140	5	1.444	- -	- -
	6	1.077	6	1.362	- -	- -
	7	1.065	7	1.215	- -	- -
	8	1.037	8	1.146	- -	- -
	9	1.022	9	1.105	- -	- -
	10	0.996	10	1.074	- -	- -
5	1	13.284	1	23.752	1	16.352
	2	1.538	2	2.386	2	2.011
	3	1.262	3	1.854	3	1.423
	4	1.205	4	1.615	4	1.340
	5	1.127	5	1.444	5	1.261
	6	1.095	6	1.311	6	1.251
	7	1.061	7	1.235	7	1.142
	8	1.048	8	1.136	8	1.132
	9	1.041	9	1.106	9	1.111
	10	1.008	10	1.059	10	1.084

	Reading		Mathematics		Science	
Grade	**Number**	**Eigenvalue**	**Number**	**Eigenvalue**	**Number**	**Eigenvalue**
6	1	16.519	1	25.445	- -	- -
	2	1.810	2	2.540	- -	- -
	3	1.379	3	1.858	- -	- -
	4	1.287	4	1.467	- -	- -
	5	1.202	5	1.384	- -	- -
	6	1.120	6	1.298	- -	- -
	7	1.096	7	1.230	- -	- -
	8	1.070	8	1.209	- -	- -
	9	1.010	9	1.113	- -	- -
	10	0.959	10	1.067	- -	- -
7	1	14.355	1	28.000	- -	- -
	2	1.418	2	2.181	- -	- -
	3	1.277	3	1.750	- -	- -
	4	1.174	4	1.431	- -	- -
	5	1.119	5	1.182	- -	- -
	6	1.062	6	1.136	- -	- -
	7	1.032	7	1.122	- -	- -
	8	1.018	8	1.063	- -	- -
	9	1.010	9	1.038	- -	- -
	10	0.975	10	1.018	- -	- -
8	1	12.683	1	26.064	1	18.351
	2	1.644	2	3.139	2	1.812
	3	1.273	3	1.954	3	1.466
	4	1.163	4	1.577	4	1.261
	5	1.143	5	1.346	5	1.194
	6	1.116	6	1.228	6	1.102
	7	1.071	7	1.136	7	1.089
	8	1.061	8	1.115	8	1.057
	9	1.042	9	1.053	9	1.016
	10	1.030	10	1.023	10	1.014
11	1	11.205	1	26.815	1	21.051
	2	1.668	2	2.934	2	2.186
	3	1.184	3	1.876	3	1.573
	4	1.113	4	1.406	4	1.355
	5	1.079	5	1.261	5	1.268
	6	1.024	6	1.147	6	1.211
	7	0.995	7	1.117	7	1.139
	8	0.950	8	1.105	8	1.126
	9	0.940	9	1.049	9	1.111
	10	0.913	10	1.009	10	1.077

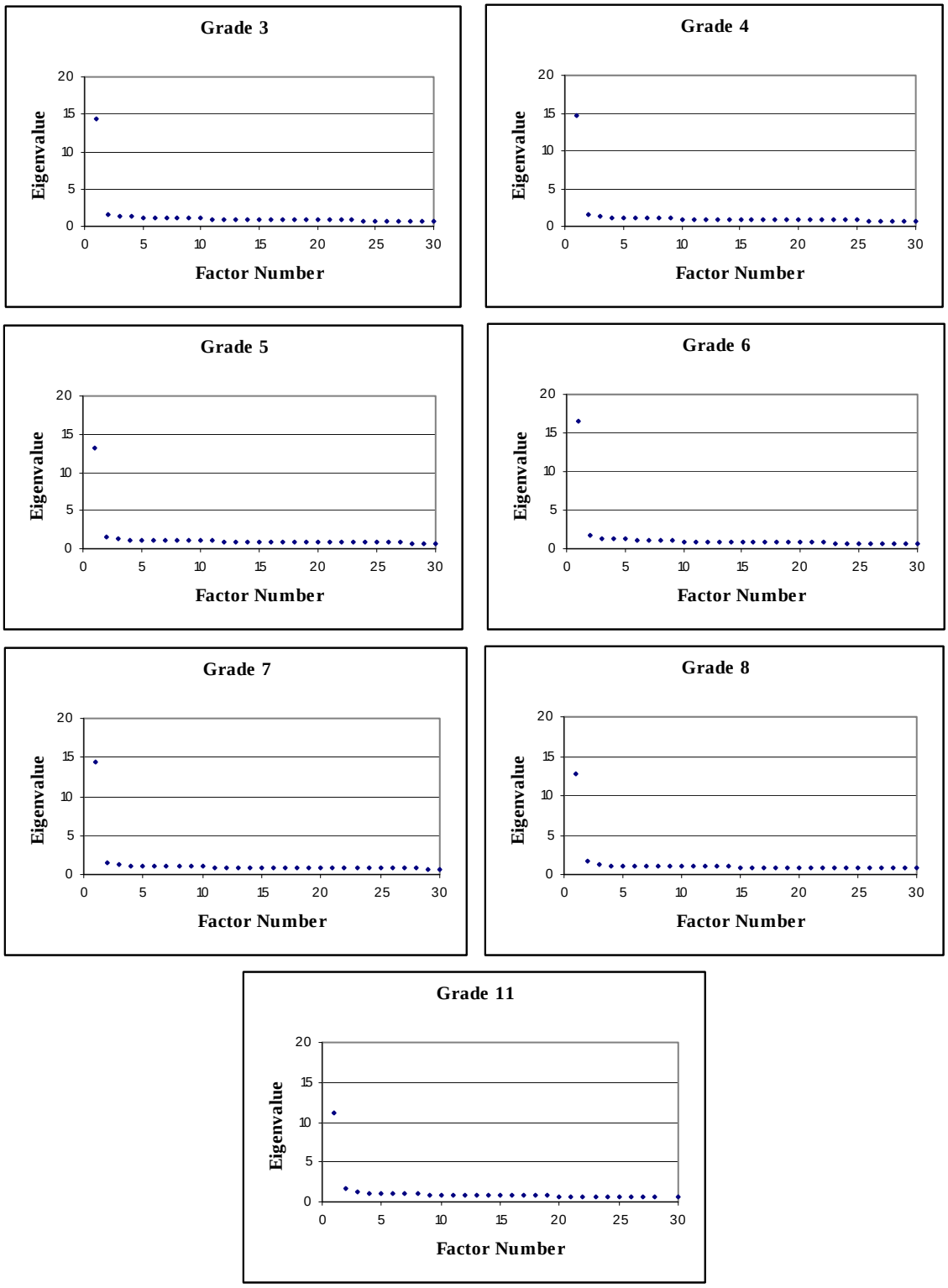
Data file 07/30/08 public schools only.

Table 8.3: Variance Explained by the First Three Principal Components

		Principal Component		
Test	**Grade**	**1**	**2**	**3**
Reading	3	14.37	1.68	1.37
	4	14.71	1.63	1.31
	5	13.28	1.54	1.26
	6	16.52	1.81	1.38
	7	14.36	1.42	1.28
	8	12.68	1.64	1.27
	11	11.20	1.67	1.18
Mathematics	3	27.39	2.22	2.05
	4	24.68	2.17	1.66
	5	23.75	2.39	1.85
	6	25.44	2.54	1.86
	7	28.00	2.18	1.75
	8	26.06	3.14	1.95
	11	26.81	2.93	1.88
Science	5	16.35	2.01	1.42
	8	18.35	1.81	1.47
	11	21.05	2.19	1.57

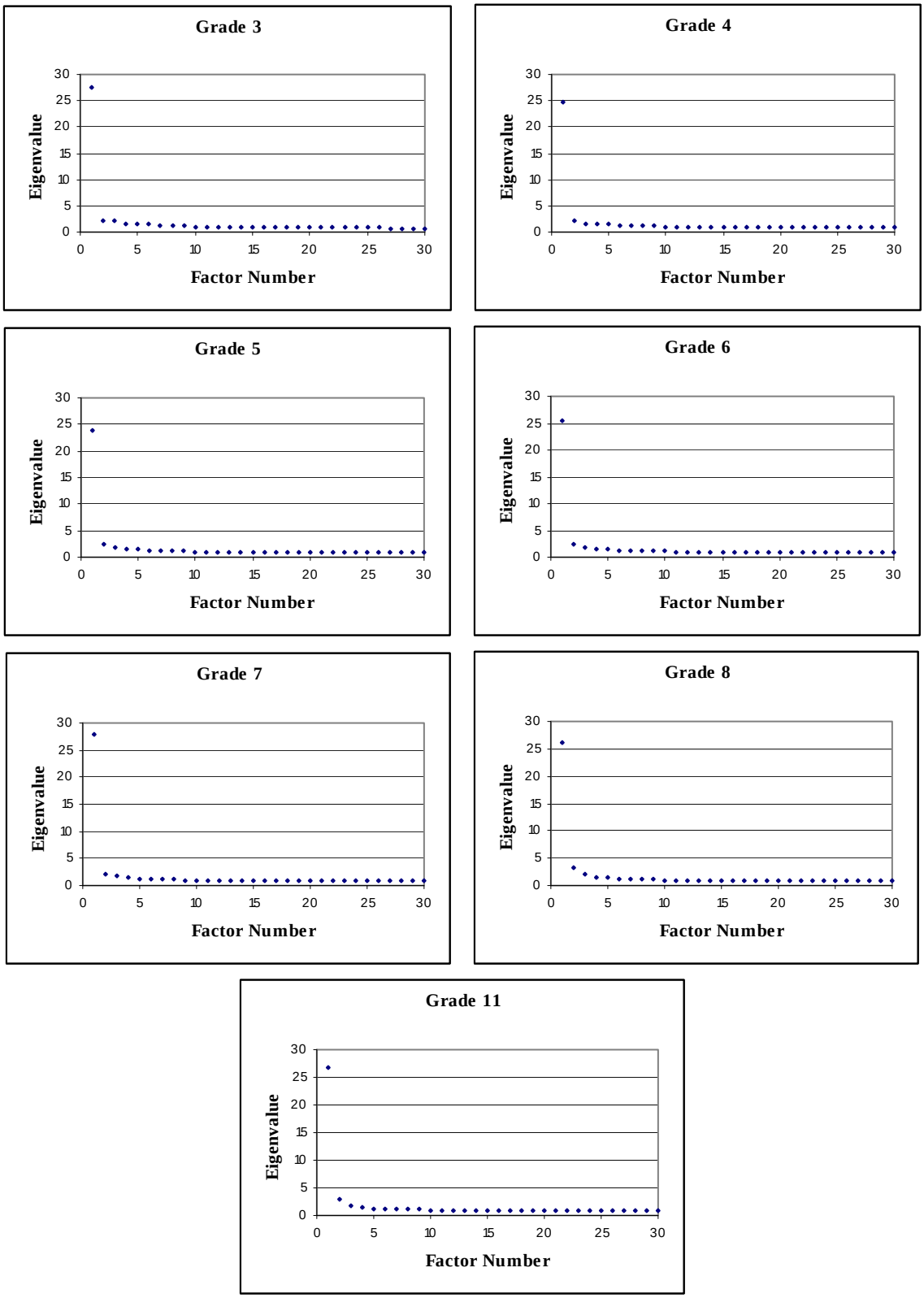
Data file 07/30/08 public schools only.

Figure 8.1: Reading Assessment PCA Scree Plot



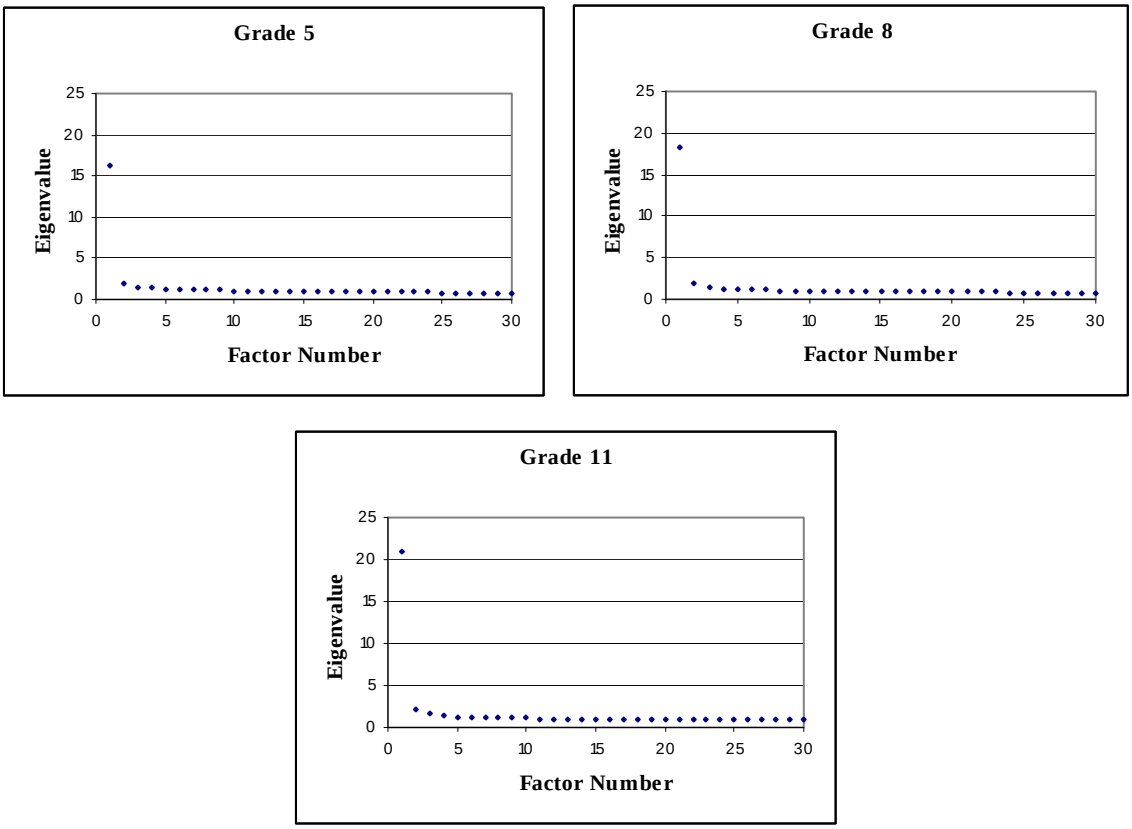
Data file 07/30/08 public schools only.

Figure 8.2: Mathematics Assessment PCA Scree Plot



Data file 07/30/08 public schools only.

Figure 8.3: Science Assessment PCA Scree Plot



Data file 07/30/08 public schools only.

CHAPTER 9: QUALITY CONTROL PROCEDURES

9.1 Quality Control for Item Development and Test Construction

The processes and guidelines used in test and item development described in Chapter 2, Item Development and Test Construction, were designed to ensure that the resulting items met specific criteria for content and format and serve as the primary means of controlling the quality of items. Three additional quality control measures for items and test forms are described below.

Pearson Item Review

Before Pearson submits any item to a review committee or the SD DOE, a lead assessment specialist within the content area (outside of the project) reviews each item for content accuracy and adherence to item format guidelines. If any inconsistencies or errors are found, the reviewer works with the assessment specialist assigned to the project to refine the items. This ensures that mistakes are identified and corrected before items are reviewed in committee.

Final Form Review

After the test forms have been approved by the SD DOE, Pearson conducts a final form review of test forms, pull lists, and test maps (scoring keys) to identify any remaining errors that would render an item invalid. Senior-level assessment specialists (outside of the project) review each test form for errors in content, alignment, and form building. If errors are found, the reviewer notifies the appropriate content lead on the project for resolution.

Publishing Review

Pearson's Publishing Quality Assurance group is responsible for performing final editorial checks of test materials prior to their release for printing. These checks include performing a cold read, taking the test, and cross-checking interrelated documents (i.e., test booklets, answer documents, and *Directions for Administering*) for accuracy, consistency, and functionality. Once all issues have been addressed, the Quality Assurance group provides final authorization to print test materials.

9.2 Quality Control for the Receipt of Non-Scannable Documents

Pearson's Scoring Center has established a precise check-in procedure for the receipt of non-scannable documents, which include test materials other than answer documents, such as used and unused student test booklets (except grade 3 booklets, which are scannable), *Directions for Administering*, and Mathematics reference sheets and rulers. At the time of receipt, the number of non-scannable documents received from schools is compared to the expected receipts based on the materials delivered to the schools. The materials that can be reused are placed back into stock. The materials that cannot be reused are destroyed by shredding. Prior to shredding any non-scannable materials, a manual check of all materials is performed to ensure that no scannable documents are destroyed.

9.3 Quality Control for Data Preparation

The Scoring Center uses a check-in procedure for answer documents similar to the one used for non-scannable documents described above. At the time of receipt, the number of answer

documents received from schools is compared to the expected receipts as indicated by the enrollment data for each school to determine that all shipments have been received. In addition, answer documents are checked to determine that no unusual circumstances exist (i.e., ink marks, water damage, paper tears, etc.) and that the documents have been correctly assembled. This is the *receiving edit*. If all specified conditions have been met, the order will be entered into a document-tracking computer program, the SCORFLOW Management System, to show that it has arrived and met all conditions necessary for processing. If an order has not met conditions for processing, it will be placed on *alert status*. Most alerts are capable of being resolved by Pearson's Alert Hotline, which works directly with the concerned school or district. Certain designated alert conditions will be reported to Pearson's Program Management team for resolution. As soon as an order's alert has been resolved, the order will be cleared for processing and placed in the SCORFLOW Management System. Alerts not resolved after three attempts by Pearson will result in scores for the concerned school/district being invalidated.

9.4 Quality Control in Document Tracking

Boxes containing all scannable documents are labeled with unique tracking bar codes to identify and track each box throughout the scanning process. The labels are then logged on scan logs, which list the ID number of each box scanned, the count of scanned documents in each, the name of the scan program used, the name of the scanner operator, and the date and time of scanning. The scan boxes are stored in sequential order on storage shelves in a secure warehouse.

In addition to the ID number for each box, the scanners are programmed to assign a unique number to each student answer document within a box, resulting in a "scan/batch" number, which can be used to verify the processing of all batches and to find a particular student record for editing and verification if needed. The number is printed on the edge of each scanned sheet as it runs through the scanner.

A complete listing of all scan boxes will be generated periodically and used by the Program Manager to ensure that all material received was scanned. These same lists will be used by the data processing staff for file verification during the editing and scoring process.

9.5 Quality Control in Editing and Data Input

All computer software used in editing, scoring, and merging is double-checked through a process called parallel processing, which compares the output from two different programs created by two different people. If the results of both programs are the same, the data are expected to be correct. Care is taken to derive the programs independently without sharing misassumptions or formulas, which may be in error. When the data from the two different programs are not identical, then both programmers examine their programs, the input, and the output to establish the source of the difference. Once the error is detected, the programs are run in parallel again and the output is checked one more time to verify that other errors were not introduced into the process.

Extensive editing steps are used to check all questionable scanned records for multiple marks, poor erasures, and incomplete data. Multiple marks are recorded in scanning as an asterisk (*) in the data file, which can then be flagged for further editing either in data correction or during the editing phase of the data file. Poor erasures can be detected through the scanner application,

which records the total number of erasures (discriminations made by the scanner during the resolution of each bubble). This data may be analyzed further and the actual answer documents can be pulled for further verification. Generally, these types of analyses can be done to detect inconsistencies.

The scanner programs are designed to detect incorrect test forms, document pages out of order, incomplete data fields (if desired), foreign materials, excessive light marks that may indicate miscalibration, and other important errors that may affect the validity and integrity of the scanning. Operators are trained to respond to the error messages and to alert a manager if necessary.

Incomplete data are detected through the use of data rules. The data rules determine if the test is complete or, if incomplete, whether the student record should be invalidated. Additionally, the data rules check for valid ranges, generally in demographic fields such as birth date, student ID, and grade. Data rules are used to establish the merging procedure for the demographic data from the pre-ID file submitted by the district versus the demographic data that may be scanned from the student document. If demographic data exists from both sources, the data rules establish which data to use.

Routines run in the editing phase also check for completeness of data files to prevent any data from being dropped, erased, or not included. Although scan logs and editing logs should prevent an undetected loss of data records, a complete listing of all student counts by school and LEA is done and checked against the counts from the school Scoring Services Identification (SSID) sheet submitted by the school and also against the original enrollment file (to check for reasonableness).

In order to safeguard against the mishandling of data files during the editing phase, all original scan files are archived and the editing staff work only with a copy of the file. Only the lead programming staff has access to the original data files. This step also ensures that in the event of an editing error, original data can be recovered rapidly. At every step, hard copies of the problem records (called “error listings”) are generated and maintained and only after correction are the changes merged back into the work files.

Editing logs are maintained in which detailed information about each editing step is recorded. Results of the duplicate records check, the pre-code merge, the error listings, and error resolutions are entered in these logs. These editing logs track the movement of the scan files in the editing system. Complete and accurate edit checks on questionable answer documents are performed after scanning.

9.6 Quality Control in Scoring and Reporting

Scoring

To verify the accuracy of item scoring, a mock set of test materials with known data is created and scored for each form of the test. This mock set of materials is known as a test deck. The scoring output from the test deck is checked using an *item analysis report* that records both the frequency of the scores by item distractor and the point-biserial correlations. These two statistics

are analyzed to verify that the proper scoring keys are being used and that the scoring programs are properly assigned the correct response for each key. Results of these analyses are provided to Pearson for a second analysis and for retaining as documentation of the process.

During the actual scoring process, the scoring keys are rechecked by running the item analysis report again on the first early sets of data for each form (approximately the first 500–1,000 records per form).

The *scoring output file* is formatted in such a way as to retain all information about the student responses, not just whether the student responded correctly or incorrectly, providing data with which additional analyses and checks can be done. The item information includes a code if the item is correct (the key retains the distractor information) and the actual distractor value for the student response if incorrect.

Reporting

When the system has determined that a school's anticipated documents have been edited and scanned, data for that order are run through the scoring program and the appropriate reports are produced. The printouts of a sufficient number of reports are examined to verify the validity and completeness of the data. It is important to note that table data, scoring algorithms, and reporting programs have already been verified as accurate in the quality control software, data have been verified as correct through the post-edit program, and tracking enrollment counts have been updated to reflect actual student counts. Printed reports for each school are grouped together for assembly and packaging.

As each school's reports are assembled in pre-mailing, shipments are inspected to verify that they include all reports ordered. All possible problem situations have specific procedures to be followed for resolution. In addition to these control procedures in the pre-packaging phase, there are complete monitoring controls of the packaging process.

GLOSSARY OF TERMS

Achievement Levels (see Performance Levels)

Accommodations–special testing conditions and methods allowed for certain students, primarily those with disabilities or with limited English proficiency

Adequate Yearly Progress (AYP)–set of accountability measures for states, districts, and schools contained in the *No Child Left Behind Act of 2001* (NCLB*)* covering student achievement, based on each state's academic content and student achievement standards and statewide assessments

Alignment Study–analysis of the degree to which test items address the indicators and standards contained in the State Content Standards

Balance of Representation–the alignment criterion for Balance of Representation is that items should be distributed among all of the standards at least to some degree.

Bias–advantage or disadvantage conferred upon groups of students because of certain personal characteristics (such as gender, race, ethnicity, religion, socioeconomic status, disability, or geographic region), unrelated to mastery of the content

Blueprints–recommendations and guidelines for the construction of the *Dakota STEP*, the development of items that specify the indicators and standards to be covered, and the number of items per standard or indicator for each. The number of items per indicator or standard must be sufficient to provide psychometrically reliable and valid measures of student achievement of them.

Calibration–the process of estimating the difficulty of each item on a test and the achievement levels of students who took the test. For the *Dakota STEP*, calibration is accomplished using the Joint Maximum Likelihood (JML) estimation method.

Categorical Concurrence–the alignment criterion for Categorical Concurrence is that at least six items measuring content from each standard are present.

Classical Test Theory (CTT)–a set of theories and methods based on the idea that an observed test score is the result of the test-taker's true score and a quantifiable degree of error. Although the *Dakota STEP* is scored using Item Response Theory (IRT), CTT statistics are nevertheless calculated as a means of quality control.

Content Area (Content Domain)–the information or skills contained in an area of study. The content areas (or subject areas) assessed on the *Dakota STEP* are Reading, Mathematics, and Science.

Criterion-Referenced Test (CRT)–an assessment in which an individual's performance is compared to a specific learning objective or achievement standard and not to the performance of other students. Criterion-referenced tests assess how well students perform on specific goals or standards rather than how their performance compares to a norm group of students.

Cut Score (Cut point)–a score that marks the threshold between Achievement Levels on the *Dakota STEP* (Below Basic, Basic, Proficient, Advanced). Cut Scores are established by the Department of Education, based on recommendations of a standard-setting committee.

Depth of Knowledge–the alignment criterion for Depth of Knowledge is that at least 50% of the items corresponding to a standard have to be at or above the cognitive complexity of the standard.

Differential Item Functioning (DIF)–used to identify gender and ethnic bias. Items with DIF exhibit differences in scores between groups that cannot be explained by differences between those groups in overall achievement. DIF is investigated using the Mantel-Haenszel method.

Domain (see Content Area)

Equating (see Linking)

Field Testing–the process of testing newly developed items before they are used as operational items to assess student performance

Goal–a level of content specification used in South Dakota's State Content Standards, more specific than Domains but more general than Indicators

Indicator–a level of content specification used in South Dakota's State Content Standards, more specific than Goals but more general than Standards

Individual Education Plan (IEP)–describes special education services provided. IEP also specifies the testing accommodations a student needs for classroom instruction and assessments.

Item–any test question or task for which a score point is awarded; a stem or stimuli and responses for which a score or set of scores is to be recorded

Item Response Theory (IRT)–a set of psychometric theories and methods based on the idea that a person's true level of ability can be estimated by examining the characteristics of items on a test and the test-takers' responses to them. IRT is widely used in large-scale testing because of the flexibility it provides in test construction and scoring.

Joint Maximum Likelihood (JML) Estimation–the method of item calibration used for the *Dakota STEP*, whereby the difficulty of each item and the achievement of each student are estimated simultaneously

Linking (Equating)–the process of placing student test scores on a scale that allows comparisons of tests from one year to scores from another year in which different test questions were used. For the *Dakota STEP*, the Stanford 10 items used as CRT items across test years are used to link test scores.

Mean *p*-value–the average of the *p*-values for all items in a cluster, subtest, or domain total

Mean Scale Score–the average of Scale Scores earned by a group taking a given subtest or domain total

NCLB–the *No Child Left Behind Act of 2001*

***No Child Left Behind Act of 2001* (NCLB)**–the 2001 reauthorization of the *Elementary and Secondary Education Act* of 1965 (ESEA). NCLB requires each state to adopt challenging academic content and achievement standards for all public school students and to implement a set of high-quality assessments to measure *Adequate Yearly Progress* toward meeting these standards.

Operational Items–items that count toward a student's score; "live" items (compare to field test items)

***p*-value**–the percentage of a group of students answering a test question correctly

Performance Levels (Achievement Levels, Proficiency Levels)–levels of achievement on the *Dakota STEP* that correspond to the four levels of achievement described in general terms in the State Content Standards: Below Basic, Basic, Proficient, and Advanced. Performance levels are bound by Cut Scores on the *Dakota STEP* score scale.

Point-biserial Correlation–measures the correlation between students' success on an item and their success on the test as a whole

Proficiency Levels (see Performance Levels)

Range of Knowledge–the alignment criterion for Range of Knowledge is that 50% of the standards have to have at least one related assessment item.

Rasch Model–a one-parameter IRT model used in calibrating and linking *Dakota STEP* CRT scores

Raw Score–the sum of the items each student answers correctly

Reliability–desired characteristic of a test; achieved when measurement error is minimized

Reporting Categories–the levels of content specification from the State Content Standards–content *domain* (subject), *goal*, *indicator*, and *standard*–for which test results can be reported

Scale Score (SS)–a standard score derived from the Raw Score through a process of calibrating, linking (equating), and scaling. Scale scores allow comparisons of achievement across different versions (years) of a test. *Dakota STEP* scores are placed on the Stanford 10 score scale.

Scaling–the process of placing scores resulting from calibrating and linking on the scale used for reporting scores, the Stanford 10 scale

Scoring Service Identification Sheet (SSID or Header Sheet)–completed by test coordinators and/or teachers/examiners after testing is complete. SSIDs must accompany answer documents that are sent to the Pearson Scoring Center. The placement of this form determines whether score reports are generated by classroom or by grade within each school.

Section 504–special classification of students as defined in Section 504 of the Rehabilitation Act of 1973. Testing accommodations are permitted for students who meet the Section 504 criteria.

Standard–a statement of what students should know and be able to do. A standard is more specific than a Goal or Indicator.

Standard Deviation–a measure of the variability in a group of scores

Standard Error of Measurement (SEM)–the standard deviation of the error score contained in the Raw Score, representing the uncertainty related to it

TCH–*Test Coordinator's Handbook*

Validity–desired characteristic of a test; achieved when the test actually measures what it is intended to measure

REFERENCES

Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores in Mathematics.* CSE Technical Report 428. http://www.cse.ucla.edu/CRESST/pages/reports.htm.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th edition). Upper Saddle River, NJ: Prentice Hall.

Buros Center for Testing (BIACO). *Evaluating the Alignment of the Full 2008 Dakota STEP Reading Assessment Forms with South Dakota's Academic Content and Achievement Standards at Grades 3–8 and High School.* Report prepared for the South Dakota Department of Education, August 17, 2008.

Buros Center for Testing (BIACO). *Evaluating the Alignment of the Full 2008 Dakota STEP Mathematics Assessment Forms with South Dakota's Academic Content and Achievement Standards at Grades 3–8 and High School.* Report prepared for the South Dakota Department of Education, August 13, 2008.

Buros Center for Testing (BIACO). *Evaluating the Alignment of the Full 2008 Dakota STEP Science Assessment with South Dakota's Academic Content and Achievement Standards at Grades 5, 8 and 11.* Report prepared for the South Dakota Department of Education, November 8, 2006.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* NY: Pearson Brace Jovanovich College Publishers.

Divgi, D. R. (1980). *Dimensionality of Binary Items: Use of a Mixed Model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publication, Inc.

Pearson (2002). *Stanford Achievement Test Series tenth edition spring techincal data report.*

Kolen, M. J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd edition).* Springer.

Linacre, J.M. (2003). *User's Guide and Program Manual: WINSTEPS, MINISTEP, Rasch-Model Computer Programs*. Chicago: Mesa Press.

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. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

MacCann, R. G. & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research, & Evaluation, 9(5)*, 1–12.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning: Theory and Practice*. (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.