Technical Manual for Minnesota Title I and Title III Assessments

For the Academic Year 2014–2015

September 2015

Minnesota Department of Education

Prepared by Pearson

Table of Contents

Index of Tables

Table of Figures

Purpose

This technical manual provides information about the development and measurement characteristics of the Minnesota Assessment System. It is organized into two parts: (1) chapters providing general information about the measurement process and (2) yearly appendices providing the specific data for a given year. The chapters outline general information about the construction of the Minnesota assessments, statistical analysis of the results, and the meaning of scores on these tests. The appendices, organized as Yearbooks, provide detailed statistics on the various assessments for a given academic year.

Improved student learning is a primary goal of any educational assessment program. This manual can help educators use test results to inform and improve instruction, thereby enhancing student learning. In addition, this manual can serve as a resource for educators in explaining assessment information to students, parents, teachers, school boards, and the general public.

A teacher constructing a test meant to provide immediate feedback on classroom instruction desires the most accurate assessment possible but typically does not need to identify the technical measurement properties of the test before or after administering it. However, a large-scale standardized assessment does require evidence to support the meaningfulness of the inferences made from the scores (validity) and the consistency with which the scores are derived (reliability, equating accuracy, and freedom from processing errors). That evidence is reported in this manual.

This manual does not include all the information available regarding the assessment program in Minnesota. Additional information can be found on the Minnesota Department of Education (MDE) website at http://education.state.mn.us/MDE/index.html. Questions may also be directed to the Division of Statewide Testing at MDE by email: mde.testing@state.mn.us.

MDE is committed to following generally accepted professional standards when creating, administering, scoring, and reporting test scores. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) is one source of professional standards. As evidence of our dedication to responsible and fair testing practices, an annotated table of contents linking the sections of this manual to the *Standards* is provided immediately after the glossary.

Chapter 1: Background

With the enactment of the No Child Left Behind Act (NCLB) in 2002, Minnesota accountability and statewide assessment requirements were dramatically increased. Under NCLB Title I, the State must develop academic content standards in the core academic areas, measure those standards and define student proficiency levels—minimum scores that students must obtain on a state assessment in order to be considered academically proficient—in the core subjects. According to NCLB, by 2005–2006, all students must take annual reading and mathematics tests in grades 3–8 and once during high school. By 2007–2008, students must be tested in science at least once in each of the following grade spans: grades 3–5, 6–9, and 10–12. The overall goal of NCLB is to have all students proficient in reading and mathematics by 2014. Title I accountability assessments (reading and mathematics) include a state's responsibility to establish Annual Measurable Objectives (AMO) for schools to determine Adequate Yearly Progress (AYP).

Under NCLB Title III, the State must develop and assess English Language Proficiency (ELP) standards for all students identified as English Learners (EL). Title III accountability assessments include a state's responsibility to establish Annual Measurable Achievement Objectives (AMAO) for EL students. This requirement establishes additional tests for EL students.

Minnesota Assessment System History

Prior to NCLB, Minnesota had already developed an accountability system. The standards movement began in Minnesota in the late 1980s and evolved into a comprehensive assessment system with the development of test specifications and formal content standards during the 1990s. State and federal legislation has guided this process.

A Brief History of the Program

1995

The Minnesota legislature enacted into law a commitment to "establishing a rigorous, results-oriented graduation rule for Minnesota's public school students [. . .] starting with students beginning ninth grade in the 1996–1997 school year" (Minn. Stat. §120B.30.7c). The Minnesota Department of Education (MDE) developed a set of test specifications to measure the minimum skills needed in order to be successful in the workforce. This was the basis for the Minnesota Basic Skills Test (BST), the first statewide diploma test. To establish higher academic standards, teachers, parents, and community members from across Minnesota collaborated to develop the Profile of Learning, Minnesota's first version of academic standards, as well as classroom-based performance assessments to measure these standards. Minnesota developed its assessment program to evaluate student progress toward achieving academic excellence, as measured by the BST and performance assessments of the Profile of Learning.

1997

The Minnesota legislature mandated a system of statewide testing and accountability for students enrolled in grades 3, 5, and 7 (Minn. Stat. §120B.30). This legislation required all Minnesota students in those grades to be tested annually using a single statewide test by grade and subject for the purpose of system accountability.

1998

MDE developed the Minnesota Comprehensive Assessments (MCAs) to fulfill the mandates of the statewide testing statute enacted in 1997. The statewide testing law also required that high school students be tested on selected standards within the required learning areas beginning in the 1999–2000 school year (see Minnesota Statute 120B.30; https://www.revisor.mn.gov/statutes/?id=120b.30). Special education students were required to participate in testing according to the recommendations of their Individualized Education Program (IEP) or 504 plan. English Learners who were in the United States for less than three years were exempted from the BST.

Since 1998, all Minnesota grades 3 and 5 students have been tested annually using a single statewide test for the purpose of statewide system accountability.

2001

The Division of Special Education Policy developed Alternate Assessments (AA)—checklists for mathematics, reading, writing, and functional skills—to be used in place of the MCA or BST for students whose Individualized Education Program (IEP) and 504 plan teams determined it was appropriate.

2004

Grade 10 students were administered the MCA Reading, and grade 11 students were tested with the MCA Mathematics. This year also marked the first operational administration of the MCA Reading and Mathematics to grade 7 students.

2006

In 2005–2006, in response to NCLB legislation, the Minnesota Assessment System was expanded. Students in grades 3–8, 10, and 11 were assessed with the first Minnesota Comprehensive Assessments-Series II (MCA-II) in mathematics and reading. Information from these tests was used to determine proficiency levels in each school and district for the purpose of determining Adequate Yearly Progress (AYP) and to evaluate student, school, and district success in Minnesota's standards-based education system for NCLB. This assessment system would be expanded in future years to meet further requirements under NCLB.

2007

The Minnesota legislature provided for the Graduation-Required Assessment for Diploma (GRAD) as the retest option for high school students to fulfill their graduation exam requirement. The GRAD measured the writing, reading, and mathematics proficiency of high school students. The Mathematics Test for English Language Learners (MTELL) was first introduced as an alternate assessment for those students learning English. Also in this year, students with the most significant cognitive disabilities participated in the Minnesota Test of Academic Skills (MTAS) for the first time.

2008

Students in grades 5, 8, and high school took the Science MCA-II using an interactive computer assessment. In those same grades, students with the most significant cognitive disabilities participated in the Science MTAS for the first time. The grade 10 Reading MCA-II included the initial operational administration of the embedded Reading GRAD. Reading and Mathematics MTAS were lengthened and scoring procedures clarified.

2009

The grade 11 Mathematics MCA-II included the initial operational administration of the embedded Mathematics GRAD. The Minnesota legislature provided an alternate pathway for meeting the GRAD requirement in mathematics: after making three unsuccessful attempts at the Mathematics GRAD, followed by remediation, a student would be considered to have met the GRAD requirement.

2010

Items for construction of the Minnesota Comprehensive Assessments-Modified assessments in mathematics and reading were field-tested. Technology-enhanced items for the Mathematics MCA-III were field-tested. A study was conducted to link scores on the Reading MCA-II and GRAD to the Lexile® scale in order to permit inferences about Lexile® reading scores based on scores from Minnesota reading assessments. This year saw the final administration of the Mathematics Test for English Language Learners (MTELL).

2011

This year saw the first operational administrations of Mathematics MCA-III as well as the Minnesota Comprehensive Assessments-Modified assessments in mathematics and reading. Districts chose to administer the Mathematics MCA-III either on computer or on paper. The computer version included technology-enhanced items. Grades 5–8 of Mathematics MCA-Modified were computer delivered. Mathematics MCA-III, grades 5–8 of Mathematics MCA-Modified, and grades 3–8 of Mathematics MTAS assessed the *2007 Minnesota K–12 Academic Standards in Mathematics.*

2012

For districts opting to participate online, the Mathematics MCA-III assessments in grades 3–8 were administered as a multi-opportunity adaptive test, offering students up to three testing opportunities, with the highest score used for score reporting and accountability purposes. In addition, this year saw the first operational administration of the Science MCA-III assessments in grades 5, 8, and high school, which continued to be administered online and assessed the 2*008 Minnesota K–12 Academic Standards in Science.* A new Title III assessment was introduced in 2012, Assessing Comprehension and Communication in English State-to-State for English Language Learners (the ACCESS for ELLs). Administration of the Test of Emerging Academic English (TEAE) and the Minnesota Student Oral Language Observation Matrix (MN SOLOM) was discontinued. ACCESS for ELLs is an English language proficiency (ELP) assessment given to students identified as English Learners in grades K–12. It is administered annually in states that are members of the World-Class Instructional Design and Assessment (WIDA) consortium. Test development for ACCESS for ELLs is performed by the Center for Applied Linguistics (CAL), and MetriTech, Inc. manages the printing, scoring, reporting, and distribution of all ACCESS test materials.

2013

This year saw the first operational administration of the Reading MCA-III, MCA-Modified, and MTAS aligned to *2010 Minnesota K–12 Academic Standards in English Language Arts.* Districts chose to administer the Reading MCA-III either on computer or on paper. The computer version included technology-enhanced items, whereas the paper version included only multiple choice items. A study was done to link MCA-III Reading scores to the Lexile® scale in order to permit inferences about Lexile® reading scores based on scores from Minnesota reading assessments. Grades 5–8 and 10 Reading MCA-Modified assessments were delivered on computer. This year also saw the first operational administration of the Optional Local Purpose Assessment (OLPA) for mathematics administered as a multi-opportunity adaptive test, offering students up to two testing opportunities. The administration of the Mathematics MCA-III was changed in spring 2013 to be a single-opportunity test.

2014

This year saw the first operational administration of the grade 11 Mathematics MCA-III, MCA-Modified, and MTAS aligned to *2007 Minnesota K–12 Academic Standards in Mathematics.* Districts chose to administer the grade 11 Mathematics MCA-III either on computer or on paper. The computer version included technology-enhanced items. This year also marked the last operational administration of the Mathematics and Reading MCA-Modified.

2015

The Mathematics, Reading, and Science MCA-III were administered online only (with the exception of accommodated paper forms for special need students). Census administrations of career and college readiness exams in grades 8 and 10, Explore and Plan, took place in fall 2014. The college entrance exam, ACT Plus Writing, was administered to all grade 11 students in spring 2015. A college placement diagnostic exam, Compass, was given to some students after grade 10 Plan and prior to grade 11 ACT Plus Writing. Students who participated had been determined to be not yet academically ready for career and college based on their performance on the grades 8 and 10 assessments. This was the last academic year in which the GRAD retests were still available as an option to meet graduation assessment requirements for students who first enrolled in grade 8 through 2010-2011. The first administration of the Reading OLPA as a single-opportunity fixed form online test took place. Reading MCA-III was being developed as a computer adaptive assessment, which will be first administered in spring 2016.

The timeline in Table 1.1 highlights the years in which landmark administrations of the various Minnesota assessments have occurred.

Table 1.1. Minnesota Assessment System Chronology

Date	Event
1995–96	• First administration of Minnesota Basic Skills Test (BST) Mathematics and Reading in grade 8 • First administration of Minnesota BST Written Composition in grade 10
1997–98	• First administration of Minnesota Comprehensive Assessments (MCAs) at grades 3 and 5
1998–99	• Development of High School Test Specifications for MCAs in grades 10–11 • Field test of Test of Emerging Academic English (TEAE)
2000–01	• First administration MCA/BST Written Composition • Field test of Reading MCA in grade 10 and Mathematics MCA in grade 11
2001–02	• Second field test of Reading MCA in grade 10 and Mathematics MCA in grade 11
2002–03	• First administration of Reading MCA in grade 10 and Mathematics MCA in grade 11 • Field test of grade 7 Reading and Mathematics MCA • Revision of grade 11 Mathematics Test Specifications
2003–04	• First field test of Reading and Mathematics MCA in grades 4, 6, and 8 • First operational administration (reported) of MCA Mathematics and Reading in grade 7, Reading in grade 10, and Mathematics in grade 11
2004–05	• Second field test of MCA Reading and Mathematics in grades 4, 6, and 8
2005–06	• First operational administration of Mathematics and Reading MCA-II in grades 3–8, 10, and 11
2006–07	• First administration of Written Composition Graduation-Required Assessments for Diploma (GRAD) test in grade 9 • Last year of BST Written Composition in grade 10 as a census test • Field test of Mathematics Test for English Language Learners (MTELL) and Minnesota Test of Academic Skills (MTAS) • First operational administration of Mathematics and Reading MTAS • First operational administration of MTELL
2007–08	• Field test of MTAS • First administration of Science MCA-II in grades 5, 8, and high school • First administration of Reading GRAD • First operational administration of Science MTAS
2008–09	• First operational administration of Mathematics GRAD
2009–10	• Field test of technology enhanced Mathematics MCA-III items • Field test of Mathematics and Reading Minnesota Comprehensive Assessments-Modified • Lexile® linking study
2010–11	• First operational administration of Mathematics MCA-III in grades 3–8 • Districts given choice of computer or paper delivery of Mathematics MCA-III • First operational administration of Mathematics and Reading MCA-Modified
2011–12	• First operational administration of Science MCA-III • First year of Mathematics MCA-III online assessments being delivered as a multi-opportunity computer adaptive assessment • First operational administration of ACCESS for ELLs as a Title III assessment

Table 1.1. Minnesota Assessment System Chronology (continued)

Date	Event
2012–13	• First operational administration of Reading MCA-III, MCA-Modified and MTAS aligned to 2010 Minnesota K–12 English Language Arts Standards • Districts given choice of computer or paper delivery of Reading MCA-III • Lexile® linking study for Reading MCA-III • First operational administration of the Optional Local Purpose Assessment (OLPA) for Mathematics being delivered as a multi-opportunity computer adaptive assessment • Online Mathematics MCA-III reverts to being a single-opportunity assessment • First operational administration of Alternate ACCESS for ELLs as Title III assessment • Census administration of Reading GRAD in grade 10 discontinued • Last year of Written Composition GRAD in grade 9 as a census test
2013–14	• First operational administration of grade 11 Mathematics MCA-III, MCA-Modified, and MTAS aligned to 2007 Minnesota K–12 Mathematics Standards • Districts given choice of computer or paper delivery of grade 11 Mathematics MCA-III • Final operational administration of Mathematics and Reading MCA-Modified • Census administration of Mathematics GRAD in grade 11 was discontinued
2014–15	• Census administrations of Explore, Plan and ACT Plus Writing • Final administrations of Mathematics, Reading and Written Composition GRAD retests • First operational administration of Reading OLPA as a single-opportunity, fixed-form online test • First year of developing Reading MCA-III as a computerized adaptive assessment

Organizations and Groups Involved

A number of groups and organizations are involved with the Minnesota assessment program. Each of the major contributors listed below serves a specific role, and their collaborative efforts contribute significantly to the program's success. One testing vendor constructs and administers all tests, while other vendors provide other independent services.

Assessment Advisory Committee

As mandated by Minnesota Statutes section 120B.365, the Assessment Advisory Committee must review all statewide assessments. View full text of Minnesota Statutes section 120B on the Office of the Revisor's website (https://www.revisor.mn.gov/statutes/?id=120B.365). As the statute states, "The committee must submit its recommendations to the commissioner and to the committees of the legislature having jurisdiction over kindergarten through grade 12 education policy and budget issues. The commissioner must consider the committee's recommendations before finalizing a statewide assessment."

> Subdivision 1. Establishment. An Assessment Advisory Committee of up to 11 members selected by the commissioner is established. The commissioner must select members as follows:
> (1) two superintendents;
> (2) two teachers;
> (3) two higher education faculty; and
> (4) up to five members of the public, consisting of parents and members of the business community.
>
> Subdivision. 2. Expiration. Notwithstanding section 15.059, subdivision 5, the committee expired on June 30, 2014.
> (Minn. Stat. §120B.365)

Table 1.2. Assessment Advisory Committee

Name	Position	Organization
Barb Ziemke	Parent	PACER Center, MN PTI
Barbara Hunter	Teacher	St. Paul Public Schools
Amy Jones	Teacher	Minneapolis Public Schools
Paul Carney	Higher Education	Fergus Falls Community College
Sandra G. Johnson	Higher Education	St. Cloud State University
Mo Amundson	Public	Governor's Workforce Development Council
Christopher Moore	Public	Minneapolis Schools

Human Resources Research Organization (HumRRO)

HumRRO is a separate vendor working with MDE to complete quality assurance checks associated with elements of the Minnesota Assessment System and accountability program. In collaboration with MDE and Minnesota's testing contractor, HumRRO conducts quality checks during calibration, equating, and scoring of Minnesota's Title I assessments, including MCA and MTAS. HumRRO has also conducted alignment studies to evaluate the congruence between the items on Minnesota assessments and the skills specified in the Minnesota Academic Standards.

Local Assessment and Accountability Advisory Committee

The Local Assessment and Accountability Advisory Committee (LAAAC) advises MDE on assessment, accountability, and technical issues.

Table 1.3. Local Assessment and Accountability Advisory Committee

Name	Position	Organization
Sherri Dahl	District Assessment Coordinator, Title I	Red Lake Schools
John Lindner	Assessment Specialist	St. Paul Public Schools
Johnna Rohmer-Hirt	District Research, Evaluation, and Testing Achievement Analyst	Anoka-Hennepin Public Schools
Justin Treptow	Assistant Principal	Minnesota Virtual Academy High School
Scott Fitzsimonds	Director of Technology, Teaching, and Learning	Watertown-Mayer High School

Minnesota Department of Education

MDE’s Division of Statewide Testing has the responsibility of carrying out the requirements in the Minnesota statute and rule for statewide assessments and graduation standards testing. The division oversees the planning, scheduling, and implementation of all major assessment activities and supervises the agency’s contracts with Pearson. In addition, the MDE Statewide Testing staff, in collaboration with an outside vendor, conducts quality control activities for every aspect of the development and administration of the assessment program. The Statewide Testing staff, in conjunction with MDE’s Compliance and Assistance Division, is also active in monitoring the security provisions of the assessment program.

Minnesota Educators

Minnesota educators—including classroom teachers from K–12 and higher education, curriculum specialists, administrators, and members of the Best Practice Networks, who are working groups of expert teachers in specific content areas—play a vital role in all phases of the test-development process. Committees of Minnesota educators review the test specifications and provide advice on the model or structure for assessing each subject. They also work to ensure that test content and question types align closely with good classroom instruction.

Draft benchmarks were widely distributed for review by teachers, curriculum specialists, assessment specialists, and administrators. Committees of Minnesota educators assisted in developing drafts of measurement specifications that outlined the eligible test content and test item formats. MDE refined and clarified these draft benchmarks and specifications based on input from Minnesota educators. After the development of test items by professional item writers, committees of Minnesota educators review the items to judge appropriateness of content and difficulty and to eliminate potential bias. Items are revised based on input gathered from these committee meetings. After items are field-tested, Minnesota educator committees are convened to review each item and its associated data for appropriateness for inclusion in the item bank from which the test forms are built.

To date, more than 2,000 Minnesota educators have served on one or more of the educator committees involved in item development for Minnesota assessments. Sign up to participate by registering on the MDE website (http://education.state.mn.us/MDE/EdExc/Testing/RegAdvPanel/).

Minnesota's Testing Contractors

Pearson served as a testing contractor for MDE beginning in 1997 and as the primary contractor for all Minnesota assessments from 2005 through the close of the 2010–2011 test administration cycle.

After that, the American Institutes for Research (AIR) served as MDE's primary testing contractor through the close of the 2013–2014 test administration cycle. AIR worked with Data Recognition Corporation (DRC), a subcontractor primarily responsible for printing, distribution, and processing of testing materials, to manage all Title I assessments in Minnesota.

Beginning with the 2014–2015 test administration cycle, Pearson has become the primary contractor. Currently Pearson provides Minnesota's Standards-Based Accountability Assessments, as well as resources for District Assessment Coordinators, including assessment manuals, Test Monitor and Student Directions, training resources, and item samplers.

MDE's testing contractors are responsible for the development, distribution, and collection of all test materials as well as for maintaining security for tests. Contractors work with MDE to develop test items and forms, produce ancillary testing materials that include test administration manuals and interpretive guides, administer tests to students on paper and online, collect and analyze student responses, and report results to the field. Contractors are responsible for scoring all student test forms, including written composition exams that are human-scored, paper tests that utilize scannable answer documents, and online tests that employ both multiple-choice items and items that utilize machine-scored rubrics.

The testing contractor may also conduct standard setting activities, in collaboration with panels of Minnesota educators, to determine the translation of scores on Minnesota assessments into performance levels on the Minnesota Academic Standards. Previously, AIR conducted standard setting procedures for Science MCA-III and MTAS on June 25–29, 2012; Reading MCA-III, MCA-Modified, and MTAS on June 24–28, 2013; and Grade 11 Mathematics MCA-III, MCA-Modified and MTAS on June 18–19, 2014.

National Technical Advisory Committee

The National Technical Advisory Committee (TAC) serves as an advisory body to MDE. It provides recommendations on technical aspects of large-scale assessment, including item development, test construction, administration procedures, scoring and equating methodologies, and standard setting workshops. The National TAC also provides guidance on other technical matters, such as practices not already described in the *Standards for Educational and Psychological Testing,* and continues to provide advice and consultation on the implementation of new state assessments and meeting the federal requirements of NCLB.

Table 1.4. National Technical Advisory Committee

Name	Position	Organization
Dr. E. Roger Trent	Trent Consulting	Columbus, Ohio
Dr. Gregory J. Cizek	Professor of Educational Measurement and Evaluation, School of Education	University of North Carolina at Chapel Hill
Dr. Claudia Flowers	Associate Professor in Educational Research and Statistics	University of North Carolina at Charlotte
Dr. S. E. Phillips	S.E. Phillips, Consultant	Mesa, Arizona
Dr. Mark Reckase	Professor of Measurement and Quantitative Methods, College of Education	Michigan State University

State Assessments Technology Work Group

The State Assessments Technology Work Group (SATWG) ensures successful administration of computer-delivered assessments by developing a site readiness workbook, testing software releases, and providing feedback to the Minnesota Department of Education and to vendors during and after online test administrations.

Table 1.5. State Assessments Technology Work Group

Name	Position	Organization
Andrew Baldwin	Director of Technology	South Washington County Schools
Tina Clasen	District Technology Supervisor	Roseville Public Schools
Delonna Darsow	Director of Assessment & Data Analysis	Burnsville-Eagan-Savage Public Schools
Josh Glassing	System Support Specialist III	St. Paul Public Schools
Sue Heidt	Director of Technology	Monticello Public Schools
Kathy Lampi	Technology/Testing	Mounds View Public Schools
John Lindner	Research, Evaluation, and Assessment	St. Paul Public Schools
Darin Marcussen	District Technology Coordinator	North Branch Public Schools
Sharon Mateer	District Assessment Coordinator	Anoka-Hennepin Public Schools
Hai Nguyen	IT Services	Minneapolis Public Schools
Don Nielsen	IT Support—Online Assessment	Minneapolis Public Schools
Mary Roden	Coordinator of Assessment and Evaluation	Mounds View Public Schools
Jeanne Sorsen	Research, Evaluation, and Testing Technology Support Technician	Anoka-Hennepin Public Schools
Douglas Tomhave	IT Services	South St. Paul Public Schools
Jim Varian	Technology Director	Big Lake Schools Public Schools
Annette Zacharias	Technology Director	Willow River Public Schools

Minnesota Assessment System

The Minnesota Department of Education (MDE) provides general information about statewide assessments at http://education.state.mn.us/MDE/SchSup/TestAdmin/index.html. Minnesota's test vendor also maintains a website that provides information about Minnesota assessments. Material available on these websites includes such documentation as

1. testing schedules;
2. rubrics and descriptions of student performance at various levels of mathematics, reading, science, and writing proficiency;
3. test specifications and technical manuals; and
4. information for parents.

The No Child Left Behind Act (NCLB) and reauthorization of the Elementary and Secondary Education Act (ESEA) reshaped the Minnesota system of standards, assessments, and school accountability. Three classes of assessments have been developed to measure the educational progress of students: Title I assessments, Title III assessments, and Minnesota diploma assessments.

The Title I assessments are used to evaluate school and district success toward Adequate Yearly Progress (AYP) related to the Minnesota Academic Standards for mathematics, reading, and science. Additional alternate assessments exist for special populations of students, such as students with disabilities or English Learners (EL). All students in grades 3–8, 10, and 11 are required to take a Title I assessment according to their eligibility status. Minnesota's Title I assessments are listed in Table 1.6 and described in the paragraphs below.

Title I Assessments

Table 1.6. Title I Accountability Tests in 2014–15

Test	**Subject**	**Grades**
MCA-III	Mathematics	3–8, 11
	Reading	3–8, 10
	Science	5, 8, 9–12[1]
MTAS	Mathematics	3–8, 11
	Reading	3–8, 10
	Science	5, 8, 9–12

Mathematics

Minnesota Comprehensive Assessments-Series III

The Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) is an exam aligned with the *2007 Minnesota K–12 Academic Standards in Mathematics* that has been given in grades 3–8 since spring 2011 and in grade 11 since spring 2014. Students are asked to respond to questions involving mathematical problem-solving. They answer questions about concepts and skills in four different math content strands: (1) numbers and operations; (2) algebra; (3) geometry and measurement, and (4) data

[1] The high school Science MCA-III is given to students in the year they complete their instruction in life science.

analysis and probability. Originally, the Mathematics MCA-III could be administered in either online or paper modes, according to district choice. Currently, only online administration is available (except for accommodated paper forms). The online Mathematics MCA-III exams include technology-enhanced item types similar to those given on the Science MCA-III tests. These item types allow measurement of higher-level thinking and concepts. The 2011 online and paper administrations were fixed forms that included 50 scored items. Beginning in 2012, the online test was administered adaptively and included 42 scored items. The paper version included 50 scored multiple-choice and (in grades 5 and above) gridded-response items. A unique feature of the 2012 online Mathematics MCA-III administration was that students were permitted to take this computer adaptive assessment up to three times and use their highest score for accountability purposes. Since 2013, only a single testing opportunity has been allowed.

Minnesota Test of Academic Skills

The Mathematics Minnesota Test of Academic Skills (MTAS) is given in grades 3–8 and 11. Each test contains a set of nine performance tasks designed to measure mathematical problem-solving. The Mathematics MTAS has been aligned with the *Minnesota K-12 Academic Standards.* The math content strands are the same as those tested by the Mathematics MCA-III in grades 3–8 and 11 and mirror their pattern of emphasis but with a reduction in the depth or complexity of concepts measured. The performance tasks can be administered on different days according to the needs of the student.

Reading

Minnesota Comprehensive Assessments-Series III

The Reading MCA-III is an exam aligned with the *2010 Minnesota K–12 Academic Standards in English Language Arts* that has been given in grades 3–8 and 10 since spring 2013. Students are asked to read both literature and informational text. For literature, students use strategies to analyze, interpret, and evaluate fiction such as short stories, fables, poetry, and drama. For informational text, students use strategies to analyze, interpret, and evaluate nonfiction such as expository and persuasive text and literary nonfiction. Originally, the Reading MCA-III could be administered in either online or paper modes, according to district choice. Currently, only online administration is available (except for accommodated paper forms).The online Reading MCA-III exams include technology-enhanced item types similar to those given on the Mathematics MCA-III and Science MCA-III exams. These item types allow measurement of higher-level thinking and concepts. The total number of scored items for the 2014 paper and online administrations for Reading MCA-III were as follows: for grades 3–5, there were 48 items; for grades 6–8, there were 54 items; and for grade 10, there were 60 items. The online exam is divided into sections by passage and each passage's associated items; the paper exam is administered in four or five separate segments that may be given on different days.

Minnesota Test of Academic Skills

The Reading MTAS is given in grades 3–8 and 10. Each test contains a set of nine performance tasks designed to measure student understanding of text. The Reading MTAS has been aligned with the *Minnesota K-12 Academic Standards*. The reading content strands are the same as those tested by the Reading MCA-III and mirror their pattern of emphasis but with a reduction in the depth or complexity of concepts measured. The Reading MTAS passages feature simple sentence structure, repetition of words and ideas, and high frequency, decodable words. The passages may be read aloud to students, signed manually, represented tactilely, and/or accompanied by objects, symbols, and illustrations. The complexity of grade-level passages increases from grades 3–8 to high school by using grade- and age-appropriate vocabulary and subject matter, as well as increases in word count and length. The performance tasks can be administered on different days according to the needs of the student.

Science

Minnesota Comprehensive Assessments-Series III

The computer-delivered Science MCA-III assessments, which are administered in grades 5 and 8 and once in high school, are aligned with the *Minnesota K-12 Academic Standards.* The grade 5 assessment covers the content standards taught in grades 3, 4, and 5, and the grade 8 assessment covers the standards for grades 6, 7, and 8. Students in grades 9–12 are expected to take the high school MCA if, in the current academic year, they are enrolled in a life science or biology course and/or have received instruction on all strands and standards that fulfill the life science credit for graduation.

The MCA-III tests for grades 5, 8, and high school were initially administered operationally in spring 2012. The assessments were administered online in fixed forms. The assessments had 41, 51, and 68 operational items respectively in grades 5, 8, and high school. Item types included multiple-choice, constructed-response and figural-response items. Figural-response items allow students to respond by selecting one or more points on a graphic image or moving objects around within the image.

Minnesota revised its academic standards in science in 2009, and the new standards were implemented in May 2010. Most notably, the revised standards explicitly include engineering knowledge and skills so that they align with the emphasis on science, technology, engineering, and mathematics (STEM) necessary for success in the twenty-first century. In grades 5 and 8, students answer questions about concepts and skills in four different science strands:

1. Nature of Science and Engineering
2. Physical Science
3. Earth and Space Science
4. Life Science

In high school, students answer questions about concepts and skills in two different science strands:

1. Nature of Science and Engineering
2. Life Science

Minnesota Test of Academic Skills

The Science MTAS is given in grades 5, 8, and high school. Each test contains a set of nine performance tasks designed to measure student understanding of science concepts. The Science MTAS has been aligned with the Minnesota Academic Standards. The science content strands are the same as those tested by the Science MCA-III and mirror their pattern of emphasis but with a reduction in the depth or complexity of concepts measured. The performance tasks can be administered on different days according to the needs of the student.

Title III Assessments

These assessments are designed to help evaluate school and district success toward Annual Measurable Achievement Objectives (AMAOs) related to Title III of NCLB. They also serve as evidence of proficiency for district funding for EL programming by the state. All EL students are required to take the Title III assessments.

ACCESS for ELLs

The ACCESS for ELLs (Assessing Comprehension and Communication in English State-to-State for English Language Learners) is a test of English language proficiency in reading, writing, listening, and speaking administered to EL in grades K–12. The tests for all four skill domains are aligned to WIDA's English language development standards, which describe performance in five areas:

1. Communication for social and instructional purposes within the school setting
2. Communication of information, ideas, and concepts necessary for academic success in the content area of Language Arts
3. Communication of information, ideas, and concepts necessary for academic success in the content area of Mathematics
4. Communication of information, ideas, and concepts necessary for academic success in the content area of Science
5. Communication of information, ideas, and concepts necessary for academic success in the content area of Social Studies

The tests are administered in five grade clusters:

1. Kindergarten
2. Grades 1–2
3. Grades 3–5
4. Grades 6–8
5. Grades 9–12

Six levels of English language proficiency are measured by the ACCESS for ELLs. Students performing at Level 1 have acquired very little English and communicate using single words, phrases, and simple statements or questions. The performance of students at Level 6 is on par with that of their peers whose first language is English. Language is no longer a barrier to academic achievement for these students. Test forms of the ACCESS for ELLs for grades 1–12 are selected by tier, depending upon students' current approximate proficiency level. The Tier A form is for beginners, and it includes test content intended to measure Levels 1–3. The majority of students take the Tier B form, which measures Levels 2–4. Students who may be ready to exit an English language program take Tier C, which measures Levels 3–6. The kindergarten test is not tiered but is adapted to each student's performance during the administration.

In grades 1–12, the listening, reading, and writing tests are administered to groups of students. In the listening test, students listen to audio stimuli and answer multiple-choice questions. The reading test also contains multiple-choice questions that are related to passages that increase in length and complexity by targeted proficiency level. The writing test requires students to write to prompts that range from single letters, words, and sentences at the lowest proficiency level to several paragraphs at the high proficiency levels and grades. The speaking tests are administered one-on-one. Test Administrators follow a script to conduct a conversation with the student. For all four skill domains, the stimuli and questions are aligned to the content areas listed above and represent the range of proficiency levels included in a given form's tier. Graphic stimuli and supports play a large role in the tests for all skill domains, tiers, and grades.

All four domains in the Kindergarten ACCESS for ELLs are administered one-on-one. The listening and speaking domains are assessed together, followed by the writing and reading domains. The entire test takes about 45 minutes to administer.

Alternate ACCESS for ELLs

The Alternate ACCESS for ELLs (Assessing Comprehension and Communication in English State-to-State for English Language Learners) is a test of English language proficiency in reading, writing, listening, and speaking administered to ELs with significant cognitive disabilities in grades 1–12. The tests for all four skill domains are aligned to the WIDA English language development standards describing performance in social and instructional language, language of Language Arts, language of Mathematics, and language of Science.

The tests are administered in four grade clusters:

1. Grades 1–2
2. Grades 3–5
3. Grades 6–8
4. Grades 9–12

Six levels of English language proficiency are measured by the Alternate ACCESS for ELLs. Students performing at Levels A1, A2, and A3 may have minimal, basic communication skills in English. These three levels describe performance that is below Level 1 on the ACCESS for ELLs. A student performing at Level A1 may communicate using gestures, eye gaze, and imitations of sounds. Students performing at Level A3 may use familiar words and practiced, routine phrases. Levels P1, P2, and P3 describe performance that shares some characteristics of performance at Levels 1, 2, and 3 on ACCESS for ELLs, but these levels are not equivalent on the two tests. Students taking the Alternate ACCESS for ELLs may achieve up to a Level P2 in Reading, Listening, and Speaking, and up to a Level P3 in Writing. Level P1 performance is characterized by phrase-level communication. Students performing at Levels P2 and P3 can communicate using sentence-level discourse.

Unlike the ACCESS for ELLs, the Alternate ACCESS for ELLs has only one form per grade cluster. Test Administrators follow a script to administer all four domains of the test to students one-on-one. In the listening test, students listen to prompts read by the Test Administrator and demonstrate comprehension of multiple-choice questions by pointing to images in the test booklet. The reading test contains single word to sentence-length prompts. Answer options in the multiple-choice questions consist of text supported by graphics. In the listening and reading tests, each task is made up of three cues, and a student may need only one cue to respond to the task or all three. The cues provide increasingly more support so that a student responding to Cue A demonstrates more ability than a student responding to Cue B, and so on. The writing test requires students to write to prompts instructing the student to perform tasks ranging from demonstrating the ability to use a pencil to writing a word or sentence based on information provided in a graphic. In the speaking test, Test Administrators follow a script to elicit spoken language ranging from single sounds to one or more sentences. For all four skill domains, the stimuli and questions are aligned to the content areas of social and instructional language and the language of Language Arts, Mathematics or Science. Graphic stimuli and supports play a large role in the tests for all skill domains and grades.

Graduation Assessment Requirements

In order to be eligible for a diploma from a Minnesota public high school, all students must fulfill graduation assessment requirements. There are different routes to meeting graduation assessment requirements depending on what year students were first enrolled in grade 8. Based on the revisions to Minnesota Statute 120B.30 enacted in 2013, the graduation assessment requirements transitioned from the GRAD requirements to the Career and College Assessments. Note: Technical information on the GRAD is found in a separate technical manual.

Students First Enrolled in Grade 8 in 2010–2011 or Earlier

Districts determine which routes are offered and used to meet the graduation assessment requirements in reading, mathematics, and writing for students first enrolled in grade 8 in 2010–2011 or earlier (likely grade 12 students and older in school year 2014–2015). Students can continue to meet the graduation assessment requirements through the high school Standards-Based Accountability Assessments, GRAD retests, or GRAD alternate routes, which include:

- Earning a proficient score on a high school Minnesota Standards-Based Accountability Assessment. If students were proficient (achieve Meets or Exceeds the Standards) on the grade 10 Reading MCA, MCA-Modified, or MTAS or grade 11 Mathematics MCA, MCA-Modified, or MTAS, they have met their graduation assessment requirement for that subject. The MCA-Modified was last administered in 2013–2014.
- Earning a passing score on the Written Composition, Reading, and Mathematics GRAD retests or have the Minnesota Alternate Assessment: Writing completed.
    - For the Written Composition GRAD retests, a student writes to one prompt, and his or her essay is assigned a score between 0 and 6 based on the rater’s overall (holistic) impression of the writing. A score of 3 or higher on the Written Composition GRAD is passing. The holistic scoring rubric used for scoring the Written Composition GRAD is included in the test specifications.
    - For the Reading and Mathematics GRAD retests, the passing score corresponds to a scale score of 50 (or above) on a scale score range of 15 to 85 for both Reading and Mathematics GRAD retests.
    - The Minnesota Alternate Assessment: Writing is the alternate assessment for the Written Composition GRAD. Students who have had this assessment completed for them have met the graduation assessment requirement for writing.
- Meeting GRAD alternate routes, which includes the following:
    - Meet mathematics alternate pathway requirements (only for students enrolled in grade 8 through 2009–2010).
    - Receive an individual passing score (for students on an IEP or 504 plan).
    - Receive an English Learner (EL) exemption.
    - Passing an accountability assessment from another state approved by MDE (reciprocity).

For students first enrolled in grade 8 in 2010–2011 or earlier, the revisions to Minnesota Statute 120B.30 also allow students to take the ACT, the WorkKeys, the ACT Compass, or the Armed Services Vocational Aptitude Battery (ASVAB) to meet graduation assessment requirements in reading, mathematics, and writing. In addition districts are able to substitute a score from an alternative, equivalent assessment to satisfy the graduation assessment requirements. The selection of an equivalent assessment is a district decision, but students must meet requirements in writing, reading, and mathematics.

Students First Enrolled in Grade 8 in 2011–2012

Students first enrolled in grade 8 in 2011–2012 (likely grade 11 students in school year 2014–2015), who were in attendance at the start of the school day on April 28, were expected to take the ACT Plus Writing during the statewide administration in order to meet graduation assessment requirements. Students unable to take the ACT Plus Writing on April 28 were to take the test on the statewide make-up date, May 12. A specific passing score is not required because these assessments are designed to provide information about career and college readiness. If a student was unable to participate in the statewide administration of the grade 11 ACT Plus Writing, students can meet the graduation assessment requirements in reading, mathematics, and writing through any combination of the options outlined above under Students First Enrolled in Grade 8 in 2010–2011 or Earlier as long as requirements are met in all three subjects. The ACT Plus Writing is administered on paper.

Students First Enrolled in Grade 8 in 2012–2013 and Later

Students first enrolled in grade 8 in 2012–2013 and later (likely grade 10 students and younger in school year 2014–2015) meet graduation assessment requirements through participation in the Career and College Assessments; a specific passing score is not required because these assessments are designed to provide information about career and college readiness.

Districts must provide the opportunity for all students to participate in the Career and College Assessments. A student will not be denied a diploma because the student did not participate in a Career and College Assessment.

The **grade 8 Career and College Assessment** is a required graduation assessment that provides information to grade 8 students, their families, and educators about students’ achievement in Reading, English, Mathematics, and Science. In 2014–2015, the grade 8 Career and College Assessment was the ACT Explore, which was administered on paper. Note: the grade 8 Career and College Assessment was not available for grade 8 students in school years 2012–2013 and 2013–2014 so these students are not required to take this assessment to meet graduation assessment requirements.

The **grade 10 Career and College Assessment** is a required graduation assessment that provides information to grade 10 students, their families, and educators about students’ achievement in Reading, English, Mathematics, and Science. In 2014–2015, the grade 8 Career and College Assessment was the ACT Plan, which was administered on paper.

ACT Plus Writing is a required graduation assessment that provides information to grade 11 students, their families, and educators regarding the level of preparedness for postsecondary success on a nationally recognized college entrance exam. The ACT Plus Writing is a paper assessment that must be administered on state-assigned test dates.

Minnesota Statute 120.30B allows students with an IEP who are eligible for the MTAS to take the MTAS in one or more subjects in place of the Career and College Assessments to meet graduation assessment requirements.

Chapter 2: Test Development

The test-development phase of each Minnesota Assessment includes several activities designed to ensure the production of high-quality assessment instruments that accurately measure the achievement of students with respect to the knowledge and skills contained in the Minnesota Academic Standards. The Standards are intended to guide instruction for students throughout the state. Tests are developed according to the content outlined in the *Minnesota Academic Standards* at each grade level for each tested subject area. In developing the *Standards*, committees review curricula, textbooks, and instructional content to develop appropriate test objectives and targets of instruction. These materials may include the following:

- National curricula recommendations by professional subject matter organizations
- *College and Work Readiness Expectations*, written by the Minnesota P-16 Education Partnership working group
- Standards found in the American Diploma Project of Achieve, Inc. (http://www.achieve.org)
- Recommended Standards for Information and Technology Literacy from the Minnesota Educational Media Organization (MEMO; http://aect.site-ym.com/?page=minnesota_educationa)
- Content standards from other states

Test-Development Procedures

The following steps summarize the process followed to develop a large-scale criterion-referenced assessment such as the Minnesota Comprehensive Assessments-Series III (MCA-III) and Minnesota Test of Academic Skills (MTAS):

1. *Development of Test Specifications.* Committees of content specialists develop test specifications that outline the requirements of the test, such as eligible test content, item types and formats, content limits, and cognitive levels for items. These specifications are published as a guide to the assessment program. Committees provide advice on test models and methods to align the tests with instruction. Information about the content, level of expectation, and structure of the tests is based on judgments made by Minnesota educators, students, and the public. Minnesota educators guide all phases of test development.
2. *Development of Items, Stimuli (Passages and Scenes), and Tasks.* Using the *Standards* and test specifications, the Minnesota Department of Education (MDE) Statewide Testing Division staff and Minnesota's testing contractor work with the item development contractor to develop items, stimuli (including Reading passages and Science scenes), and tasks.
3. *Item (and Stimulus) Content Review.* All members of the assessment team review the developed items (and stimuli for Reading and Science), discuss possible revisions, and make changes when necessary.
4. *Item (and Stimulus) Content Review Committee.* Committees of expert teachers review the items (some of which are revised during content review) for appropriate difficulty, grade-level specificity, and potential bias and sensitivity issues.
5. *Field-Testing.* Items are taken from the item content review committees, with or without modifications, and are field-tested as part of the assessment program. Data are compiled regarding student performance, item difficulty, discrimination, reliability, and possible bias.

6. *Data Review.* Committees review the items in light of the field-test data and make recommendations regarding the inclusion of the items in the item bank from which forms are built.
7. *New Forms Construction.* Items are selected for the assessment according to test specifications. Selection is based on content requirements as well as statistical (equivalent passing rates and equivalent test form difficulty) and psychometric (reliability, validity, and fairness) considerations.

More detailed information regarding each step is provided in subsequent sections of this chapter.

Test Specifications

Criterion-referenced tests such as Minnesota's statewide tests are intended to estimate student knowledge within a domain such as mathematics, reading, or science proficiency. The characteristics of the items making up the domain must be specified and are known as the *test specifications.* They provide information to test users and test constructors about the test objectives, the domain being measured, the characteristics of the test items, and how students will respond to the items. Test specifications are unique for each test and lay the framework for the construction of a test.

Test specifications developed by MDE since 2005 have been designed to be consistent in format and content, thereby making the testing process more transparent to the education community. The tests being developed are based on content standards defined by committees of Minnesota teachers. Thus, the content standards and their strands, substrands, and benchmarks serve as the basis for the test specifications. Item types, cognitive levels of understanding to be tested, range in the number of items, and content limits are assigned to each benchmark within the standards.

The item formats are constrained by the test delivery system (paper or online). The item format determines how the student responds to the item, such as selecting an answer, writing a response, or manipulating images on a computer screen.

The cognitive level of understanding for an item is determined by the type of cognition required for a correct response to the item. Teacher committees consider what types of cognition are appropriate for different content in order to determine the assigned cognitive levels for each benchmark. Cognitive levels for benchmarks are determined independently of the item formats and difficulty of the content; this runs counter to many people's perceptions that cognitive level and content difficulty are equivalent concepts. For example, a benchmark measured at a high cognitive level could be assessed with different item formats, such multiple-choice or technology enhanced item.

Similarly, the ranges in number of items and content limits are based on discussion among the educators in the committees about two things: the emphasis a benchmark is given in the classroom and the type of curriculum content regularly taught to students in a grade level. This discussion guides the final information entered in the test specifications.

Test specifications facilitate building a technically sound test that is consistent from year to year. They demonstrate MDE's respect for teacher concerns about the time students spend taking tests, and they take into account the grade and age of students involved as well as other pedagogical concerns. Test specifications define, clarify, and/or limit how test items will be written. They can be used by schools and districts to assist in the planning of curricula and instruction to implement the Minnesota standards. The test specifications also provide a basis for interpreting test results.

The remainder of this section provides details about the development of test specifications for each test in the Minnesota Assessment System.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

To develop the Minnesota Comprehensive Assessments-Series III (MCA-III), MDE held meetings with Minnesota educators to define general test specifications for each grade. Minnesota classroom teachers, curriculum specialists, administrators, and university professors served on committees organized by grade and subject area. MDE chose committee members to represent the state in terms of geographic region, type and size of school district, and the major ethnic groups found in Minnesota.

The committees identified strands, standards, and benchmarks of the Minnesota Academic Standards to be measured in the tests. Some strands, standards, or benchmarks were not suitable for the large-scale assessments. These were clearly identified as content to be assessed in the classroom.

After the measurable components of the standards were identified, teacher committees set item formats, cognitive levels, and content limits for each benchmark. Item prototypes were developed as part of the development of the test specifications.

Committees of Minnesota educators reviewed drafts of these specifications, and their suggestions were incorporated into the final versions of the test specifications. The complete MCA-III test specifications document for each subject is available on the MDE website at http://www.education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html.

Minnesota Test of Academic Skills

Criteria outlined by the National Alternate Assessment Center served as a guide in the development of the Minnesota Test of Academic Skills (MTAS) to help ensure that items were based on the Minnesota grade-level academic standards. All the content of the MTAS is academic and derived directly from the Minnesota grade-level academic standards in reading, mathematics, and science.

A systematic and iterative process was used to create the MTAS test specifications. Prior to the on-site benchmark extensions meetings, MDE met with stakeholder groups and their vendors (Minnesota's testing contractor and ILSSA) to identify preliminary benchmarks at each grade level that would be finalized after a public comment period. The process was guided by test alignment criteria and balanced by characteristics of students with significant cognitive disabilities, as listed below.

1. The grade-level benchmark was assessed on the MCA.
2. Proficiency on the benchmark will aid future learning in the content area for students with significant cognitive disabilities.
3. Proficiency on the benchmark will help the student in the next age-appropriate environment for students with significant cognitive disabilities (that is, the next grade in school or a post-school setting).
4. A performance task can be written for the benchmark without creating a bias against a particular student population.

The benchmark contributed to the pattern of emphasis on the test blueprint for the MTAS, including multiple substrands, cognitive levels, and benchmarks.

The recommended benchmarks were taken to teacher groups who were tasked with developing the extended benchmarks. Benchmark extensions represent a reduction in the depth or complexity of the benchmark while maintaining a clear link to the grade-level content standard. During the meetings, the teachers scrutinized the recommended benchmarks using their professional expertise and familiarity with the target student population and made changes to a subset of the recommended benchmarks in reading, mathematics, and science.

Content limits had been written and approved for the MCA-III but needed to be reviewed and further revised for the MTAS for each of the recommended benchmarks. During the benchmark extension writing sessions, the groups were instructed to review the content limits for the general assessment. If those content limits were sufficient, no other content limits were needed. However, if the groups felt strongly that only certain components of a benchmark should be assessed in this student population, they added this information to the content limits.

The next step for Minnesota educators who served on the benchmark extension panel was to determine the critical learner outcome represented by each prioritized benchmark in reading, mathematics and science. The critical outcome is referred to as the essence of a benchmark and can be defined as the most basic skill inherent in the expected performance. These critical outcomes are called *essence statements.* Panel members wrote sample instructional activities to show how students with the most significant cognitive disabilities might access the general education curriculum represented by the essence statement. Once panel members had a clear picture of how a skill might be taught, they wrote benchmark extensions. Three extensions were written for each benchmark to show how students who represent the diversity within this population could demonstrate proficiency on the benchmark.

MDE recognizes that the students who take the MTAS are a heterogeneous group. To help ensure that every student in this group has access to the test items, student communication modalities were considered and accommodations were made. Six teacher groups composed of curriculum experts and both special and general educators were convened to write these entry points for three grade bands in reading and mathematics and each grade-level assessment in science. After approximately one-half day of training, the teacher groups wrote entry points, in essence each of the selected benchmarks included on the MTAS. The process included the following steps:

1. A curriculum specialist described the intent or underlying essence of the benchmark.
2. A general educator described a classroom activity or activities in which the benchmark could be taught.
3. A special educator described how the activity or activities could be adapted to include a student with significant cognitive disabilities.

At each step, the group verified that the benchmark was still being addressed, the general education activity was still appropriate, and the student could still access the content in a meaningful way. The groups then developed an assessment activity for each type of learner, including the different types of supports that might be used. After writing each assessment activity, the group reviewed the activity to check that it maintained the integrity of the original instructional activity and the essence of the benchmark.

The specifications were published on the MDE website for public review in December 2006. In order to update the assessment to align to the *2007 Minnesota K–12 Academic Standards in Mathematics,* test specifications for grades 3–8 Mathematics MTAS were published in 2011 and test specifications for

grade 11 were published in 2013. The complete MTAS test specification documents are available on the MDE website at
http://www.education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html.

Title III Assessments

ACCESS for ELLs

The forms of the 2013 ACCESS for ELLs are based on the 2012 WIDA English language development standards (Social and Instructional Language, Language of Language Arts, Mathematics, Science and Social Studies) and WIDA performance definitions, which describe the linguistic complexity, language forms and conventions, and vocabulary used by students at six proficiency levels. The 2013 forms of the Alternate ACCESS for ELLs are based on four of the WIDA English language development standards (Social and Instructional Language, Language of Language Arts, Mathematics and Science) and alternate performance definitions at six proficiency levels. Documents describing these standards and performance definitions are available in the Download Library on the WIDA website at http://www.wida.us/downloadLibrary.aspx. Refer to the *WIDA 2012 ELD Standards* under the "Standards" heading.

WIDA does not publish test blueprints or specifications to its website, but district staff who have logins to the secure portal of the WIDA website can access the Test Administration Manuals. These manuals contain limited information about the organization of the tests. The Test Administration Manuals are available under the "ACCESS for ELLs" heading on the same page as indicated above.

Item Development

This section describes the item writing process used during the development of test items (including stimuli) and, in the case of the Minnesota Test of Academic Skills (MTAS), performance tasks. Minnesota's testing contractor has the primary role for item and task development; however, MDE personnel and state review committees also participate in the item development process. Item and task development is a complex multistage process.

Items and tasks are written and internally reviewed by the testing contractor before submission to MDE. For each subject and grade, MDE receives an item tally sheet displaying the number of test items by benchmark and target. Item tallies are examined throughout the review process. Additional items are written by the testing contractor, if necessary, to complete the requisite number of items per benchmark.

Content Limits and Item Specifications

Content limits and item specifications identified in the test specifications are strictly followed by item writers to ensure accurate measurement of the intended knowledge and skills. These limits were set using committee feedback, Minnesota Department of Education (MDE) input, and use of the standards, as mandated by federal and state law.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Item specifications are provided for each assessed benchmark for the Minnesota Comprehensive Assessments-Series III (MCA-III). The item specifications provide restrictions of numbers, notation, scales, context, and item limitations/requirements. The item specifications also list symbols and vocabulary that may be used in items. This list is cumulative in nature. For example, symbols and vocabulary listed at grade 3 are eligible for use in all grades that follow (grades 4–8).

Minnesota Test of Academic Skills

The content limits of the MTAS provide clarification of the manner in which the depth, breadth, and complexity of the academic standards have been reduced. In mathematics, this might concern the number of steps required of a student to solve a problem. In reading, this could involve a restriction in the number of literary terms assessed within a benchmark. In science, this might be addressed by requiring knowledge of only major aspects of the water cycle.

Title III Assessments

ACCESS for ELLs

The complexity of tasks that are called for by the ACCESS for ELLs is limited by the WIDA performance definitions, which describe the linguistic complexity, language forms and conventions, and vocabulary used by students at five proficiency levels. The sixth proficiency level represents the end of the proficiency scale continuum and is characterized by performance that meets all criteria through Level 5. The tasks for the Alternate ACCESS for ELLs are based on WIDA's alternate model performance indicators, which describe the linguistic complexity, language forms and conventions, and vocabulary usage in the performance of EL with significant cognitive disabilities. Performance for the six levels of proficiency is described in the Alternate ACCESS for ELLs Performance Definitions. Documents describing the performance definitions for both assessments are available in the Download Library on the WIDA website at http://www.wida.us/downloadLibrary.aspx. Refer to the *WIDA 2007 ELP Standards* under the "Standards" heading.

Item Writers

Minnesota's testing contractor uses item writers who have extensive experience developing items for standardized achievement tests. The contractor selects item writers for their knowledge of the specific content area and for their experience in teaching or developing curricula for the relevant grades.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Minnesota's testing contractor employs item writers who are accomplished and successful in meeting the high standards required for large-scale assessment items. Most item writers are former teachers who have substantial knowledge of curriculum and instruction for their content area and grade levels. Item writers must go through rigorous training and are retained only after demonstrating competency during this training.

Minnesota Test of Academic Skills

In addition to meeting the standards described above, item writers must have experience with and a clear understanding of the unique needs of students with significant cognitive disabilities with respect to their ability to provide responses to the performance tasks.

MTAS item writers comprise both general and special education teachers. Item writing assignments for each grade level and subject area are divided between both general and special education teachers to ensure coverage of the content breadth as well as ensuring maximum accessibility for students with significant cognitive disabilities. Item writer training includes an overview of the requirements for alternate assessments based on alternate achievement standards, characteristics of students with significant cognitive disabilities, descriptions of performance-based tasks, principles of universal design, the MTAS Test Specifications, and the MTAS Essence Statements. Throughout the item writing process, evaluative feedback is provided to item writers by contractor content and alternate assessment specialists to ensure submission of performance tasks that meet the grade level, content, and cognitive requirements.

Title III Assessments

ACCESS for ELLs

The Center for Applied Linguistics (CAL) is contracted by WIDA to develop items and construct test forms for the ACCESS for ELLs and the Alternate ACCESS for ELLs. CAL has extensive experience in language proficiency test development and has item writers on staff dedicated to the WIDA consortium and its assessments.

Item Writer Training

Minnesota's testing contractor and MDE provide extensive training for writers prior to item or task development. During training, the content benchmarks and their measurement specifications are reviewed in detail. In addition, Minnesota's testing contractor discusses the scope of the testing program, security issues, adherence to the measurement specifications, and avoidance of economic, regional, cultural, and ethnic bias. Item writers are instructed to follow commonly accepted guidelines for good item writing.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Minnesota's testing contractor conducts comprehensive item writer training for all persons selected to submit items for the MCA-III. Training includes an overview of the test-development cycle and specific training in the creation of high quality multiple-choice, constructed-response, and figural-response items. Experienced contractor staff members lead the trainings and provide specific and evaluative feedback to participants.

Minnesota Test of Academic Skills

Minnesota's testing contractor conducts item writer training for the MTAS that focuses on including students with significant cognitive disabilities in large-scale assessments. Item writers are specifically trained in

- task elements
- vocabulary appropriateness
- fairness and bias considerations
- significant cognitive disability considerations

Minnesota's testing contractor recruits item writers who have specific experience with special populations, and the focus of the training is on the creation of performance tasks and reading passages.

Performance tasks must

- match the expected student outcomes specified in the Benchmark Extensions document;
- follow the format of the template provided by the testing contractor;
- clearly link to the essence statement and be unique;
- represent fairness and freedom from bias;
- represent high yet attainable expectations for students with the most significant cognitive disabilities;
- include clearly defined teacher instructions and student outcomes; and
- lend themselves to use with assistive technology and other accommodations.

Item Review

Contractor Review

Experienced testing contractor staff members, as well as content experts in the grades and subject areas for which the items (including stimuli) or performance tasks were developed, participate in the review of each set of newly developed items. This annual review for each new or ongoing test checks for the fairness of the items and tasks in their depiction of minority, gender, and other demographic groups. In addition, Minnesota's testing contractor instructs the reviewers to consider other issues, including the appropriateness of the items and tasks to the objectives of the test, difficulty range, clarity, correctness of answer choices, and plausibility of the distractors. Minnesota's testing contractor asks the reviewers to consider the more global issues of passage appropriateness, passage difficulty, and interactions between items within and between passages, as well as artwork, graphs, or figures. The items are then submitted to the Minnesota Department of Education (MDE) for review.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Before an item may be field-tested for the Minnesota Comprehensive Assessments-Series III, it must be reviewed and approved by the Content Committee and the Bias and Fairness Committee. The Content Committee's task is to review the item content and scoring rubric to assure that each item

- is an appropriate measure of the intended content (strand, substrand, standard, and benchmark);
- is appropriate in difficulty for the grade level of the examinees;
- has only one correct or best answer (for multiple-choice items); and
- has an appropriate and complete scoring guideline (for technology-enhanced items).

The Content Committees can make one of three decisions about each item: approve the item and scoring rubric as presented; conditionally approve the item and scoring rubric with recommended changes or item edits to improve the fit to the strand, substrand, standard, and benchmark; or eliminate the item from further consideration.

Each test item is coded by content area and item type (for example, multiple-choice, technology-enhanced) and presented to MDE Assessment Specialists for final review and approval before field-testing. The final review encompasses graphics, artwork, and page layout.

The Bias and Fairness Committee reviews each item to identify language or content that might be inappropriate or offensive to students, parents, or community members or that contain stereotypical or biased references to gender, ethnicity, or culture. The Bias and Fairness Committee accepts, edits, or rejects each item for use in field tests.

Minnesota Test of Academic Skills

The Minnesota Test of Academic Skills (MTAS) has been aligned with the academic content standards (i.e., Minnesota Academic Standards) established for all students.

Assessments have been developed in grades 3–8 and high school for both mathematics and reading; assessments in science have been developed for grades 5 and 8 and high school. The science and mathematics tests consist of a series of discrete items. In reading, the tasks are designed to assess comprehension of the MTAS passages. Reading passages for the MTAS differ from those appearing on the MCA-IIIs. The MTAS passages are shorter (approximately 200 words or less), and the overall difficulty level is reduced. The content of the passages is less complex. Passages are written to include simple sentence structures, high frequency words, decodable words, and repeated words and phrases. MTAS passages feature clear, concise language. In general, passages mirror high-interest/low-level materials that are accessible for instruction for this population.

The Reading MTAS includes both fiction and nonfiction passages. Passage topics are age appropriate and generally familiar to the population assessed. Concepts presented in the passages are literal.

Before a passage or item may be field-tested, it must be reviewed and approved by the Content Committee and the Bias and Fairness Committee. The Content Committee’s task is to review the item content and scoring rubric to assure that each item

- is an appropriate measure of the intended content;
- is appropriate in difficulty for the grade level of the examinees; and
- has only one correct or best answer for each multiple-choice item.

The Content Committees can make one of three decisions about each item: (1) approve the item and scoring rubric as presented; (2) conditionally approve the item and scoring rubric with recommended changes or item edits to improve the fit to the strand, substrand, standard, and benchmark; or (3) eliminate the item from further consideration.

The Bias and Fairness Committee reviews each passage and item to identify language or content that might be inappropriate or offensive to students, parents, or community members or that contain stereotypical or biased references to gender, ethnicity, or culture. The Bias and Fairness Committee accepts, edits, or rejects each item for use in field tests.

Each test item is coded by content area and presented to MDE Alternate Assessment Specialists for final review and approval before field-testing. The final review encompasses graphics, artwork, and page layout.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

The Center for Applied Linguistics (CAL) contacts WIDA member state education agencies (SEAs) to recruit educators from across the consortium to participate in item and bias reviews. Following field-testing of new items, CAL also asks SEAs to recommend educators to serve on data review panels.

MDE Review

Staff at MDE and Minnesota's testing contractor review all newly developed items and tasks prior to educator committee review. During this review, content assessment staff scrutinize each item for content-to-specification match, difficulty, cognitive demand, plausibility of the distractors, rubrics, and sample answers and for any ethnic, gender, economic, or cultural bias.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Content assessment staff from MDE and Minnesota's testing contractor discuss each item, addressing any concerns during this review. Edits are made accordingly, prior to item review with teachers.

Minnesota Test of Academic Skills

Assessment staff with both content and students-with-disabilities expertise from MDE and Minnesota's testing contractor discuss each item, addressing any concerns during this review. Edits are made accordingly, prior to item review with teachers.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

All development and review for the ACCESS for ELLs and Alternate ACCESS for ELLs is performed by the Center for Applied Linguistics and WIDA. Consortium member states do not review items as a matter of course, although they may send SEA staff to participate in item, bias, and data reviews.

Item Committee Review

During each school year, MDE convenes committees composed of K–12 and higher-education teachers, curriculum directors, and administrators from across Minnesota to work with MDE staff in reviewing test items (including stimuli) and performance tasks developed for use in the assessment program.

MDE seeks recommendations for item review committee members from Best Practice Networks, district administrators, district curriculum specialists, and subject-area specialists in MDE's Curriculum Division and other agency divisions. MDE selects committee members based on their recognized accomplishments and established expertise in a particular subject area. Committee members represent the regions of the state and major ethnic groups in Minnesota, as well as various types of school districts (such as urban, rural, large, and small districts).

Each school year, Minnesota educator committees review all newly developed test items and tasks and all new field-test data. Approximately 40 committee meetings are convened, involving Minnesota educators who represent school districts statewide.

MDE Research and Assessment staff, along with measurement and content staff from Minnesota's testing contractor, train committee members on the proper procedures and the criteria for reviewing newly developed items. Reviewers judge each item for its appropriateness, adequacy of student preparation, and any potential bias. Prior to field-testing, committee members discuss each test item and recommend whether the item should be field-tested as written, revised, or rejected. During this review, if committee members judge an item questionable for any reason, they may recommend the item be removed from consideration for field-testing. During their reviews, all committee members consider the potential effect of each item on various student populations and work toward eliminating bias against any group.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

Item review committees are composed of content teachers in English language arts (ELA), mathematics, and science. Within a given content area, teachers are invited so that the committee appropriately represents the state in terms of geography, ethnicity, and gender. Teachers are also selected to represent English as a second language (ESL) and special education licensures. Content area educators serving on these committees are familiar with the Minnesota Academic Standards. Items are reviewed according to an 11-point checklist (presented below) to ensure alignment to *the Standards*. Teachers' discussion of the test items is facilitated by MDE and its testing contractor.

Item Review Checklist

1. Does the item have only one correct answer?
2. Does the item measure what it is intended to measure?
3. Is the cognitive level appropriate for the level of thinking skill required?
4. Is the item straightforward and direct with no unnecessary wordiness?
5. Are all distractors plausible yet incorrect?
6. Are all answer options homogeneous?
7. Are there any clues or slang words used that may influence the student's responses to other items?
8. Is the intent of the question apparent and understandable to the student without having to read the answer options?
9. Do all items function independently?
10. Are all items grammatically correct and in complete sentences whenever possible?
11. Reading items: Does the item require the student to read the passage in order to answer the question?

Minnesota Test of Academic Skills

Item review committees are composed of special education and content teachers in English language arts, mathematics, and science. Within a given content area, these two areas of expertise are equally represented, to the extent possible, and MDE makes a special effort to invite teachers who are licensed in both areas. Many content area educators serving on these committees have also served on item review panels for the MCA-III and are therefore very familiar with the Minnesota Academic Standards. The collaboration between special education and content area teachers ensures that MTAS assesses grade-level standards that have been appropriately reduced in breadth, depth, and complexity for students with the most significant cognitive disabilities.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

Item review committees are convened by the Center for Applied Linguistics and WIDA. These organizations follow industry standards when conducting item review committee meetings.

Bias and Sensitivity Review

All items placed on Minnesota assessments are evaluated by a panel of teachers and community experts familiar with the diversity of cultures represented in Minnesota. This panel evaluates the fairness of passages, storyboards, and test items for Minnesota students by considering issues of gender, cultural diversity, language, religion, socioeconomic status, and various disabilities.

Field-Testing

Before an item can be used on a live test form, it must be field-tested. MDE uses two approaches to administer field-test items to large, representative samples of students: embedded items and stand-alone administrations.

Embedded Field-Testing

Whenever possible, MDE embeds field-test items in multiple forms of operational tests so that the field-test items are randomly distributed to students across the state. This ensures that a large representative sample of responses is gathered under operational conditions for each item. Past experience has shown that these procedures yield sufficient data for precise statistical evaluation of a large number of field-test items in an authentic testing situation. The number of students responding to each item is listed among the item analysis data presented to the data review committees. Currently, responses to most field-test items are obtained from thousands of students. Enough forms are produced annually to result in a number of items sufficient for replenishing and improving the item pools.

Performance on field-test items does not contribute to a student’s scores on the operational tests. The specific locations of the embedded items on a test form are not disclosed. These data are free from the effects of differential student motivation that may characterize stand-alone field-test designs because the items are answered by students taking actual tests under standard administration procedures.

Stand-Alone Field-Testing

When MDE implements testing at new grade levels or for new subject areas, it is necessary to conduct a separate stand-alone field test in order to obtain performance data. When this type of field-testing is required, MDE requests volunteer participation from the school districts. MDE has been successful in obtaining volunteer samples that are representative of the state population.

To make certain that adequate data are available to appropriately examine each item for potential ethnic bias, MDE designs the sample selection in such a manner that the proportions of minority students in the samples are representative of their total student populations in Minnesota. School districts are notified in advance about which schools and classes are chosen for the administration of each test form so that any problems related to sampling or to the distribution of materials can be resolved before the test materials arrive.

Data Review

Data Review Committees

MDE convenes data review committees composed of Minnesota teachers and curriculum and assessment specialists. Much effort goes into ensuring that these committees of Minnesota educators represent the state demographically with respect to ethnicity, gender, size of school district, and geographical region. These committees receive training on how to interpret the psychometric data compiled for each field-test item. Minnesota's testing contractor supplies psychometricians (typically persons with an advanced degree in the application of statistical analyses to measurement), content experts (usually former teachers and item writers), and group facilitators for the data review committee meetings.

Data obtained from the field test include

- numbers of students by ethnicity and gender in each sample;
- percentage of all students choosing each response;
- students distributed into thirds based on performance on the overall test and that group of students' distribution choosing each response;
- percentage of students, by gender and by major ethnic group, choosing each response;
- point-biserial correlations summarizing the relationship between a correct response on a particular test item and the score obtained on the total subject area test; and
- item response theory (IRT) and Mantel-Haenszel statistical indices to describe the relative difficulty and discrimination of each test item and to identify greater-than-expected differences in performance on an item associated with gender and ethnicity.

Specific directions are provided on the use of the statistical information and review booklets. Committee members evaluate each test item with regard to benchmark and instructional target match, appropriateness, level of difficulty, and bias (cultural, ethnic, gender, geographic and economic) and then recommend that the item be accepted, rejected, or revised and field-tested again. Items that pass all stages of development—item review, field-testing, and data review—are placed in the "item bank" and become eligible for use on future test forms. Rejected items are noted and precluded from use on any test form.

Statistics Used

In order to report the field-test results, MDE requires that various statistical analyses, based on classical test theory and item response theory, be performed. Item response theory, more completely described in Chapter 6, comprises a number of related models, including Rasch-model measurement (Wright, 1977; Masters, 1982), the two-parameter and three-parameter logistic models (Lord & Novick, 1968), and the generalized partial credit model (Muraki, 1992). An outline is given to each committee member about the types of field-test data they review in order to determine the quality of each item. Two types of differential item functioning (DIF)—i.e., item bias—data are presented during committee review: Mantel-Haenszel Alpha and its associated chi-square significance and item response distributions for each analysis group.

The Mantel-Haenszel Alpha statistic is a log-odds ratio indicating when it is more likely for one of the demographic groups to answer a particular item correctly than for another group at the same ability level. When this probability is significantly different across the various ability strata, the item is flagged for further examination.

Response distributions for each demographic group give an indication of whether members of a group were drawn to one or more of the answer choices for the item. If a large percentage of a particular group selected an answer chosen significantly less often by other groups, the item should be inspected carefully.

Several pieces of summary statistical information are also provided. The item mean and item-total correlation are general indicators of item difficulty and quality. The response distribution for all students is used by the data review committee to evaluate the attractiveness of multiple-choice distractors and determine the effectiveness of the constructed-response items in identifying and awarding partial credit responses.

Finally, the IRT item parameters and a fit index are provided. The IRT model must fit student responses for the scaling and equating procedures used by MDE to be valid. The primary item parameters provided measure the item’s relative difficulty and the item’s capability of separating low performers from high performers. The review committee uses these values to identify items that might be undesirable for inclusion in the item pool.

Title I Assessments

Minnesota Comprehensive Assessments-Series III

The first data review meetings for the Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) grades 3–8 were held in March 2010. Items reviewed at these meetings were field-tested in a stand-alone online field test conducted in fall 2009. MCA-III data reviews use the procedures described previously. Panelists are invited to the workshops according to procedures established by MDE that attempt to provide broad representation of expertise, ethnicity, school size, and geography.

Minnesota Test of Academic Skills

The Minnesota Test of Academic Skills (MTAS) data reviews use the procedures described previously. Emphasis is placed on inviting panelists who have content and/or special education expertise. In addition to the data displays common to all Minnesota assessments, the MTAS data review panels also consider disaggregated information about performance of students most likely to participate in the MTAS. This disaggregation includes additional score level analysis for students in three categories of disabilities:

1. Developmentally Cognitively Disabled—Mild
2. Developmentally Cognitively Disabled—Severe
3. Autism Spectrum Disorder

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

Data review committees are convened by the Center for Applied Linguistics (CAL) and WIDA. These organizations follow industry standards when conducting data review committee meetings.

Item Bank

Minnesota’s testing contractor maintains the item bank for all tests in the Minnesota assessment program and stores each test item and its accompanying artwork in a database. Additionally, the Minnesota Department of Education (MDE) maintains paper copies of each test item. This system allows test items to be readily available to MDE for test construction and reference and to the testing contractor for test booklet design and printing.

In addition, Minnesota’s testing contractor maintains a statistical item bank that stores item data, such as a unique item number, grade level, subject, benchmark/instructional target measured, dates the item has been administered, and item statistics. The statistical item bank also warehouses information obtained during the data review committee meetings indicating whether a test item is acceptable for use, acceptable with reservations, or not acceptable at all. MDE and Minnesota’s testing contractor use the item statistics during the test construction process to calculate and adjust for differential test difficulty and to check and adjust the test for content coverage and balance. The files are also used to review or print individual item statistics as needed.

Test Construction

The Minnesota Department of Education (MDE) and Minnesota’s testing contractor construct test forms from the pool of items or performance tasks deemed eligible for use by the educators who participated in the field-test data review committee meetings. Minnesota’s testing contractor uses operational and field-test data to place the item difficulty parameters on a common item response theory scale (see Chapter 6, “Scaling”). This scaling allows for the comparison of items, in terms of item difficulty, to all other items in the pool. Hence, Minnesota’s testing contractor selects items within a content benchmark not only to meet sound content and test construction practices but also to maintain comparable item difficulty from year to year.

Minnesota’s testing contractor constructs tests to meet the specifications for the number of test items included for each test benchmark as defined on the test specifications. The Minnesota Academic Standards are arranged in a hierarchical manner where the strand is the main organizational element (e.g., number sense or patterns, functions and algebra). The substrand is the secondary organizational element (e.g., patterns and functions or vocabulary). Each substrand contains one or more standards. Each standard contains one or more benchmarks. Each year’s assessment will assess items in each strand but not necessarily every benchmark. To do so would create a very lengthy assessment. The tests are constructed to measure the knowledge and skills as outlined in the specifications and the standards, and they are representative of the range of content eligible for each benchmark being assessed.

In the cases of Braille and large-print accommodations, it is the goal of MDE to keep all items on an operational form. Items are replaced if they cannot be placed into Braille translation or large-print mode appropriately. To date, Minnesota has been able to meet this goal in all assessments since the current program began in 1997.

Chapter 3: Test Administration

Eligibility for Assessments

As a result of the No Child Left Behind Act (NCLB) and the Elementary and Secondary Education Act (ESEA), all public school students enrolled in grades 3–8, 10, and 11 must be annually assessed with accountability tests. This requirement includes students who receive special education services. In addition, public school English Learners (ELs) in grades K–12 are annually assessed with language proficiency tests.

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series III

General education students and students in special populations—i.e., ELs and students with disabilities (SWDs) who are able to do so—take the Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) to fulfill their mathematics requirement.

Minnesota Test of Academic Skills

Students with IEPs who meet the eligibility criteria of the MTAS as defined in the annually published *Procedures Manual for Minnesota Assessments* are eligible to participate in the Mathematics MTAS to fulfill their mathematics requirement.

Reading

Minnesota Comprehensive Assessments-Series III

General education students—and students with disabilities who are able to do so—take the Reading MCA-III to fulfill their reading requirement.

Minnesota Test of Academic Skills

Students with IEPs who meet the eligibility criteria of the Minnesota Test of Academic Skills (MTAS) are eligible to participate in the Reading MTAS to fulfill their reading requirement.

Science

Minnesota Comprehensive Assessments-Series III

General education students—and students with disabilities who are able to do so—take the Science MCA-III to fulfill their science requirement.

Minnesota Test of Academic Skills

Students with IEPs who meet the eligibility criteria of the MTAS are eligible to participate in the Science MTAS to fulfill their science requirement.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

English Learners in grades K–12 must participate in the WIDA language proficiency assessments. Most ELs take the ACCESS for ELLs. English Learners in grades 1–12 who have significant cognitive disabilities may instead take the Alternate ACCESS for ELLs.

Administration to Students

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series III

The grades 3–8 and 11 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) are administered online, with paper or Braille forms available for students requiring an accommodated form. For grades 3–8, the computer and paper versions are both divided into four segments, allowing for districts administering the test on paper to administer it over two or more days. For grade 11, the computer version is divided into six segments, and the paper version is divided into four segments, allowing for districts administering the test on paper to administer it over two or more days. Students may use a calculator on the entire test, and handheld calculators may be used. Students testing online have the ability to exit the test at the end of each segment, allowing administration over multiple class periods or days. The Minnesota Department of Education (MDE) allows district staff to determine how many test segments will be administered during each testing session. The multiple segments can be administered in a number of ways: all segments at one time, one segment per day, two segments per day, etc. Some segments of the Mathematics MCA-III do not allow calculators to be used in answering questions.

The grades 3–8 and 11 Mathematics MCA online and paper versions are administered any time within the eight-week online testing window.

Minnesota Test of Academic Skills

Any district employee who has received Minnesota Test of Academic Skills (MTAS) test administration training may administer the MTAS. However, the Test Administrator should be a person who is familiar with the student’s response mode and with whom the student is comfortable. All MTAS Test Administrators must be trained or review training materials prior to each test administration. Training can be completed by attending an in-person MDE training (when available), attending a district-provided training or viewing training materials online. The Mathematics MTAS is administered to students in a one-on-one setting. Therefore, Test Administrators must schedule times to administer the tasks.

Although the MTAS is administered in a one-on-one setting, the administration of the assessment is still considered standardized. The design of the assessment and its administration are specified in the *MTAS Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to examinees.

Reading

Minnesota Comprehensive Assessments-Series III

The grades 3–8 and 10 Reading Minnesota Comprehensive Assessments-Series III (MCA-III) is administered online, with paper or Braille forms available for students requiring an accommodated form. The online exam is divided into sections by passage and each passage's associated items; the paper exam is administered in four or five separate segments allowing for districts administering the test on paper to administer it over two or more days. Students testing online have the ability to exit at the end of each segment and return to the test to complete the remaining segments later, allowing administration over multiple class periods or days. Students must complete all items in a segment before exiting. For the paper version, the Minnesota Department of Education (MDE) allows district staff to determine how many test segments will be administered during each testing session. Administration of the multiple segments can be done in a number of ways: all segments at one time, one segment per day, two segments per day, etc.

The grades 3–8 and 10 Reading MCA online and paper versions are administered any time within the eight-week online testing window.

Minnesota Test of Academic Skills

Any district employee who has received MTAS test administration training may administer the MTAS. However, the Test Administrator should be a person who is familiar with the student's response mode and with whom the student is comfortable. All MTAS Test Administrators must be trained or review training materials prior to each test administration. Training can be completed by attending a district-provided training or viewing training materials online. The Reading MTAS is administered to students in a one-on-one setting. Therefore, Test Administrators must schedule times to administer the tasks.

For the Reading MTAS, students may interact with the passage text in one of several presentations: the passage text, a picture-supported passage, a symbolated image representation, or other accommodations appropriate for students' needs. When using one of these presentations, students may read the passage independently, read along as the Test Administrator reads the passage, or have the passage read to them. As a part of the data-collection process, teachers identify what support, if any, students had with the passage. This passage support was used to create the alternate achievement level descriptors (ALDs) and determine performance levels in spring 2008. This level of passage support is also reported on the student report presented to parents.

Prior to allowing students to have these levels of passage support on the Reading MTAS, MDE consulted with national experts on alternate assessments—including staff from the National Alternate Assessment Center as well as the National Center on Educational Outcomes—about the appropriateness of those accommodations. These assessment experts supported MDE's desire to allow for appropriate passage support on the Reading MTAS.

Although the Reading MCA-III does not allow for a read-aloud accommodation, the Reading MTAS is used to assess a very different population. Disallowing an MTAS read-aloud accommodation would make assessment difficult, particularly since the intended population includes students who are communicating at pre-emerging and emerging levels of symbolic language use. Facilitating students' progress toward symbolic language use is essential to reading and literacy. Language development is essential for reading, and the MTAS is designed to assess language development using age- and/or grade-appropriate language passages as documented in the communication literature. Recent research supports this decision. A study by Towles-Reeves, Kearns, Kleinert, and Kleinert (2009) suggests that this reading passage support is appropriate:

> For each of the five options under reading and math, teachers were asked to select the option that best described their students' present performance in those areas. In States 1 and 3, teachers noted that over 2% of the population read fluently with critical understanding in print or Braille. This option was not provided on the inventory in State 2. Almost 14% of the students in State 1, 12% in State 2, and 33% in State 3 were rated as being able to read fluently, with basic (literal) understanding from paragraphs or short passages with narrative or informational texts in print or Braille. The largest groups from all three states (50%, 47%, and 33% in States 1, 2, and 3, respectively) were rated as being able to read basic sight words, simple sentences, directions, bullets, and/or lists in print or Braille, but not fluently from text with understanding. Smaller percentages of students (17%, 14%, and 18%) were rated as not yet having sight word vocabularies but being aware of text or Braille, following directionality, making letter distinctions, or telling stories from pictures. Finally, teachers noted that 15% of students in State 1, 25% of students in State 2, and 13% of students in State 3 had no observable awareness of print or Braille. (p. 245)

Towles-Reeves et al. (2009) go on to cite other research that supports their findings:

> Our results appear consistent with those of Almond and Bechard (2005), who also found a broad range of communication skills in the students in their study (i.e., 10% of the students in their sample did not use words to communicate, but almost 40% used 200 words or more in functional communication) and in their motor skills (students in their sample ranged from not being able to perform any components of the task because of severe motor deficits to being able to perform the task without any supports). Our findings, together with those of Almond and Bechard, highlight the extreme heterogeneity of the population of students in the AA-AAS, making the development of valid and reliable assessments for these students an even more formidable task. (p. 250)

Other research also supports Minnesota's decision to allow students to have the reading passages read to them for the MTAS. In an article for the journal *Remedial and Special Education,* Browder et al. (2009) propose a conceptual foundation for literacy instruction for students with significant cognitive disabilities. The conceptual foundation discussed includes accessing books through listening comprehension. As Browder et al. (2009) note, "To use literature that is grade and age appropriate, books will need to be adapted, including the use of text summaries and key vocabulary. Students who do not yet read independently will need either a technological or human reader" (p. 10).

Although the MTAS is administered in a one-on-one setting, the administration of the assessment is still considered standardized. The design of the assessment and its administration are specified in the *MTAS Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to examinees.

Science

Minnesota Comprehensive Assessments-Series III

The Science MCA-III is a computer-delivered assessment administered as two segments. Students have the ability to exit the assessment at the end of each segment, allowing the test to be administered in more than one testing session.

Minnesota Test of Academic Skills

Any district employee who has received MTAS test administration training may administer the MTAS. However, the Test Administrator should be a person who is familiar with the student's response mode and with whom the student is comfortable. All MTAS Test Administrators must be trained or review training materials prior to each test administration. Training can be completed by attending an in-person district-provided training or viewing training materials online. The Science MTAS is administered to students in a one-on-one setting. Therefore, Test Administrators must schedule times to administer the tasks.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

The ACCESS for ELLs includes English language proficiency tests in reading, writing, listening and speaking. Each of these domains is assessed individually in grades 1–12, and the four tests may be administered over multiple days. The grades 1–12 reading, writing, and listening tests are taken in groups, while the speaking test is a one-on-one conversation in which Test Administrators follow a script to elicit speech from students. The Kindergarten ACCESS for ELLs is administered entirely one-on-one, and all four domains are intended to be assessed in one 45-minute session. For ELs in grades 1–12 who have significant cognitive disabilities, the Alternate ACCESS for ELLs may be selected by IEP teams as the more appropriate assessment of students' developing language proficiency. The Alternate ACCESS for ELLs includes reading, writing, listening and speaking tests, which are entirely scripted and administered one-on-one.

Test Security

The recovery of testing materials after each administration is critical. Test booklets must be returned in order to preserve the security and confidential integrity of items that will be used on future tests.

Minnesota's testing contractor assigns secure test booklets to school districts by unique eight-digit barcoded security numbers. School districts complete answer document packing lists to assist Minnesota's testing contractor in determining whether student answer documents are missing. Minnesota's testing contractor compares barcode scan files of returned test booklets with test booklet distribution files to determine whether all secure materials have been returned from each school and district. School districts are responsible for ensuring the confidentiality of all testing materials and their secure return. Minnesota's testing contractor contacts any district with unreturned test booklets.

The Minnesota Department of Education's (MDE's) internal security procedures are documented in the *Policy and Procedures* appendix of the *Procedures Manual for the Minnesota Assessments.*

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series III

The grades 3–8 and 11 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) are delivered online, with paper or Braille forms available for students requiring an accommodated form. For the computer-delivered assessments, there are no secure materials to return. For students taking accommodated forms, which are paper-based, secure materials include large-print (18- and 24-point) test books, braille test books and scripts and CDs. Districts enter student responses into the corresponding Data Entry form in TestNav. All used and unused test books and accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Test of Academic Skills

Secure test materials for the Mathematics Minnesota Test of Academic Skills (MTAS) include the Task Administration Manuals, Presentation Pages, and Response Option Cards shipped to the district. Following administration, all used and unused Task Administration Manuals and Presentation Pages must be returned to Minnesota's testing contractor. All Response Option Cards must be securely destroyed at the district.

Reading

Minnesota Comprehensive Assessments-Series III

The grades 3–8 Reading Minnesota Comprehensive Assessments-Series III (MCA-III) are delivered online, with paper or Braille forms available for students requiring an accommodated form. For the computer-delivered assessments, there are no secure materials to return. For students taking accommodated forms, which are paper-based, secure materials include large-print (18- and 24-point) test books and answer books and braille test books. Districts enter student responses into the corresponding Data Entry form in TestNav. All used and unused test books and accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Test of Academic Skills

Secure test materials for the Reading MTAS include the Task Administration Manuals, Presentation Pages, and Response Option Cards shipped to the district. Following administration, all used and unused Task Administration Manuals and Presentation Pages must be returned to Minnesota's testing contractor. All Response Option Cards must be securely destroyed at the district.

Science

Minnesota Comprehensive Assessments-Series III

Since the Science MCA-III is a computer-delivered assessment, the only secure test materials for the Science MCA-III are accommodated materials, including large-print (18- and 24-point) test books and answer books, braille test books and scripts and CDs. All used and unused accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Test of Academic Skills

Secure test materials for the Science MTAS include the Task Administration Manuals, Presentation Pages, and Response Option Cards shipped to the district. Following administration, all used and unused Task Administration Manuals and Presentation Pages must be returned to Minnesota's testing contractor. All Response Option Cards must be securely destroyed at the district.

Title III Assessments

ACCESS for ELLs and Alternate ACCESS for ELLs

Secure test materials for the English language proficiency assessments include:

- Student test books
- Test Administrator's script
- Large-print materials, if ordered
- Test administration manuals

In addition to the above materials, the ACCESS for ELLs shipment includes:

- Listening test stimuli on CD
- Speaking flip charts
- Kindergarten ancillary kit
- Braille forms of the reading and writing tests, if ordered

Districts return **all** materials—used and unused—to MetriTech, Inc.

Accommodations

Some students who have disabilities or are English Learners (ELs) require special testing accommodations in order to fully demonstrate their knowledge and skills. Such accommodations allow these students to be assessed in the testing program without being disadvantaged by a disability or lack of English language experience. The available accommodations for each group of students are documented in chapters 5 and 6 of the *Procedures Manual for the Minnesota Assessments,* which is updated annually and available upon request from MDE.

Accommodation Eligibility

Students with Individualized Education Programs (IEPs), 504 plans, or EL status are eligible for testing accommodations. Districts are responsible for ensuring that accommodations do not compromise test security, difficulty, reliability, or validity and are consistent with a student's IEP or 504 plan. If the student has limited English proficiency, then accommodations or interpretations of directions may be provided. The decision to use a particular accommodation with a student should be made on an individual basis. This decision should take into consideration the needs of the student as well as whether the student routinely receives the accommodation during classroom instruction.

Typically, accommodations allow for a change in one or more of the following areas:

- Presentation
- Timing/scheduling

- Response

Not every accommodation is appropriate or permitted for every subject area.

For the Minnesota Test of Academic Skills (MTAS), any accommodation listed on a student's IEP may be used so long as it does not invalidate the test. Some administration activities that are allowed for the MTAS include:

- Familiarizing the student with the format of the MTAS prior to administration using the item samplers found on the MDE website
- Adapting the materials presented to meet student need, which includes enlarging materials or incorporating texture
- Using manipulatives unless otherwise specified in the task script
- Reading passages aloud to the student
- Using assistive technology devices, including calculators
- Refocusing and repeating as needed

Available Accommodations and Rationales

Presentation

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, multisensory, tactile and visual.

Assistive Technology

Description:

Assistive Technology refers to technology that is used to maintain, increase or improve the functional capabilities of students with disabilities who take online assessments.

Rationale:

According to MacArthur and Cavalier (1999):

> The results demonstrate that dictation helped students with LD produce better essays than they could produce by handwriting. The best essays were produced when dictating to a scribe. Essays composed by students with LD by dictating to speech recognition software were not as good as when using a scribe but were better than their handwritten essays. The performance of students without LD was equivalent in all three conditions.

MacArthur and Cavalier (2004) found the following:

> Results demonstrate that both dictation conditions helped students with learning disabilities produce better essays. Students with learning disabilities produced higher quality essays when using a scribe, than when using speech recognition software. Both adapted conditions were better in quality than handwritten essays.

Allowable Assessments:

- Science MCA
- Mathematics MCA
- Limited use on Writing subtest of ACCESS for ELLs
- Assistive technologies used by students with significant cognitive disabilities are generally allowed on the Mathematics, Reading, and Science MTAS and the Alternate ACCESS for ELLs.

Braille Versions of Assessment

Description:

Braille versions are available to students who are blind or partially sighted and are competent in the Braille system as determined by the student’s IEP Team. Student responses may be recorded in one of the following ways:

- In the answer book by a proctor
- In the test book by the student
- With a typewriter or word processor by the student
- Dictated to a scribe by the student
- With a Braille writer, slate and stylus used by the student

A regular-print version of the Braille tests for paper tests is provided at the time of testing to Test Monitors working with students. Test Monitors will need to view a computer screen for online tests.

Rationale:

As found by Wetzel and Knowlton (2000):

> Average print-reading rate ranged from 30% to 60% faster than the average Braille reading rate. Less than one-third of the Braille readers read slower than the print readers. Based on their performances in the different modes (for example, oral, silent, studying), it appears that Braille and print readers employ similar strategies for different tasks.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs (reading and writing tests only)

Handheld Calculator for Online Assessment

Description:

A handheld calculator is provided in an individual setting for items where the online calculator is available.

Allowable Assessments:

- Mathematics MCA (if indicated in IEP or 504 plan; must be used with paper form of test)

Note: For grade 11 Mathematics MCA, all students may use a calculator on the entire test, and handheld calculators may be used.

Large-Print Test Book

Description:

Large-print test books are for students with low vision who need a large-print test book to see the test items. Students write responses directly into the test book, and district staff transfer the answers into a Data Entry form in TestNav, which also includes the names and email addresses of school personnel involved. Large-print test materials may not be ordered for alternate assessments for students with significant cognitive disabilities, but Test Administrators may enlarge test materials to meet individual student needs.

Rationale:

Beattie, Grise, and Algozzine (1983) state:

> The results suggested that the competence of students with learning disabilities was enhanced by the use of tests which include the modifications such as large print.

As noted by Bennett, Rock, and Jirele (1987):

> With respect to performance level, the groups of students with visual impairments achieved mean scores that approximated or slightly exceeded those of students without disabilities. Students with physical disabilities scored lower on two of the three test scales. Students with physical disabilities and visual impairments taking timed, national administrations were slightly less likely to complete selected test sections than in the other conditions. The reliability of the General Test was found to be comparable to the reference population for all groups with students with disabilities.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- Mathematics MTAS
- Reading MTAS
- Science MTAS
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Made Tape

Description:

Tape recorders may be used by the student to record and edit answers if the student is unable to mark a scannable answer book. School testing personnel must transfer answers to a Data Entry form in TestNav.

Rationale:

According to Koretz (1997):

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- Reading subtest of ACCESS for ELLs

Mathematics Scripts Presented in English to Student via CD

Description:

Mathematics Scripts on CD may be provided to special education students as documented in their IEP or to English Learner (EL) students who need this accommodation.

Rationale:

A study by Helwig, Rozek-Tedesco, and Tindal (2002) found the following:

> The result suggest that Low reading students performed significantly better when test items were read aloud on only one of the two forms and in only one grade level. The accommodation did not seem to benefit High readers. No significant interaction was found between basic skill level and testing format.

According to Helwig, Rozek-Tedesco, Tindal, Heath, and Almond (1999):

> Students with low mathematical ability (regardless of reading ability) scored significantly higher under the video accommodation condition. There appeared to be little or no association between how many words, syllables, long words, or other language variables were present in a given test item and the difference in success rate on the standard or video version of the test. However, students with combined low reading fluency and above-average performance on the mathematics skills test experienced notable improvements when the selected items were read aloud.

Allowable Assessments:

- Mathematics MCA

Mathematics and Science Scripts Presented in English to Student via Online Audio (Text-to-Speech)

Description:

Mathematics or Science Scripts in audio may be provided to special education students as documented in their IEP or to English Learner (EL) students who need this accommodation.

Rationale:

A study by Helwig et al. (2002) found the following:

> The results suggest that Low reading students performed significantly better when test items were read aloud on only one of the two forms and in only one grade level. The accommodation did not seem to benefit High readers. No significant interaction was found between basic skill level and testing format.

According to Helwig et al. (1999):

> Students with low mathematical ability (regardless of reading ability) scored significantly higher under the video accommodation condition. There appeared to be little or no association between how many words, syllables, long words, or other language variables were present in a given test item and the difference in success rate on the standard or video version of the test. However, students with combined low reading fluency and above-average performance on the mathematics skills test experienced notable improvements when the selected items were read aloud.

Allowable Assessments:

- Mathematics MCA
- Science MCA

Mathematics and Science Scripts Presented to Student in Sign Language

Description:

Signed interpretation of the Mathematics or Science MCA scripts may be provided for deaf or hard-of-hearing students. The script along with the corresponding test book or accommodated form for online must be used during administration to maintain the validity of the test. Only the literal interpretation is acceptable.

Rationale:

According to a study by Johnson, Kimball, and Brown (2001):

> The results from the study suggest that the use of sign language as an accommodation presents political, practical, and psychometric challenges. The data showed that sign language translation can result in the omission of information required to answer a test item correctly.

Note: MDE continues to evaluate the efficacy of this accommodation for future administrations.

Allowable Assessments:

- Mathematics MCA
- Science MCA

Mathematics and Science Scripts Read in English to Student

Description:

Mathematics or Science MCA scripts may be read to special education students as documented in their IEPs or to EL students who need this accommodation.

Rationale:

As found by Huynh, Meyer, and Gallant (2004):

> It was found that the test structure remained rather stable across the three groups. Controlling for student background variables, disabled students under oral administration performed better than disabled students on the non-accommodated format. On the non-accommodated format, students with disabilities fared worse than general education students.

Allowable Assessments:

- Mathematics MCA
- Science MCA

Noise Buffer

Description:

Noise buffers may include individual study carrels, headsets, earplugs, individual portable buffers set on the student’s desk or an audio player that generates white noise or instrumental music. Audio players must be school-owned and the audio must be provided by the school. The noise buffer can be accessed through headphones or in an individual setting.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Templates to Reduce Visual Print Field, Magnification, and Low-Vision Aids

Description:

Templates to reduce the visual print field may be used by students competent in their use. Templates are not available from the state. Magnification or low-vision aids may be used as documented in an IEP or 504 Plan. Examples of low-vision aids are magnifying glasses, electronic magnifiers, cardboard cutouts and colored paper.

Rationale:

As noted by Robinson and Conway (1990):

> Subjects demonstrated significant improvements in reading comprehension and reading accuracy, but not in rate of reading, when assessed using the Neale Analysis of Reading Ability at 3-, 6-, and 12-month intervals after lens fitting. Students demonstrated a significant improvement in attitude to school and to basic academic skills.

Zentall, Grskovic, Javorsky, and Hall (2000) state:

> Students with attention deficits read as accurately as other students when color was added, read worse in the standard (black-and-white) condition, and improved reading accuracy during the second test administration with color added.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Translated Directions (Oral, Written, or Signed) into Student's First Language

Description:

Directions translated (oral, written, or American Sign Language [ASL]) into the student's first language.

Rationale:

As noted by Ray (1982):

> Deaf students taking the adapted version of the test scored similarly to students without hearing impairments on the WISC-R performance scale overall. The author suggests that when factors related to test administration are controlled (that is the child's comprehension of the task), deaf children score on the average the same as the normal population.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Voice Feedback Device (Whisper Phone)

Description:

Voice feedback devices or whisper phones are allowed for students with an IEP or 504 plan. These devices allow students to vocalize as they read and work problems. The use of whisper phones must not be audible to other students.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Word-to-Word Dual-Language Dictionary

Description:

A word-to-word dual-language dictionary contains mathematical and scientific terms in English and in the first language of a given learner. In a word-to-word dictionary, no definitions are provided, only direct translations of the mathematical and scientific words.

Rationale:

Idstein (2003) found the following:

> Qualitative results show the better students do well in less time than it takes weaker students to achieve lower grades. Weaker students rely excessively on their dictionaries and do not trust themselves. Dictionary use does not affect the scores or test time of the better students, and may actually slow down and negatively affect the scores of weaker students.

Allowable Assessments:

- Mathematics MCA
- Science MCA

Timing and Scheduling

Timing and scheduling accommodations increase the allowable length of time to complete an assessment or assignment and perhaps change the way the time is organized. While extended time or frequent breaks may be specified as accommodations in a student's IEP or 504 plan, they are considered an accommodation only for a student taking the ACCESS for ELLs, which is timed, although with some flexibility allowed. For all other Minnesota assessments, extended time and frequent breaks are considered a general practice and are available to all students.

Extended Testing Time (Same Day)

Description:

Extended testing time (same day) for the ACCESS for ELLs is available to EL students who have an IEP. Other EL students must finish the segment(s) on the day scheduled. Extended time on the Alternate ACCESS for ELLs is allowed and is not considered an accommodation.

Rationale:

According to Antalek (2005):

> While the majority of subjects used additional time to complete the writing task, no relationships were found between demographic factors such as gender, age, school type and grade and the completion of the task within the allotted time. Also, all subjects produced tests faster when given extended time. Subjects may feel compelled "to wrap it up," spent more time planning, or gained momentum during the task. Additional time contributed to improved performance. A significant relationship was noted between the quality of sentence structure and extended time testing conditions.

Allowable Assessment:

- ACCESS for ELLs

Extended Testing Time (Multiple Days)

Description:

Extended testing is considered an accommodation for assessments when testing is extended over multiple days. However, extended testing is not considered an accommodation for online assessments with pausing capability or for MTAS and Alternate ACCESS for ELLs.

Rationale:

According to Antalek (2005):

> While the majority of subjects used additional time to complete the writing task, no relationships were found between demographic factors such as gender, age, school type and grade and the completion of the task within the allotted time. Also, all subjects produced tests faster when given extended time. Subjects may feel compelled "to wrap it up," spent more time planning, or gained momentum during the task. Additional time contributed to improved performance. A significant relationship was noted between the quality of sentence structure and extended time testing conditions.

Allowable Assessment:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Response

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.

Answer Orally or Point to Answer

Description:

Students dictate their answers to a scribe or point to their answer in the test book.

Rationale:

A study done by Koretz (1997) found the following:

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Assistive Technology

Description:

Assistive technology refers to technology that is used to maintain, increase or improve the functional response capabilities of students with disabilities. The use of augmentative communication devices is not considered an accommodation on MTAS and Alternate ACCESS for ELLs, which are assessments for students with significant cognitive disabilities.

Rationale:

MacArthur and Cavalier (1999) note:

> The results indicate that two-thirds (68%) of the students achieved 85% accuracy and more than one-third (40%) achieved 90% accuracy using dictation to a scribe or speech recognition software. Only 3 students (10%) were below 80% accuracy. Results for adults have been reported between 90% and 98%.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Braille Writers

Description:

Braille note-taking devices may be used by students competent in their use as determined by their IEP or 504 team. School testing personnel must transfer answers to a Data Entry form in TestNav.

Rationale:

As Wetzel and Knowlton (2000) state:

> Average print-reading rate ranged from 30% to 60% faster than the average Braille reading rate. Less than one third of the Braille readers read slower than the print readers. Based on their performances in the different modes (for example, oral, silent, studying), it appears that Braille and print readers employ similar strategies for different tasks.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA

Large-Print Answer Book

Description:

Large-print answer books may be provided for students who need more space to accommodate their large handwriting when completing constructed-response items. Large-print test materials may not be ordered for alternate assessments for students with significant cognitive disabilities, but Test Administrators may enlarge test materials to meet individual student needs.

Rationale:

A study done by Beattie et al. (1983) found the following:

> The results suggested that the competence of students with learning disabilities was enhanced by the use of tests which include the modifications such as large print.

As suggested by Bennett et al. (1987):

> With respect to performance level, the groups of students with visual impairments achieved mean scores that approximated or slightly exceeded those of students without disabilities. Students with physical disabilities scored lower on two of the three test scales. Students with physical disabilities and visual impairments taking timed, national administrations were slightly less likely to complete selected test sections than in the other conditions. The reliability of the General Test was found to be comparable to the reference population for all groups with students with disabilities.

Allowable Assessments:

- Mathematics MCA
- Reading MCA

- Science MCA
- ACCESS for ELLs

Scratch Paper or Graph Paper (Always Allowed for Online Assessments)

Description:

For most tests, scratch paper is only available for students with IEP or 504 Plans. The exceptions are the online assessments, for which all students may use scratch paper, and alternate assessments for students with significant cognitive disabilities. Other students use the margins and other white space in the test book, but grade 3 students should be very careful not to write over the bubble areas in the MCA or ACCESS for ELLs.

Rationale:

As Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) note:

> General education students performed significantly higher than special education students in reading and in math. For both tests, performance was not higher when students were allowed to mark the booklet directly than when they had to use a separate bubble sheet.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Scribes

Description:

Scribes may be provided to students in those rare instances when visual or motor difficulties, including injuries, prevent them from writing their answers. The student's IEP must document the need for a scribe except in injury situations. The students should be competent in the use of scribes as determined by the student's IEP Team. Scribes must be impartial and experienced in transcription. Students must be given time, if desired, to edit their document. Students do not need to spell out words or provide punctuation.

Rationale:

Koretz (1997) states the following:

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to

> have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs
- Alternate ACCESS for ELLs

Voice-Activated Computer

Description:

Voice-activated computers may be used by students who are competent in their use as determined by student’s IEP Team. The student must be given the time needed to edit the documents.

Rationale:

As noted by MacArthur and Cavalier (1999):

> The results demonstrate that dictation helped students with LD produce better essays than they could produce by handwriting. The best essays were produced when dictating to a scribe. Essays composed by students with LD by dictating to speech recognition software were not as good as when using a scribe but were better than their handwritten essays. The performance of students without LD was equivalent in all three conditions.

A study by MacArthur and Cavalier (2004) found the following:

> Results demonstrate that both dictation conditions helped students with learning disabilities produce better essays. Students with learning disabilities produced higher quality essays when using a scribe than when using speech recognition software. Both adapted conditions were better in quality than handwritten essays.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Word Processor or Similar Assistive Device

Description:

Word processors, computers, or similar computerized devices may be used if the IEP or 504 Team determines that a student needs it. For example, a student may use a portable note taker, such as an AlphaSmart or a related program (such as a spellchecker or word prediction software or device) commonly used in a student’s academic setting, if it is included in the IEP and the student has demonstrated competency in its use. Assistive devices are allowed on assessments for students with significant cognitive disabilities and are not considered accommodations.

Rationale:

According to Hollenbeck, Tindal, Harniss, and Almond (1999):

> Differences between handwritten students' essays and computer-generated essays were non-significant. Significant differences were found between ratings for essays of computer-last day group and computer last day with spell-check group. Students with disabilities performed significantly poorer when composing with a computer than when handwriting their stories.

Hollenbeck, Tindal, Stieber, and Harniss (1999) found that:

> Analysis showed that the original handwritten compositions were rated significantly higher than the typed composition on three of the six writing traits for the total group. Further, five of the six mean trait scores favored the handwritten essays.

Note: MDE continues to evaluate the efficacy of this accommodation for future administrations.

Allowable Assessments:

- Mathematics MCA
- Reading MCA
- Science MCA
- ACCESS for ELLs

Other Accommodations Not Listed

If an IEP or 504 team desires to use an accommodation not on the approved list, it may contact MDE for consideration of that accommodation for the current administration and in future administrations pending literature and research reviews.

Accommodations Use Monitoring

Minnesota uses a data audit system—as well as selected field audits—to monitor the use of accommodations on its assessments. At a state level, data is reviewed for all accommodations for those students who are (1) receiving special education or identified as disabled under Section 504 of the Rehabilitation Act of 1973 and (2) ELs.

Data Audit

The data collection is intended to provide MDE with the information about districts' use of accommodations on state assessments. This information will allow MDE to analyze the accommodation data to draw conclusions about the use and overuse of accommodations and will inform future policy decisions and training needs regarding the use of accommodations.

The Yearbook provides an annual review of percentages of accommodations used against the number of assessments scored without accommodations. MDE continually reviews these numbers both in overall percentage and in percentage expected in specific disability categories based on past data.

Field Audit

MDE annually conducts monitoring visits through its Division of Compliance and Assistance to review the use of accommodations on state assessments. During the course of these visits, IEPs are reviewed for a variety of state and federal requirements and statutes. For the state assessments, IEPs are reviewed so that MDE can

- verify that accommodations used on state assessments are documented in the IEP; and
- monitor the provisions of accommodations used during testing.

The field audit reviews the IEP to ensure that any accommodations used during state or district testing are appropriately documented in the student’s IEP as well as the rationale for the accommodation.

Chapter 4: Reports

After each test administration, a number of reports are provided. The reports include individual student paper reports and labels, online reports, and an electronic District Student Results (DSR) file containing individual student records with demographics and multiple scores used to prepare all other reports. Summary reports are also created that provide test results aggregated at school, district, or state levels. The reports focus on three types of scores: scale scores, raw scores, and achievement levels. This chapter provides an overview of the types of scores reported and a brief description of each type of report. Also provided in this chapter are guidelines for proper use of scores and cautions about misuse.

As with any large-scale assessment, the Minnesota Assessments provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Because all tests measure a finite set of skills with a limited set of item types, placement decisions and decisions concerning student promotion or retention should be based on multiple sources of information, including but not limited to test scores.

Information about student performance is provided on individual student reports and summary reports for schools, districts and the state. This information may be used in a variety of ways. Interpretation guidelines were developed and published as a component of the release of public data; this document, the *Interpretive Guide,* is available on the Manuals tab of the PearsonAccess Resources page.

Description of Scores

Scores are the end product of the testing process. They provide information about how each student performed on the tests. Three different types of scores are used on the Minnesota Assessment reports: scale scores, raw scores and achievement levels. These three scores are related to each other. The following section briefly describes each type of score.

Raw Score

The raw score is the sum of points earned across items on a subject-area test. In addition to total raw scores, raw scores for items that constitute a specific strand or substrand may be reported. By themselves, these raw scores have limited utility. They can be interpreted only in reference to the total number of items on a subject-area test or within a stand or substrand. They cannot be compared across tests or administrations. Several values derived from raw scores are included to assist in interpreting the raw scores: maximum points possible and aggregate averages (for school-, district-, and state-level reports). Note that for the Mathematics and Reading Minnesota Comprehensive Assessments-Series III (MCA-III), total and strand scores are computed using measurement model–based pattern scoring (i.e., scores depend on the pattern of correct/incorrect responses for the particular items taken by the student). Thus, the sum of points earned is not used to determine scale scores. Therefore, raw number-correct scores are not reported for Mathematics and Reading MCA-III.

Scale Score

Scale scores are statistical conversions of raw scores or model-based scores that maintain a consistent metric across test forms and permit comparison of scores across all test administrations within a particular grade and subject. Because scale scores adjust for different form difficulties, they can be used to determine whether a student met the standard or achievement level in a manner that is fair across forms and administrations. Schools can also use scale scores to compare the knowledge and skills of groups of students within a grade and subject across years. These comparisons can be used in assessing the impact of changes or differences in instruction or curriculum.

The scale scores for a given Minnesota Comprehensive Assessments-Series III (MCA-III) subject and grade range from *G*01 to *G*99, where *G* is the grade tested. For each MCA assessment, scale score *G*50 is the cut score for *Meets the Standards*, and *G*40 is the cut score for *Partially Meets the Standards*. The MCA cut score for *Exceeds the Standards* can vary by grade and subject. The scale score metric for each grade and subject is determined independently of that for other grades and subjects; comparisons should not be made across grades or subjects. In the case of the Science MCA-III exams, scale scores are transformations of raw number-correct scores. More than one raw score point may be assigned the same scale score, except at cut scores for each achievement level or at the maximum possible scale score. Pattern scoring is used to determine scale scores for Mathematics and Reading MCA-III.

The range of observed scale scores for the Minnesota Test of Academic Skills (MTAS) varies somewhat from year to year. The scale metric for the MTAS was originally set in 2007, with the cut scores for *Partially Meets* and *Meets* set at 195 and 200, respectively, for each grade and subject. In 2008, additional items were added to the MTAS to increase its reliability, and the scale metric and cut scores were adjusted. On the 2008 MTAS metric, 190 became the cut score for *Partially Meets* while 200 remained the scale cut for *Meets*. Because of the adjustments to the scale metric, 2007 MTAS scores cannot be compared directly to MTAS scores from subsequent years. As with the MCA-III, MTAS scale scores from different grades and subjects are not directly comparable.

The meaning of scale scores is tied to the content and achievement levels associated with a given set of Minnesota Academic Standards. Thus, MCA-II scores are not directly comparable to MCA-III scores, because those scores reflect different content and achievement standards. Similarly, when MTAS assessments change the academic standards to which they aligned (concomitant with the MCA in the same grade and subject), the scores from assessments based on different academic standards are not directly comparable.

Details about how scale scores are computed are given in Chapter 6, “Scaling.”

Achievement Levels

To help parents and schools interpret scale scores, achievement levels are reported. Each achievement level is determined by the student’s scale or raw score. The range for an achievement level is set during the standard setting process. Each time a new test is implemented, panels of Minnesota educators set the achievement levels. For each test, certain achievement levels are designated as proficient. Table 4.1 provides a summary of the achievement levels for the Minnesota Assessment System.

Table 4.1. Achievement Levels for Minnesota Assessment System

Test	Subject	Name of Achievement Level	Proficient
MCA-III, MTAS	Mathematics, Reading, Science	Does Not Meet the Standards	No
		Partially Meets the Standards	No
		Meets the Standards	Yes
		Exceeds the Standards	Yes
ACCESS for ELLs	Listening, Speaking, Reading, Writing, Composite	Level 1: Entering	*
		Level 2: Emerging	*
		Level 3: Developing	*
		Level 4: Expanding	*
		Level 5: Bridging	*
		Level 6: Reaching	*
Alternate ACCESS for ELLs	Listening, Speaking, Reading, Writing, Composite	Level A1: Initiating	*
		Level A2: Exploring	**
		Level A3: Engaging	**
		Level P1: Entering	**
		Level P2: Emerging	**
		Level P3: Developing (writing test only)	**

* Proficiency requires composite score ≥ 5 and all subject scores ≥ 4.

** No proficiency cut scores have been set.

Description of Reports

Reports resulting from administrations of the Minnesota Assessments fall into two general categories. Student-level reports provide score information for individual students. Summary reports provide information about test performance aggregated across groups of students (e.g., students in a school). The available student reports are listed in Table 4.2. Sample student reports can be found in MDE's *Interpretive Guide.*

Secure online reports of student and summary data are available to authorized district personnel from Minnesota's test vendor and from MDE. The secure summary reports distributed to schools and districts are not for public release; all student data is reported. MDE also makes extensive summary data available to public users (subject to filtering when cell sizes fall below 10 students) through the MDE Data Center website at http://education.state.mn.us/MDE/Data/index.html.

Table 4.2. Student Test Reports

File or Report Name	Report Format	Applies to MCA-III, MTAS, and GRAD	Applies to ACCESS for ELLs, Alternate ACCESS for ELLs
Individual Student Report (Home Copy)	Paper	X	X
Individual Student Report (School Copy)	Online or PDF	X	
Student Labels	Paper	X	
District Student Results File (DSR)	Electronic	X	X

Student Reports

Minnesota Assessment student reports provide information on a student's overall performance in each subject measured as well as a comparison of his or her performance relative to other students in the school, district, and state. For many assessments, including the MCA-III, these reports provide scaled scores as well as achievement-level designations associated with the student's performance level. Subscores at the strand or substrand level are also reported for each student. The information presented in these reports can be used by parents to help them understand their child's achievement.

Individual Student Reports

The Individual Student Report (ISR) is a document sent home to provide individual student data for student, parent, and teacher use. An individual student's earned scale score is presented in a graphic representation along with the assigned achievement level. School, district, and/or state average scale scores are presented on the same graphic for comparison. The Minnesota Comprehensive Assessments-Series III (MCA-III) ISRs provide all three averages; the Minnesota Test of Academic Skills (MTAS) provide the state average. Decisions about which average scale scores are reported for a given test are driven by the number of students generally included in the average. For example, the number of students included in the MTAS Reading school-level average scale score is small for most schools; this results in large standard errors of the mean. MDE has a policy to filter information for public release if the number of students is less than 10. For the MTAS tests, the number of students is frequently quite small for school or district populations.

The ISR presents further information about student performance, including presentation of subscores. For the MCA-III, the provided subscore information includes student strand scale scores, along with scale score range and a tolerance band for each strand score representing score precision. For the MTAS, student raw scores, maximum possible scores and state mean raw scores are reported for each strand (mathematics) or substrand (reading). A proficiency indicator is provided for each subject tested along with the achievement-level descriptors for the earned achievement level.

For grades 3–8, an across-grade Progress Score is also provided for Mathematics and Reading MCA-III. The Progress Score is explained in more detail in Chapter 6, “Scaling.”

The ISRs are provided to the district in two formats: one paper copy for sending home to parents and one Adobe PDF document for school use. Authorized district personnel can also access the test vendor’s online reporting in PearsonAccess to retrieve an online version of a student’s ISR. When appropriate, these online reports permit immediate posting of a student’s results. This online report is considered preliminary, subject to verification by MDE. The paper and PDF ISRs provided to districts reflect official accountability results for students.

Student Label

The student label contains the test name, test date, student information, scale scores, and achievement level for each subject tested for a single test. The individual student labels have adhesive backing to permit their secure attachment to a student’s permanent paper file, should the district maintain one. The purpose of the student label is to provide a compact form of individual student information for recording in student files.

Summary Reports

Summary reports provide information to schools and districts that may be used for the purpose of evaluating programs, curriculum, and instruction of students. For example, districts may use the MCA-III school summary reports of test results by subject as one line of evidence to consider in evaluating how well their curriculum and instruction is aligned with the Minnesota Academic Standards. Summary reports are available online to authorized district personnel from the test vendor’s Online Reporting System, and from MDE’s Data Center. Public summary reports are also available from the Data Center.

Online Reporting System

Minnesota’s test vendor’s PearsonAccess online reporting provides performance data to authorized district personnel that is aggregated at the district, school, teacher, and roster levels, as well as for individual students. The Online Reporting System contains three major applications: On Demand Reports, Longitudinal Reports, and Published Reports.

1. **On Demand Reports** provide the student’s score in PearsonAccess within 60 minutes after testing is completed for MCA, GRAD, and OLPA. Information on benchmark strengths and weaknesses by reporting category, Learning Locators, and Lexile scores for reading are also available within the report. Authorized users can log in to PearsonAccess to view the student’s score and access printable student reports. These reports include results only from students who tested online and reflect the latest results.

2. **Longitudinal Reports** are available for authorized users to create reports using tools to disaggregate data over multiple years by school and student groups for MCA, MTAS, and OLPA. The longitudinal system allows users to disaggregate data by subject, grade, and specific demographics. Reports are generated for score and performance level using aggregation criteria

selected by user. **Comparison of Achievement** reports are also available to compare state/district/school strands and achievement levels from year to year.

3. **Published Reports** provide users access to Benchmark Reports, which compare school- or district-level aggregate performance on items from each benchmark with that expected given overall student scores. PDF copies of the Individual Student Reports (ISRs) and district and school rosters are available after final test results are available to distribute.

PearsonAccess Online Reporting provides dynamic data that can be used to gauge students' achievement on the Minnesota assessments. However, the data and reports in this system are not to be used for official accountability purposes. The Minnesota Department of Education provides official accountability data.

MDE Data Center

A wide variety of secure online reports summarizing test results at the school, district, and state level are used to provide information to authorized school and district educators and administrators. The data are reported for all students tested. For example, a disaggregated report showing average scale scores and the percentage of students proficient at each achievement level by the subgroups used for No Child Left Behind (NCLB) provides a different perspective on the school or district performance. This allows district staff to use the reports to estimate their index points for NCLB Adequate Yearly Progress (AYP) calculations. Downloadable data files containing individual student score records (DSR/SSR files) are also available to authorized district personnel.

Although individual student scores are confidential by law, reports of group (aggregated) scores are considered public information and are available for general use from the MDE Data Center (http://education.state.mn.us/MDE/Data/index.html). These public data include interactive reports that users can query to summarize data at the school, district, or statewide level with customizable demographic breakdowns, as well as downloadable summary data files. Student confidentiality on public documents is maintained by filtering; if any specific group (for example, English Learners) consists of fewer than 10 students, mean scores and the percentage of students who are proficient are not included in reports or data files posted to the MDE website.

Appropriate Score Uses

The tests in the Minnesota Assessment System are designed primarily to determine school and district accountability related to the implementation of the Minnesota standards. They are summative measures of a student's performance in a subject at one point in time. They provide a snapshot of the student's overall achievement, not a detailed accounting of the student's understanding of specific content areas defined by the standards. Test scores from Minnesota assessments, when used appropriately, can provide a basis for making valid inferences about student performance. The following list outlines some of the ways the student scores can be used.

- *Reporting results to parents of individual students*

  The information can help parents begin to understand their child's academic performance as related to the Minnesota standards.

- *Evaluating student scores for placement decisions*

The information can be used to suggest areas needing further evaluation of student performance. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the Minnesota standards. Students may also exhibit strengths or deficits in strands or substrands measured on these tests. Because the strand and substrand scores are based on small numbers of items, the scores must be used in conjunction with other performance indicators to assist schools in making placement decisions, such as whether a student should take an improvement course or be placed in a gifted or talented program.

- *Evaluating programs, resources and staffing patterns*

  Test scores can be a valuable tool for evaluating programs. For example, a school may use its scores to help evaluate the strengths and weaknesses of a particular academic program or curriculum in their school or district as it relates to the Minnesota standards.

Individual Students

Scale scores determine whether a student's performance has met or fallen short of the proficiency criterion level. Test results can also be used to compare the performance of an individual student with the performance of a similar demographic group or to an entire school, district, or state group. For example, the score for a Hispanic student in a gifted program could be compared with the average scores of Hispanic students, gifted students, all the students on campus, or any combination of these aggregations.

Subscores provide information about student performance in more narrowly defined academic content areas. For example, individual scores on strands and/or substrands can provide information to help identify areas in which a student may be having difficulty, as indicated by a particular test. Once an area of possible weakness has been identified, supplementary data should be collected to further define the student's instructional needs.

Finally, individual student test scores must be used in conjunction with other performance indicators to assist in making placement decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.

Groups of Students

Test results can be used to evaluate the performance of student groups. The data should be viewed from different perspectives and compared to district and state data to gain a more comprehensive understanding of group performance. For example, the average scale score of a group of students may show they are above the district and/or state average, yet the percentage of students who are proficient in the same group of students may be less than the district or state percentages. One perspective is never sufficient.

Test results can also be used to evaluate the performance of student groups over time. Average scale scores can be compared across test administrations within the same grade and subject area to provide insight into whether student performance is improving across years. For example, the average scale score for students taking the Grade 8 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) in 2014 can be compared to any of the 2011–2013 grade 8 MCA-III administrations. However, whenever drawing inferences from such comparison, it is important to account for how changes in the testing program over the years may have influenced the testing population taking a specific test. For example, 2011 saw the introduction of the Minnesota Comprehensive Assessments-Modified (MCA-Modified) exams, which meant that some students who otherwise would have taken the MCA now could take MCA-Modified instead. Most of those students returned to the MCA-III in 2015, when in MCA-Modified was discontinued. Consequently, in making comparisons with past administrations it is important to consider that the population taking the MCA-III has changed in 2015.

Consideration must also be given to changes in test administration policies when interpreting year-to-year changes in test scores. For Mathematics MCA-III, students taking the test in online mode in 2012 were allowed up to three administrations of the assessment and could use the highest score for accountability purposes. By contrast, in 2011, 2013 and subsequent years, students were allowed only a single Mathematics MCA-III testing opportunity.

In making longitudinal comparisons, it is important to recognize that new testing programs cannot be compared to previous testing programs that assessed different academic standards. For example, results from the Minnesota Comprehensive Assessments-Series III (MCA-III) cannot be directly compared to previous administrations of the MCA-II, because the MCA-III assesses different academic standards than its predecessor. The same holds true for grades 3–8 Mathematics MTAS, which assesses new standards beginning in 2011 and cannot be directly compared to the grades 3–8 Mathematics MTAS from prior years, and for the Science MTAS, which was constructed to align to new academic standards in 2012. In 2013, all Reading assessments (MCA-III, MCA-Modified and MTAS) were revised to align to the 2010 Minnesota ELA Academic Standards, and thus results on them are not directly comparable to those of prior years' versions of those tests. In 2014, grade 11 Mathematic assessments (MCA-III, MCA-Modified and MTAS) were revised to align to the 2007 Minnesota Mathematics Academic Standards.

The percentages of students in each achievement level can also be compared across administrations within the same grade, subject area and test to provide insight into whether student performance is improving across years. For example, the percentage of students in each achievement level for the grade 8 Mathematics MCA-III in 2015 can be compared to any of the 2011–2014 populations, while keeping in mind changes to the testing program such as those noted above. Schools would expect the percentage of students to decrease in the *Does Not Meet the Standards* achievement level, while the percentages in the *Meets the Standards* and *Exceeds the Standards* achievement levels would be expected to increase; this will show the school or district is moving toward the NCLB goal of having 100% of students proficient by 2014. However, the caveats expressed in the previous paragraphs concerning testing program changes would also apply to achievement level comparisons across years, particularly because testing program changes in content alignment are accompanied by changes in the definition of achievement levels.

Test scores can also be used to compare the performance of different demographic or program groups (within the same subject and grade) on a single administration to determine which demographic or program group, for example, had the highest or lowest average performance, or the highest percentage of students considered proficient on the Minnesota standards. Other test scores can be used to help evaluate academic areas of relative strength or weakness. Average performance on a strand or substrand can help identify areas where further diagnosis may be warranted for a group of students.

Test results for groups of students may also be used when evaluating instructional programs; year-to-year comparisons of average scores or the percentage of students considered proficient in the program will provide useful information. Considering test results by subject area and by strand or substrand may be helpful when evaluating curriculum, instruction, and their alignment to standards because all the Minnesota assessments are designed to measure content areas within the required state standards.

Generalizations from test results may be made to the specific content domain represented by the strands or substrands being measured on the test. However, because the tests are measuring a finite set of skills with a limited set of items, any generalizations about student achievement derived solely from a particular test should be made cautiously and with full reference to the fact that the conclusions were based on only one test. All instruction and program evaluations should include as much information as possible to provide a more complete picture of performance.

Cautions for Score Use

Test results can be interpreted in many different ways and used to answer many different questions about a student, educational program, school, or district. As these interpretations are made, there are always cautions to consider.

Understanding Measurement Error

When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. That is to say, test scores are not infallible measures of student characteristics. Rather, some score variation would be expected if the same student tested across occasions using equivalent forms of the test. This effect is due partly to day-to-day fluctuations in a person's mood or energy level that can affect performance and partly a consequence of the specific items contained on a particular test form the student takes. Although all testing programs in Minnesota conduct a careful equating process (described in Chapter 7) to ensure that test scores from different forms can be compared, at an individual level one form may result in a higher score for a particular student than another form. Because measurement error tends to behave in a fairly random fashion, when aggregating over students these errors in the measurement of students tend to cancel out. Chapter 8, "Reliability," describes measures that provide evidence indicating that measurement error on Minnesota assessments is within a tolerable range. Nevertheless, measurement error must always be considered when making score interpretations.

Using Scores at Extreme Ends of the Distribution

As with any test, student scores at the extremes of the score range must be viewed cautiously. For instance, if the maximum raw score for the grade 5 Science MCA-III is 41 and a student achieves this score, it cannot be determined whether the student would have achieved a higher score if a higher score were possible. In other words, if the test had 10 more items on it, it is difficult to know how many of those items the student would have correctly answered. This is known as a "ceiling effect." Conversely, a "floor effect" can occur when there are not enough items to measure the low range of ability. Thus, caution should be exercised when comparing students who score at the extreme ends of the distribution.

Another reason for caution in interpreting student scores at extreme ends of the distribution is the phenomenon known as regression toward the mean. Students who scored high on the test may achieve a lower score the next time they test because of regression toward the mean. (The magnitude of this regression effect is proportional to the distance of the student's score from the mean and bears an inverse relationship to reliability.) For example, if a student who scored 38 out of 40 on a test were to take the same test again, there would be 38 opportunities for him or her to incorrectly answer an item he or she answered correctly the first time, while there would only be two opportunities to correctly answer items missed the first time. If an item is answered differently, it is more likely to decrease the student's score than to increase it. The converse of this is also true for students with very low scores; the next time they test, they are more likely to achieve a higher score, and this higher score may be a result of regression toward the mean rather than an actual gain in achievement. It is more difficult for students with very high or very low scores to maintain their score than it is for students in the middle of the distribution.

Interpreting Score Means

The scale score mean (or average) is computed by summing each student's scale score and dividing by the total number of students. Although the mean provides a convenient and compact representation of where the center of a set of scores lies, it is not a complete representation of the observed score distribution. For example, very different scale score distributions in two groups could yield similar mean scale scores. When a group's scale score mean falls above the scale score designated as the passing or proficient cut score, it does not necessarily follow that most students received scale scores higher than the cut score. It can be the case that a majority of students received scores lower than the cut score while a small number of students got very high scores. Only when more than half of the students score at or above the particular scale cut score can one conclude that most students pass or are proficient on the test. Therefore, both the scale score mean and percentage at or above a particular scale cut score should be examined when comparing results from one administration to another.

Using Objective/Strand-Level Information

Strand or substrand level information can be useful as a preliminary survey to help identify skill areas in which further diagnosis is warranted. The standard error of measurement associated with these generally brief scales makes drawing inferences from them at the individual level very suspect; more confidence in inferences is gained when analyzing group averages. When considering data at the strand or substrand level, the error of measurement increases because the number of possible items is small. In order to provide comprehensive diagnostic data for each strand or substrand, the tests would have to be prohibitively lengthened. Once an area of possible weakness has been identified, supplementary data should be gathered to understand strengths and deficits.

In addition, because the tests are equated only at the total subject-area test scale score level, year-to-year comparisons of strand- and/or substrand-level performance should be made cautiously. Significant effort is made to approximate the overall difficulty of the strands or substrands from year to year during the test construction process, but fluctuations in difficulty do occur across administrations. Observing trends in strand and/or substrand level performance over time, identifying patterns of performance in clusters of benchmarks testing similar skills and comparing school or district performance to district or state performance are more appropriate uses of group strand/substrand information.

Furthermore, for tests under development with new content standards, changes to the test content and the percentage of score points allotted to each standard, strand, substrand, and/or benchmark may change. Some of these changes may be significant. When changes in test content occur, comparing student performance across years is particularly difficult, and under these circumstances the advice from measurement professionals is likely to discourage making such comparisons.

Program Evaluation Implications

Test scores can be a valuable tool for evaluating programs, but any achievement test can give only one part of the picture. As addressed in Standard 13.9 in the *Standards for Educational and Psychological Testing*, "In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation." The Minnesota statewide tests are not all-encompassing assessments measuring every factor that contributes to the success or failure of a program. Although more accurate evaluation decisions can be made by considering all the data the test provides, users should consider test scores to be only one component of a comprehensive evaluation system.

Chapter 5: Performance Standards

Performance standards are provided to assist in the interpretation of test scores. Anytime changes in test content take place, development of new performance standards may be required. The discussion below provides an introduction to the procedures used to establish performance standards for Minnesota assessments.

Introduction

Test scores in and of themselves do not imply student competence. Rather, the interpretation of test scores permits inferences about student competence. In order to make valid interpretations, a process of evaluating expected and actual student performance on assessments must be completed. This process is typically referred to as standard setting (Jaeger, 1989). Standards are set to determine the level of performance students need to demonstrate to be classified into defined achievement levels. There are four levels of achievement for the Minnesota Comprehensive Assessments-Series III (MCA-III): *Does Not Meet the Standards, Partially Meets the Standards, Meets the Standards* and *Exceeds the Standards.* Student achievement on the Minnesota Test of Academic Skills (MTAS) is reported using the same names for the achievement levels as used for the MCA-III; however, for the MTAS, this performance is related to the Alternate Achievement Standards. The ACCESS for ELLs has six performance levels that range from 1 (Entering) to 6 (Reaching). The Alternate ACCESS for ELLs has five performance levels for the reading, listening and speaking tests and six performance levels for the writing test. The levels range from A1 (Initiating) to P2 (Emerging) or P3 (Developing).

Standard setting for the Reading MCA-III and MTAS aligned to 2010 Minnesota academic standards in English language arts (ELA) was conducted in June 2013. Standard setting for the Science MCA-III and MTAS was conducted in June 2012. Standard setting for grades 3–8 Mathematics MCA-III was conducted in June 2011. Standard setting for grades 3–8 Mathematics MTAS aligned to 2007 academic standards was conducted in June 2011. Standard setting for grade 11 Mathematics MCA and grade 11 MTAS aligned to 2007 academic standards was conducted in June 2014. An overview of the process for establishing the achievement levels for these tests is described in the following pages of this chapter. More detailed explanations of the standard setting activities can be found in the technical reports of these workshops, which may be found on the Minnesota Department of Education (MDE) website at: http://education.state.mn.us or obtained from MDE upon request.

Achievement Level Setting Activity Background

There are a variety of achievement-level setting methods, all of which require the judgment of education experts and possibly other stakeholders. These experts are often referred to as judges, participants or panelists (the term “panelist” will be used here). The key differences among the various achievement-level setting methods can be conceptualized in terms of exemplar dichotomies. The most cited dichotomy is *test-centered* versus *student-centered* (Jaeger, 1989). Test-centered methods focus panelists’ attention on the test or items in the test. Panelists make decisions about how important and/or difficult test content is and set cut scores based on those decisions. Student-centered methods focus panelists’ attention on the actual and expected performance of examinees or groups of examinees. Cut scores are set based on student exemplars of different levels of competency.

Another useful dichotomy is *compensatory* versus *conjunctive* (Hambleton & Plake, 1997). Compensatory methods allow examinees who perform less well on some content to “make up for it” by performing better on other important content. Conjunctive methods require that students perform at specified levels within each area of content. There are many advantages and disadvantages to methods in each of these dichotomies, and some methods do not fall neatly into any classification.

Many achievement-level setting methods perform best under specific conditions and with certain item types. For example, the popular Modified Angoff method is often favored with selected-response (SR) items (Cizek, 2001; Hambleton & Plake, 1997), whereas the policy-capturing method was designed specifically for complex performance assessments (Jaeger, 1995). Empirical research has repeatedly shown that different methods do not produce identical results; it is important to consider that many measurement experts no longer believe "true" cut scores exist (Zieky, 2001). Therefore, it is crucial that the method chosen meets the needs of the testing program and that subsequent achievement-level setting efforts follow similar procedures.

Descriptions of most standard setting methods detail how cut scores are produced from panelist input, but they often do not describe how the entire process is carried out. However, the defensibility of the resulting standards is determined by the description of the complete process, not just the "kernel" methodology (Reckase, 2001). There is no clear reason to choose one method or one set of procedures over others. Because of this fact, test developers often design the process and adapt a method to meet their specific needs.

Process Components

Selecting a Method

Different methodologies rely on different types of expertise for the facilitators and the panelists. A major consideration is the knowledge, skills and abilities (KSA) of prospective panelists. If the panel includes persons who are not familiar with instruction or the range of the student population, it may be wise to avoid methods requiring a keen understanding of what students can actually do. Selection of the method should include consideration of past efforts in the same testing program and the feasibility of carrying out the chosen method.

Selecting and Training Panelists

Panelists should be subject-matter experts, understand the examinee population, be able to estimate item difficulty, have knowledge of the instructional environment, have an appreciation of the consequences of the standards and be representative of all the stakeholder groups (Raymond & Reid, 2001). This is a demanding cluster of KSA, and it may be difficult to gather a panel where every member is completely qualified. It may be useful to aim for the panel as a whole to meet KSA qualifications, while allowing individual panelists to have a varied set of qualities. Training should include upgrading the KSA of panelists where needed, as well as method-specific instruction. Training should also imbue panelists with a deep, fundamental understanding of the purposes of the test, test specifications, item development specifications and standards used to develop the items and the test.

Carrying Out the Methodology

As stated earlier, the methods are often adapted to meet the specific needs of the program. The KSA of the panel should be considered in the adaptations.

Feedback

Certain methodologies explicitly present feedback to panelists. For example, some procedures provide examinee performance data to panelists for decision-making. Other types of feedback include consequential (impact data), rater location (panelist comparisons), process feedback and hybrid (Reckase, 2001). Experts do not agree on the amount or timing of feedback, but any feedback can have influence on the panelists' ratings. Reckase (2001) suggests that feedback be spread out over rounds in order to have impact on the panelists. Care should be taken not to use feedback to pressure panelists into decisions.

Standard Setting for Grade 11 Mathematics Minnesota Comprehensive Assessments-Series III and Mathematics Minnesota Test of Academic Skills

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was implemented for the Minnesota Comprehensive Assessments-Series III (MCA-III) standard setting held in Saint Paul, Minnesota, on June 18–19, 2014. Achievement-level cut scores were established for Mathematics MCA-III in grade 11.

The ID Matching procedure (Ferrara, Perie, & Johnson, 2002) was used to recommend performance standards for the Mathematics MTAS-III assessments held in Saint Paul, Minnesota on June 18–19, 2014. Although similar to the widely implemented Bookmark method, the ID Matching procedure asks panelists to indicate which of the achievement-level descriptors is best matched by the knowledge and skill requirements necessary to respond successfully to each test item. Achievement-level cut scores were established for Mathematics MTAS-III in grade 11.

The activities of the meeting are documented in a document titled *Minnesota Assessments Summer 2014 Standard Setting: Recommended Performance Standards for Series III Mathematics Assessments.* The report can be found under the heading "Standard Setting Technical Report: Grade 11 Mathematics (2014)" on the MDE website at:
http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html.

This section provides a summary of outcomes from the meeting. Minnesota's testing contractor, the Minnesota Department of Education (MDE) and MDE's National Technical Advisory Committee (TAC) worked together to design the standard setting activities so as to follow the same general procedures as the standard settingf meeting for science, mathematics and reading for Minnesota Comprehensive Assessments-Series III (MCA-III) and Minnesota Test of Academic Skills. Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE convened separate educator panels to recommend performance standards for the Series III Mathematics MCA and MTAS assessments. Each panel had its own facilitator and was physically separate from the other panels.

MDE invited approximately 12 participants from across Minnesota to set cut scores for each assessment. The details of the credentials and demographics of the participants can be found in the *Minnesota Assessments Summer 2014 Standard Setting: Recommended Performance Standards for Series III Mathematics Assessments.*

Table Leaders

During the standard setting, participants are divided into groups, often called “tables.” Each table had one table leader who had been previously selected by MDE. Table leaders were expected to keep track of the table-level discussion and represent their committee’s point of view during the vertical articulation meeting. Table leaders were trained about their roles and responsibilities on Day 1 of the standard setting.

Ordered Item Booklets

Central to both the Bookmarking and ID Matching procedures is the production of an Ordered Item Booklet (OIB). As noted previously, while the OIB is often produced from only those items in the first operational test, it is rarely the case that a single operational test administration provides a comprehensive sampling of items across the range of content standards and difficulty. While recommending standards on the entire item bank may in some respects be ideal, including too many items makes review of the OIB overly burdensome. For the Mathematics MCA-III OIB, operational items common to the 2014 online and paper test form administration modes served as the base. This OIB was augmented with 21 additional operational items selected from other operational forms. These additional items were selected to complement the content distribution of the selected operational form, in terms of standards and benchmarks assessed and the item types, and to fill item difficulty gaps in the OIB. This led to an OIB that included 77 items for Mathematics MCA-III.

For the MTAS, Minnesota’s testing vendor produced an OIB using both operational and field-test items to more fully represent the range of academic achievement encompassed within the MTAS item bank. The details of the OIB construction can be found in the *Minnesota Assessments Summer 2014 Standard Setting: Recommended Performance Standards for Series III Mathematics Assessments.*

The Standard Setting Meeting

Before beginning the standard setting activities, MDE and Minnesota’s testing contractor staff briefed the committees on the purpose of the panel meeting and use of the outcomes. Specifically, panelists were advised that the principal outcome was a set of cut score recommendations. The panelists were informed that the educator committees were one of many components in the complete policy-making process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the Commissioner of Education. The participants were given an overview of standard setting and were introduced to the standard setting procedure they would be using (BSSP or ID Matching). Panelists then broke into their different assessment groups. Next, panelists used the previously developed achievement-level descriptors to help them generate threshold descriptors as a group. After creating the threshold descriptors and completing standard setting training and practice activities, the committee began the process of setting standards.

The MCA-III standard setting meetings were conducted in a series of three rounds of setting bookmarks. After the Round 1 cuts were made, psychometricians evaluated results and produced feedback forms for each table and for the room as a whole. The feedback forms for each table contained summary statistics showing the median, lowest, and highest cut scores for that table, as well as all individual bookmark placements. The room feedback form contained summary statistics showing the median, lowest, and highest cut scores for each table. After completing discussions on the Round 1 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards,* followed by *Partially Meets* and *Exceeds.* After the second round, in addition to the room form an impact data sheet containing OIB pages and the percentage of students at or above the level for each possible cut score was provided to the facilitator for reference and discussion. After completing discussions on the Round 2 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards*, followed by *Partially Meets* and *Exceeds.*

The MTAS standard setting meeting was also conducted in a series of three rounds but instead used the ID Matching method. Panelists began the standard setting process by identifying the threshold region between *Partially Meets the Standards* and *Meets the Standards* achievement levels. This entailed indicating the first item in the OIB that clearly matched the *Meets the Standards* ALD and the last page that clearly matched the Partially Meets ALD. The pages in between defined the threshold region in which panelists placed their cut scores. After identifying the threshold region, panelists were instructed to examine each item in the threshold region to determine the “first item that more closely matches the ALD for Meets *the Standards* than it matches the ALD for Partially Meets.” Panelists marked that item as their cut score. Panelists were instructed to use the same process to determine the threshold region and cut scores for *Partially Meets* and *Exceeds.* The same feedback was given to the MTAS participants as was given to the MCA-III panelists.

A description of the activities of each of the three rounds is given below.

Round 1

After completion of the practice activities, panelists were provided with the OIB associated with their assessment. For security purposes, all books were numbered so that distributed materials could be easily monitored and accounted for. After a brief review of the format of the OIB, panelists were instructed to begin their independent review of the items. Specifically panelists were instructed to do the following:

- Read each item in the OIB thinking about the knowledge, skills and abilities required to answer the item correctly.
- Record comments or notes about competencies required to address a given item in the OIB.
- Think about how students of different achievement levels should perform on each item.
- MTAS panelists were also asked to identify the threshold region between *Partially Meets the Standard*s and *Meets the Standards* achievement levels.

After the panelists completed their review, they were given a Readiness Survey and proceeded to make their first round of recommendations. MCA panelists did this by placing their bookmarks for *Partially Meets the Standards, Meets the Standards* and *Exceeds the Standards,* while keeping in mind their descriptions of the target students, the Achievement Level Descriptors (ALDs) and the Minnesota Academic Standards. MTAS panelists identified their threshold region, and were instructed to examine each item in the threshold region to determine the “first item that more closely matches the ALD for Meets the Standards than it matches the ALD for Partially Meets.” Panelists marked that item as their cut score.

Round 2

During Round 2, participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean, and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variation among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants again placed their bookmarks or identified their cut scores. Participants were reminded that this is an individual activity.

Round 3

During Round 3, participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 2 recommendations as well as impact data that was given to the facilitator. Each table was instructed to discuss their Round 2 recommendations with the goal of identifying major sources of variation among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants placed their final bookmarks or identified their final cut scores. Participants were reminded that this is an individual activity.

Table 5.1 summarizes feedback by round.

Table 5.1. Summary of Feedback by Round

Workshop	Round	Data Presented: Anchor Grades	Data Presented: Grades With Interpolated Cuts
MCA	Round 2	• R1 Panelist feedback data	• R1 Panelist feedback data
	Round 3	• R2 Panelist feedback data • Series-II MCA historical impact data • Series-III MCA operational impact data • College and career-ready benchmark data (grade 10 MCA panel only)	• R2 Panelist feedback data • Series-II MCA historical impact data • Series-III MCA operational impact data
MTAS	Round 2	• R1 Panelist feedback data	• R1 Panelist feedback data
	Round 3	• R2 Panelist feedback data • Series-II MTAS historical impact data • Series-III MTAS operational impact data	• R2 Panelist feedback data • Series-II MTAS historical impact data • Series-III MTAS operational impact data

Table 5.2 shows the participant-recommended cut scores for MCA-III Mathematics after Round 3. Table 5.4 shows the participant-recommended cut scores for MTAS Mathematics after Round 3. Cut scores are shown on the theta metric. For the MTAS assessments, final cut scores were identified by selecting the observed theta score nearest to the theta value associated with the panelists' recommended page number in the OIB. The nearest observed theta in the operational test form raw score to theta table is the final recommended cut. Table 5.3 and Table 5.5 show the impact data, the percentage of students in each of the four performance categories, based on the cut scores after Round 3.

Table 5.2. Participant-Recommended Cut Scores for Mathematics MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Mathematics	11	-0.5371	0.1034	0.9989

Table 5.3. Impact Data Associated with Participant-Recommended Cut Scores for MCA-III

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	11	28	22	31	19

Table 5.4. Participant-Recommended Cut Scores for Mathematics MTAS

Content Area	Grade	Cut Scores (Theta Metric) Partially Meets	Cut Scores (Theta Metric) Meets	Cut Scores (Theta Metric) Exceeds
Mathematics	11	1.0260	1.6731	2.8329

Table 5.5. Impact Data Associated with Participant-Recommended Cut Scores for MTAS

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	11	30	21	39	10

Commissioner-Approved Results

After the standard setting meeting, the Minnesota Commissioner of Education reviewed the recommended cut scores for overall consistency and continuity. All of the panelist-recommended cut scores were approved by the commissioner for the 2014 MCA-III administration.

Standard Setting for Reading Minnesota Comprehensive Assessments-Series III and Reading Minnesota Test of Academic Skills

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was implemented for the Minnesota Comprehensive Assessments-Series III (MCA-III) standard setting held in Roseville, Minnesota, on June 24–26, 2013. Achievement-level cut scores were established for MCA reading in grades 3–8 and 10.

The ID Matching procedure (Ferrara, Perie, & Johnson, 2002) was used to recommend performance standards for the Reading MTAS-III assessments held in Roseville, Minnesota, on June 27–28, 2013. Although similar to the widely implemented Bookmark method, the ID Matching procedure asks panelists to indicate which of the achievement level descriptors is best matched by the knowledge and skill requirements necessary to respond successfully to each test item. Achievement-level cut scores were established for reading in grades 3–8 and 10.

The activities of the meeting are documented in a paper titled *Minnesota Assessments Summer 2013 Standard Setting: Recommending Performance Standards for Series-III Reading Assessments*. The report can be found under the heading “Standard Setting Technical Report: Reading (2013) on the MDE website at: http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html.

This section provides a summary of outcomes from the meeting. Minnesota’s testing contractor, the Minnesota Department of Education (MDE) and MDE’s National Technical Advisory Committee (TAC) worked together to design the standard setting activities so as to follow the same general procedures as the standard setting meeting for Science, Mathematics and Reading for Minnesota Comprehensive Assessments-Series II (MCA-II) and Minnesota Test of Academic Skills. Minnesota’s testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE convened separate educator panels to recommend performance standards for the Series III Reading MCA and MTAS assessments. Each panel was further divided into subpanels by grade band (3–4, 5–6, 7–8 and 10). Each sub-panel had its own facilitator and was physically separate from the other subpanels.

MDE invited approximately 10 participants from across Minnesota to set cut scores in each grade band. Each grade band had a lower grade and an upper grade for which panelists set standards. The details of the credentials and demographics of the participants can be found in the *Minnesota Assessments Summer 2013 Standard Setting: Recommending Performance Standards for Series-III Reading Assessments.*

Table Leaders

During the standard setting, participants are divided into groups, often called “tables.” Each table had one table leader who had been previously selected by MDE. Table leaders were expected to keep track of the table-level discussion and represent their committee’s point of view during the vertical articulation meeting. Table leaders were trained about their roles and responsibilities on Day 1 of the standard setting.

Ordered Item Booklets

Central to both the Bookmarking and ID Matching procedures is the production of an Ordered Item Booklet (OIB). As noted previously, while the OIB is often produced from only those items in the first operational test, it is rarely the case that a single operational test administration provides a comprehensive sampling of items across the range of content standards and difficulty. While recommending standards on the entire item bank may in some respects be ideal, including too many items makes review of the OIB overly burdensome. For the MCA-III Reading OIB, operational items from one of the 2013 test administration online fixed forms served as the base. For grades 3–8 the OIB was augmented with two additional operational passages and corresponding items, selected from other operational forms. MDE selected additional passages for inclusion in the OIB that complemented the content distribution of the selected operational form, in terms of standards and benchmarks assessed and the item types, and that targeted test information gaps in the OIB. This led to OIBs that included 59–70 items across grades 3–8. At grade 10, performance standards were recommended based on the single paper form, so that the core of the OIB comprised the operational items contained in that form. The grade 10 paper OIB was augmented using one field-test passage and associated items from that form, which led to a total of 57 items. For the grades 5–8 an OIB was created based on one of the two forms and additional field test questions that had been administered on the 2013 tests. This led to OIBs with 47–53 items across the grades. For the MTAS, Minnesota's testing vendor similarly produced an OIB using both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS item bank. The details of the OIB construction can be found in the *Minnesota Assessments Summer 2013 Standard Setting: Recommending Performance Standards for Series-III Reading Assessments.*

The Standard Setting Meeting

Before beginning the standard setting activities, MDE and Minnesota's testing contractor staff briefed the committees on the purpose of the panel meeting and use of the outcomes. Specifically, panelists were advised that the principal outcome was a set of cut score recommendations. The panelists were informed that the educator committees were one of many components in the complete policy-making process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the Commissioner of Education. The participants were given an overview of standard setting and were introduced to the standard setting procedure they would be using (BSSP or ID Matching). Panelists then broke into their grade-level groups. Next, panelists used the previously developed achievement level descriptors to help them generate threshold descriptors as a group. After coming up with the threshold descriptors and completing standard setting training and practice activities, the committee began the process of setting standards.

The MCA-III standard setting meetings were conducted in a series of three rounds of setting bookmarks. After the Round 1 cuts were made, psychometricians evaluated results and produced feedback forms for each table and for the room as a whole. The forms for each table contained summary statistics showing the median, lowest, and highest cut scores for that table, as well as all individual bookmark placements. The room form contained summary statistics showing the median, lowest, and highest cut scores for each table. After completing discussions on the Round 1 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards*, followed by *Partially Meets* and *Exceeds.* After the second round, in addition to the room form an impact data sheet containing OIB pages and the percentage of students at or above the level for each possible cut score was provided to the facilitator for reference and discussion. After completing discussions on the Round 2 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards*, followed by *Partially Meets* and *Exceeds.*

The MTAS standard setting meeting was also conducted in a series of three rounds but alternatively used the ID Matching method. Panelists began the standard setting process by identifying the threshold region between *Partially Meets the Standards* and *Meets the Standards* achievement levels. This entailed indicating the first item in the OIB that clearly matched the *Meets the Standards* ALD and the last page that clearly matched the Partially Meets ALD. The pages in between defined the threshold region in which panelists placed their cut scores. After identifying the threshold region, panelists were instructed to examine each item in the threshold region to determine the "first item that more closely matches the ALD for Meets the Standards than it matches the ALD for Partially Meets." Panelists marked that item as their cut score. Panelists were instructed to use the same process to determine the threshold region and cut scores for *Partially Meets* and *Exceeds.* The same feedback was given to the MTAS participants as was given to the MCA-III panelists.

A description of the activities of each of the three rounds is given below.

Round 1

After completion of the practice activities, panelists were provided with the OIB associated with their grade. For security purposes, all books were numbered so that distributed materials could be easily monitored and accounted for. After a brief review of the format of the OIB, panelists were instructed to begin their independent review of the items. Specifically panelists were instructed to do the following:

1. Read each item in the OIB thinking about the knowledge, skills and abilities required to answer the item correctly.
2. Record comments or notes about competencies required to address a given item in the OIB.
3. Think about how students of different achievement levels should perform on each item.
4. MTAS panelists were also asked to identify the threshold region between *Partially Meets the Standards* and *Meets the Standards* achievement levels.

After the panelists completed their review, they were given a Readiness Survey and proceeded to make their first round of recommendations. MCA panelists did this by placing their bookmarks for *Partially Meets the Standards*, *Meets the Standards* and *Exceeds the Standards,* while keeping in mind their descriptions of the target students, the achievement level descriptors and the Minnesota Academic Standards. MTAS panelists identified their threshold region and were instructed to examine each item in the threshold region to determine the “first item that more closely matches the ALD for Meets the Standards than it matches the ALD for Partially Meets.” Panelists marked that item as their cut score.

Round 2

During Round 2, participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean, and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variance among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants again placed their bookmarks or identified their cut scores. Participants were reminded that this is an individual activity.

Round 3

During Round 3, participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 2 recommendations as well as impact data that was given to the facilitator. Each table was instructed to discuss their Round 2 recommendations with the goal of identifying major sources of variance among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants placed their final bookmarks or identified their final cut scores. Participants were reminded that this is an individual activity.

Table 5.6 summarizes feedback by round.

Table 5.6. Summary of Feedback by Round

Workshop	Round	Data Presented: Anchor Grades	Data Presented: Grades With Interpolated Cuts
MCA	Round 2	• R1 Panelist feedback data	• R1 Panelist feedback data
	Round 3	• R2 Panelist feedback data • Series-II MCA historical impact data • Series-III MCA operational impact data • College and career-ready benchmark data (grade 10 MCA panel only)	• R2 Panelist feedback data • Series-II MCA historical impact data • Series-III MCA operational impact data
MTAS	Round 2	• R1 Panelist feedback data	• R1 Panelist feedback data
	Round 3	• R2 Panelist feedback data • Series-II MTAS historical impact data • Series-III MTAS operational impact data	• R2 Panelist feedback data • Series-II MTAS historical impact data • Series-III MTAS operational impact data

Table 5.7 shows the participant-recommended cut scores for MCA-III Reading after final moderation. Table 5.8 shows the participant-recommended cut scores for MTAS Reading after final moderation. Cut scores are shown on the theta metric. For the MTAS assessments, final cut scores were identified by selecting the nearest observable theta to the theta value associated with the panelists' recommended page number in the OIB. The nearest observable theta in the operational test form raw score to theta table is the final recommended cut.

Table 5.9 and 5.10 show the impact data, or the percentage of students in each of the four performance categories, based on the cut scores after final moderation, for MCA-III and MTAS Reading, respectively.

Table 5.7. Participant-Recommended Cut Scores (Final Moderation) for Reading MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Reading	3	-0.6589	-0.1085	1.1921
	4	-0.8084	-0.0495	1.1556
	5	-1.1292	-0.3252	1.0237
	6	-0.8162	-0.1754	0.9008
	7	-0.6654	-0.0325	1.0741
	8	-0.6514	-0.0261	1.0228
	10	-0.9714	-0.2318	0.8172

Table 5.8. Participant-Recommended Cut Scores (Final Moderation) for Reading MTAS

Content Area	Grade	Cut Scores (Theta Metric) Partially Meets	Cut Scores (Theta Metric) Meets	Cut Scores (Theta Metric) Exceeds
Reading	3	0.6611	1.1660	2.5183
	4	1.1928	1.6441	2.6145
	5	0.8677	1.5322	3.6884
	6	0.9286	1.7583	3.5801
	7	1.1819	2.3916	3.0936
	8	1.1021	1.9319	3.7007
	10	0.8784	1.6991	2.9514

Table 5.9. Impact Data Associated with Participant-Recommended Cut Scores (Final Moderation) for MCA-III

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Reading	3	25	18	44	13
	4	21	25	40	14
	5	15	22	46	18
	6	21	21	37	21
	7	26	22	37	16
	8	26	21	35	18
	10	17	22	38	23

Table 5.10. Impact Data Associated with Participant-Recommended Cut Scores (Final Moderation) for MTAS

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Reading	3	17	12	47	24
	4	20	12	24	44
	5	16	15	52	17
	6	17	14	39	30
	7	11	24	26	39
	8	17	18	36	29
	10	18	17	28	38

Vertical Articulation and Moderation

Following Round 3 bookmarking for the initial grades, a vertical moderation session was conducted to allow table leaders to evaluate recommended cut scores in the context of a system of standards across grade levels. Following evaluation of recommended cut scores across the initial grade levels (grades 3, 5, 7, and 10 for MCA and MTAS), table leaders from each of the panels could elect to modify the recommended cut scores to better articulate performance standards across grades. Following Round 3 for the remaining grades (grades 4, 6, and 8 for MCA and MTAS), a final moderation session was held to allow table leaders to evaluate the entire system of performance standards and make any final revisions.

Commissioner-Approved Results

After the standard setting meeting, the Minnesota Commissioner of Education reviewed the recommended cut scores for overall consistency and continuity. All of the panelist-recommended cut scores were approved by the commissioner for the 2013 MCA-III administration.

Standard Setting for Science Minnesota Comprehensive Assessments-Series III and Minnesota Test of Academic Skills

The Bookmark Standard Setting Procedure (BSSP; Lewis et al., 1996) was implemented for the Minnesota Comprehensive Assessments-Series III (MCA-III) standard setting held in Roseville, Minnesota, on June 25–26, 2012. Achievement-level cut scores were established for science in grades 5, 8 and high school.

The ID Matching procedure (Ferrara, Perie, & Johnson, 2002) was used to recommend performance standards for the science MTAS assessments held in Roseville, Minnesota, on June 27–28, 2012. Although similar to the widely implemented Bookmark method, the ID Matching procedure asks panelists to indicate which of the achievement level descriptors is best matched by the knowledge and skill requirements necessary to respond successfully to each test item. Achievement-level cut scores were established for science in grades 5, 8 and high school.

The activities of the meeting are documented in a paper titled *Minnesota Assessments Summer 2012 Standard Setting: Recommending Performance Standards in Grades 5, 8, and High School Science.* The report can be found under the heading “Standard Setting Technical Report: Science MCA-III and MTAS” on the MDE website at: http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html.

This section provides a summary of outcomes from the meeting. Minnesota’s testing contractor, the Minnesota Department of Education (MDE) and MDE’s National Technical Advisory Committee (TAC) worked together to design the standard setting activities so as to follow the same general procedures as the standard setting meeting for Science, Mathematics and Reading Minnesota Comprehensive Assessments-Series II (MCA-II) and Minnesota Test of Academic Skills. Minnesota’s testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE convened one panel for the science MCA-III standard setting workshop and a second panel to recommend performance standards for the science MTAS-III assessment. Each panel was further divided into subpanels by grade (5, 8 and high school). Each subpanel had its own facilitator and was physically separate from the other subpanels.

MDE invited approximately 30 participants from across Minnesota to set cut scores in each test. The details of the credentials and demographics of the participants can be found in the report *Minnesota Assessments Summer 2012 Standard Setting: Recommending Performance Standards in Grades 5, 8, and High School Science.*

Table Leaders

During the standard setting, participants are divided into groups, often called “tables.” Each table had one table leader who had been previously selected by MDE. Table leaders were expected to keep track of the table-level discussion and represent their committee’s point of view during the vertical articulation meeting. Table leaders were trained about their roles and responsibilities on Day 1 of the standard setting.

Ordered Item Booklets

Central to both the Bookmarking and ID Matching procedures is the production of an Ordered Item Booklet (OIB). As noted previously, while the OIB is often produced from only those items in the first operational test, it is rarely the case that a single operational test administration provides a comprehensive sampling of items across the range of content standards and difficulty. And while recommending standards on the entire item bank may in some respects be ideal, including too many items makes review of the OIB overly burdensome. For the grades 5 and 8 science MCA-III assessments, Minnesota’s testing vendor therefore developed an augmented OIB that was built on a proportional test blueprint but that included 70 items. The high school science MCA-III assessment contained sufficient items that it was not necessary to augment the OIB. For the MTAS, the test vendor similarly produced an OIB using both operational and field-test items to more fully represent the range of academic achievement encompassed within the MTAS item bank. The details of the OIB construction can be found in the report *Minnesota Assessments Summer 2012 Standard Setting: Recommending Performance Standards in Grades 5, 8, and High School Science.*

The Standard Setting Meeting

Before beginning the standard setting activities, MDE and Minnesota’s testing contractor staff briefed the committees on the purpose of the panel meeting and use of the outcomes. Specifically, panelists were advised that the principal outcome was a set of cut score recommendations. The panelists were informed that the educator committees were one of many components in the complete policy-making process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the Commissioner of Education. The participants were given an overview of standard setting and were introduced to the standard setting procedure they would be using (BSSP or ID Matching). Panelists then broke into their grade-level groups. Next, panelists used the previously developed achievement level descriptors to help them generate threshold descriptors as a group. After coming up with the threshold descriptors and completing standard setting training and practice activities, the committee began the process of setting standards.

The MCA-III standard setting meeting was conducted in a series of two rounds of setting bookmarks. After the Round 1 cuts were made, psychometricians evaluated results and produced feedback forms for each table and for the room as a whole. The forms for each table contained summary statistics showing the median, lowest, and highest cut scores for that table, as well as all individual bookmark placements. The room form contained summary statistics showing the median, lowest, and highest cut scores for each table. In addition, an impact data sheet containing OIB pages and the percentage of students at or above the level for each possible cut score was provided to panelists for reference and discussion. After completing discussions on the Round 1 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards,* followed by *Partially Meets* and *Exceeds.*

The MTAS standard setting meeting was also conducted in a series of two rounds but alternatively used the ID Matching method. Panelists began the standard setting process by identifying the threshold region between *Partially Meets the Standards* and *Meets the Standards* achievement levels. This entailed indicating the first item in the OIB that clearly matched the *Meets the Standards* ALD and the last page that clearly matched the *Partially Meets* ALD. The pages in between defined the threshold region in which panelists placed their cut scores. After identifying the threshold region, panelists were instructed to examine each item in the threshold region to determine the "first item that more closely matches the ALD for *Meets the Standards* than it matches the ALD for Partially Meets." Panelists marked that item as their cut score. Panelists were instructed to use the same process to determine the threshold region and cut scores for *Partially Meets* and *Exceeds.* The same feedback was given to the MTAS participants as was given to the MCA-III panelists. After completing discussions of the Round 1 feedback, panelists again worked through the OIB, placing their cut scores for *Meets* the Standard, followed by *Partially Meets* and *Exceeds.*

A description of the activities of each of the three rounds is given below.

Round 1

After completion of the practice activities, panelists were provided with the OIB associated with their grade. For security purposes, all books were numbered so that distributed materials could be easily monitored and accounted for. After a brief review of the format of the OIB, panelists were instructed to begin their independent review of the items. Specifically panelists were instructed to do the following:

1. Read each item in the OIB thinking about the knowledge, skills and abilities required to answer the item correctly.
2. Record comments or notes about competencies required to address a given item in the OIB.
3. Think about how students of different achievement levels should perform on each item.
4. MTAS panelists were also asked to identify the threshold region between *Partially Meets the Standards* and *Meets the Standards* achievement levels.

After the panelists completed their review, they were given a Readiness Survey and proceeded to make their first round of recommendations. MCA panelists did this by placing their bookmarks for *Partially Meets the Standards*, *Meets the Standards* and *Exceeds the Standards,* while keeping in mind their descriptions of the target students, the achievement level descriptors (ALDs) and the Minnesota Academic Standards. MTAS panelists identified their threshold region and were instructed to examine each item in the threshold region to determine the "first item that more closely matches the ALD for Meets the Standard than it matches the ALD for Partially Meets." Panelists marked that item as their cut score.

Round 2

During Round 2, participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variance among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants again placed their bookmarks or identified their cut scores. Participants were reminded that this is an individual activity.

Table 5.11 shows the participant-recommended cut scores for MCA-III Science, as taken from Round 2. Table 5.13 shows the participant-recommended cut scores for MTAS Science, as taken from Round 2. Table 5.12 and Table 5.14 show the impact data associated with the cut scores shown in in their respective tables (Table 5.11 and Table 5.13). Cut scores are shown on the theta metric.

Table 5.11. Participant-Recommended Cut Scores (Round 2) for Science MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Science	5	-0.81	-0.44	0.53
	8	-0.59	0.32	1.51
	HS	-0.69	0.07	1.04

Table 5.12. Impact Data Associated with Participant-Recommended Cut Scores for Science MCA-III

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Science	5	19.2	13.7	33.4	33.7
	8	27.2	29.5	33.5	9.9
	HS	23.2	25.3	34.3	17.3

Table 5.13. Participant-Recommended Cut Scores (Round 2) for Science MTAS

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Science	5	0.82	1.64	3.68
	8	0.61	1.12	2.87
	HS	0.33	1.33	1.86

Table 5.14. Impact Data Associated with Participant-Recommended Cut Scores for Science MTAS

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Science	5	13.7	15.4	50.0	21.0
	8	13.0	9.3	49.2	28.5
	HS	13.4	21.2	20.7	44.7

Commissioner-Approved Results

After the standard setting meeting, the Minnesota Commissioner of Education reviewed the recommended cut scores for overall consistency and continuity. The final cut scores approved by the commissioner for the 2012 Science MCA-III administration are given in Table 5.15. Impact data associated with the final cut scores are reported in Table 5.16.

Table 5.15. Commissioner-Approved Science MCA-III Cut Scores

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Science	5	-0.81	-0.09	1.35
	8	-0.59	0.32	1.51
	HS	-0.69	0.07	1.04

Table 5.16. Impact Data Associated with Commissioner-Approved Science MCA-III Cut Scores

Content Area	Grade	Percentage of Students in Achievement Level: Does Not Meet (%)	Percentage of Students in Achievement Level: Partially Meets (%)	Percentage of Students in Achievement Level: Meets (%)	Percentage of Students in Achievement Level: Exceeds (%)
Science	5	20.1	23.1	44.9	11.9
	8	27.2	29.5	33.5	9.9
	HS	23.2	25.3	34.3	17.3

The final cut scores approved by the commissioner for the 2012 Science MTAS administration are given in Table 5.17. Impact data associated with the final cut scores are reported in Table 5.18.

Table 5.17. Commissioner-Approved Science MTAS Cut Scores

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Science	5	0.82	1.64	3.68
	8	0.61	1.12	2.87
	HS	0.33	1.33	2.36

Table 5.18. Impact Data Associated with Commissioner-Approved Science MTAS Cut Scores

Content Area	Grade	Percentage of Students in Achievement Level: Does Not Meet (%)	Percentage of Students in Achievement Level: Partially Meets (%)	Percentage of Students in Achievement Level: Meets (%)	Percentage of Students in Achievement Level: Exceeds (%)
Science	5	13.7	15.4	50.0	21.0
	8	13.0	9.3	49.2	28.5
	HS	13.4	21.2	31.2	34.2

Standard Setting for Grades 3–8 Mathematics Minnesota Comprehensive Assessments-Series III

The Bookmark Standard Setting Procedure (BSSP; Lewis et al., 1996) was implemented for the Minnesota Comprehensive Assessments-Series III (MCA-III) standard setting held in Roseville, Minnesota, on June 27–29, 2011. Achievement-level cut scores were established for mathematics in grades 3–8. The activities of the meeting are documented in a paper titled *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified*. The report can be found under the heading “Standard Setting Technical Report (2011)” on the MDE website at:
http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html.

This section provides a summary of outcomes from the meeting. Minnesota’s testing contractor, the Minnesota Department of Education (MDE) and MDE’s National Technical Advisory Committee (TAC) worked together to design the standard setting activities so as to follow the same general procedures as the standard setting meeting for Mathematics and Reading Minnesota Comprehensive Assessments-Series II (MCA-II). Minnesota’s testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE invited approximately 14–15 participants from across Minnesota to set cut scores in each grade band. Each grade-band had a lower grade and an upper grade for which panelists set standards. The details of the credentials and demographics of the participants can be found in the *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.*

Table Leaders

During the standard setting, participants are divided into groups, often called “tables.” Each table had one table leader who had been previously selected by MDE. Table leaders were expected to keep track of the table-level discussion and represent their committee’s point of view during the vertical articulation meeting. Table leaders were trained about their roles and responsibilities on Day 1 of the standard setting.

Ordered Item Booklets

The ordered item booklets (OIB) contained 60 operational items from the 2011 MCA-III exams that spanned the range of content, item types, and difficulty represented on a typical test. The details of the OIB construction can be found in the *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.*

The Standard Setting Meeting

Before beginning the standard setting activities, MDE and Minnesota’s testing contractor staff briefed the committees on the purpose of the panel meeting and use of the outcomes. Specifically, panelists were advised that the principal outcome was a set of cut score recommendations. The panelists were informed that the educator committees were one of many components in the complete policy-making process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the Commissioner of Education. The participants were given an overview of standard setting and were introduced to the BSSP. Panelists then broke into their grade-level groups. Next, panelists used the previously developed achievement level descriptors to help them generate threshold descriptors as a group. After coming up with the threshold descriptors and completing standard setting training and practice activities, the committee began the process of setting standards. The standard setting meeting was conducted in a series of three rounds of setting bookmarks. Rounds 1 and 2 recommendations were first completed for the lower grade followed by Rounds 1 and 2 for the upper grade. Round 3 recommendations were made for both grades concurrently after the review of Round 2 impact across grades. A description of the activities of each of the three rounds is given below.

Round 1

After completion of the practice activities, panelists were provided with the OIB associated with the lower grade in their grade band. For security purposes, all books were numbered so that distributed materials could be easily monitored and accounted for. After a brief review of the format of the OIB, panelists were instructed to begin their independent review of the items. Specifically panelists were instructed to do the following:

1. Read each item in the OIB thinking about the knowledge, skills and abilities required to answer the item correctly.
2. Record comments or notes about competencies required to address a given item in the OIB.
3. Think about how students of different achievement levels should perform on each item.

After the panelists completed their review for the lower grade they completed a Readiness Survey and proceeded to make their first round of recommendations by placing their bookmarks for *Partially Meets the Standards, Meets the Standards* and *Exceeds the Standards,* while keeping in mind their descriptions of the target students, the achievement level descriptors and the Minnesota Academic Standards.

Round 2

During Round 2, participants discussed their bookmark placements in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variance among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants again placed their bookmarks. Participants were reminded that bookmark placement is an individual activity.

Following placing bookmarks for Round 2 of the lower grade, Round 1 and Round 2 were repeated for the upper grade.

Round 3

At the beginning of Round 3, historical impact or relevant impact data were presented to the panelists as external reference. For MCA-III, 2006–2010 MCA-II impact data were presented. Then, results based on Round 2 recommendations were provided for both the lower and upper grade levels. First, table and group level summary data were distributed for the lower grade. Next, the impact data associated with the panelists' median recommendations for the lower-grade were presented for discussion. As a group, panelists were given the opportunity to discuss and react to the recommendations and impact associated with the lower grade level. They were then presented with this same information and data for the upper grade level. After the results for each grade were reviewed separately, the facilitator presented the total group impact data for the two grades side by side. Panelists were asked to think about whether the observed impact made sense in light of the ALDs, the test taking population, and the requirements of the assessment.

Table leaders were reminded to take notes throughout the impact discussions so that they could accurately represent the impressions of their committee at the vertical articulation meeting. After group discussion panelists were asked to make their final, Round 3 recommendations. Panelists were reminded that they must be able to defend any changes from a content perspective and should not arbitrarily change their rating in the hope to affect impact. After Round 3, panelists were asked to check in their materials and complete the meeting evaluation. This was the end of the regular by grade-level standard setting activities. Complete details on the standard setting process followed can be found in the *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.*

Table 5.19 shows the participant-recommended cut scores, as taken from participants' Round 3 bookmark placements. Cut scores are shown on the theta metric. Table 5.20 shows the impact data associated with the cut scores shown in Table 5.19.

Table 5.19. Participant-Recommended Cut Scores (Round 3) for Mathematics MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Mathematics	3	-1.21	-0.51	0.61
	4	-1.05	-0.43	0.42
	5	-0.86	-0.03	1.04
	6	-0.72	0.06	0.95
	7	-1.19	0.08	0.95
	8	-0.82	-0.03	0.84

Table 5.20. Impact Data Associated with Participant-Recommended Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	14	17	41	28
	4	17	17	32	34
	5	21	27	36	15
	6	25	27	30	17
	7	14	38	30	18
	8	22	26	31	21

Vertical Articulation

Articulation panelists are stakeholders in the results of the assessment system from a broad range of perspectives. Members of an articulation panel include representatives from teacher and administrator professional education organizations, business, higher education, the Minnesota state legislature, parent organizations and the community at large. The role of the articulation panel is to review the recommendations of the content experts and make further recommendations based on the effect that the results would have on the educational system and its members. A subset of the panelists who participated in standard setting, as well as other stakeholders, participated in the vertical articulation.

For the stakeholders who did not participate in the grade-level standard setting activities, an orientation was provided by Minnesota's testing contractor staff. Standard setting method, process and relevant materials were provided so that stakeholders could get an overview of the work that had been completed. Next, stakeholders joined the table leaders in the respective committees for the vertical articulation process.

The steps in the vertical articulation process were as follows:

1. Panelists reviewed the ALDs associated with all grades.
2. Panelists reviewed historical or relevant impact for the assessment.
3. As a group, the panelists discussed their expectations for impact across the grade levels in light of the ALDs and content assessed in each grade.
4. The group reviewed the impact associated with the Round 3 recommended cut scores across all grades and then discussed the extent to which the data mirrored their expectations.
5. As a group the committee discussed how/if the cut scores should be adjusted to provide for impact more consistent with their expectations.
6. Panelists were instructed that, after the meeting, their percentages recommendations would be compared to the content recommendations to make sure that the vertical articulation recommendations are within the range of variability from the content recommendations.
7. Panelists made independent recommendations as to the percentage of students testing in 2011 that they believed should fall in each level for each grade. Panelists were reminded that the goal was to make a recommendation that considered both the content-based ratings (from Round 3) and their expectations.
8. Impact recommendations were entered and the median recommended impact percentages associated with each achievement level in a grade were provided for review and discussion.
9. The panelists were asked to discuss whether the median impact percentages appropriately represented expected impact for the test taking population. The result was a final set of impact recommendations for each assessment.
10. Panelists completed evaluations.

After the completion of vertical articulation, the final recommended impact for each grade within an assessment was mapped back to the obtained 2011 frequency distribution to identify the raw scores or IRT theta values that would provide for impact as similar to that recommended as possible. Table 5.21 shows the cut scores from the vertical articulation.

Table 5.22 shows the impact data associated with the cut scores shown in Table 5.21.

Table 5.21. Vertical Articulation Panel's Smoothed Cut Scores for Mathematics MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Mathematics	3	-1.22	-0.52	0.60
	4	-1.06	-0.44	0.57
	5	-0.88	-0.04	1.01
	6	-0.75	0.03	0.96
	7	-0.91	0.03	0.94
	8	-0.83	-0.03	0.83

Table 5.22. Impact Data Associated with Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	14	17	41	28
	4	17	17	37	29
	5	21	27	36	16
	6	24	27	32	17
	7	20	30	32	18
	8	22	26	31	21

Commissioner-Approved Results

After the standard setting meeting, the Minnesota Commissioner of Education reviewed the recommended cut scores for overall consistency and continuity. The final cut scores approved by the commissioner for the 2011 MCA-III administration are given in Table 5.23. Impact data associated with the final cut scores are reported in Table 5.24.

Table 5.23. Commissioner-Approved Cut Scores for Mathematics MCA-III

Content Area	Grade	Cut Scores (Theta Metric): Partially Meets	Cut Scores (Theta Metric): Meets	Cut Scores (Theta Metric): Exceeds
Mathematics	3	-1.22	-0.52	0.60
	4	-1.06	-0.44	0.57
	5	-0.88	-0.04	1.01
	6	-0.75	0.03	0.96
	7	-0.91	0.03	0.94
	8	-0.83	-0.03	0.83

Table 5.24. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2006 Percentage of Students in Achievement Level: Does Not Meet (%)	2006 Percentage of Students in Achievement Level: Partially Meets (%)	2006 Percentage of Students in Achievement Level: Meets (%)	2006 Percentage of Students in Achievement Level: Exceeds (%)
Mathematics	3	14	17	41	28
	4	17	17	37	29
	5	21	27	36	16
	6	24	27	32	17
	7	20	30	32	18
	8	22	26	31	21

Standard Setting for Grades 3–8 Mathematics Minnesota Test of Academic Skills (MTAS)

Because the Minnesota Test of Academic Skills (MTAS) is composed of a small number of observations of student achievement, the test design is not ideal for the use of the Bookmark Standard Setting Procedure, which was used for the Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III). Instead, the Modified Angoff, a test-centered standard setting method (Jaeger, 1989) that has been used successfully in many states and by many publishers, along with some features of the Reasoned Judgment method (Kingston, Kahl, Sweeney, & Bay, 2001) was used. The standard setting meeting was held in Roseville, Minnesota, on June 29–30, 2011. Achievement-level cut scores were established for mathematics in grades 3–8. The activities of the meeting are documented in a paper titled *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.* The report can be found at the MDE website at http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html or is available from MDE upon request.

This section provides a summary of outcomes from the meeting. Minnesota's testing contractor, the Minnesota Department of Education (MDE) and MDE's National Technical Advisory Committee (TAC) worked together to design the standard setting activities. Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE invited approximately 12–14 participants from across Minnesota to set cut scores in each grade band. Each grade band had a lower grade and an upper grade for which panelists set standards. The invitation approach differed from that of the Mathematics MCA-III in that approximately half of the invited participants were educators involved in special education either through academic specialty or classroom experience. The details of the credentials and demographics of the participants can be found in the *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.*

Table Leaders

During the standard setting, participants were divided into groups, called "tables." Each table had one table leader that had been previously selected by MDE. Table leaders were expected to keep track of the table-level discussion and represent their committee's point of view during the vertical articulation meeting. Table leaders were trained about their roles and responsibilities on Day 1 of the standard setting.

Task Book

The Task Book contained all of the operational tasks from the 2011 MTAS. The tasks were ordered in the same sequence as they appeared on the test.

The Standard Setting Meeting

Before beginning the standard setting activities, MDE and Minnesota's testing contractor staff briefed the committees on the purpose of the panel meeting and use of the outcomes. Specifically, panelists were advised that the principal outcome was a set of cut score recommendations. The panelists were informed that the educator committees were one of many components in the complete policy-making process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the Commissioner of Education. The participants were given an overview of standard setting and were introduced to the Modified Angoff standard setting methodology. Panelists then broke into their grade-level groups. Next, panelists used the previously developed achievement level descriptors to help them generate threshold descriptors as a group. After coming up with the threshold descriptors and completing standard setting training and practice activities, the committee began the process of setting standards. The standard setting meeting was conducted in a series of three rounds, with the first two rounds using Modified Angoff and the third round using Reasoned Judgment. Rounds 1 and 2 recommendations were first completed for the lower grade, followed by Rounds 1 and 2 for the upper grade. Round 3 recommendations were made for both grades concurrently after the review of Round 2 impact across grades. A description of the activities of each of the three rounds is given below.

Round 1

After completion of the practice activities, panelists were provided with the Task Book associated with the lower grade in their grade band. For security purposes, all books were numbered so that distributed materials could be easily monitored and accounted for. After a brief review of the format of the Task Book, panelists were instructed to begin their independent review of the tasks. Specifically panelists were instructed to do the following:

1. Read each task in the book thinking about the knowledge, skills and abilities required to answer the item correctly.
2. Record comments or notes about competencies required to address a given task in the book.
3. Think about how students of different achievement levels should perform on each item.

After the panelists completed their review for the lower grade they completed a Readiness Survey and proceeded to make their first round of recommendations using Modified Angoff for *Partially Meets the Standards, Meets the Standards* and *Exceeds the Standards,* while keeping in mind their descriptions of the target students, the achievement level descriptors and the Minnesota Academic Standards.

Round 2

During Round 2, participants discussed their Round 1 recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variance among panelists. Understanding, rather than consensus, was the ultimate goal of the discussion.

After the discussion, participants made their Round 2 recommendations. Participants were reminded that making their recommendations is an individual activity.

Following making recommendations for Round 2 of the lower grade, Round 1 and Round 2 were repeated for the upper grade.

Round 3

At the beginning of Round 3, historical impact or relevant impact data were presented to the panelists as external reference. For MTAS, 2006–2010 MCA-II impact data were presented as well as preliminary impact data from Mathematics MCA-III. Then, results based on Round 2 recommendations were provided for both the lower and upper grade levels. First, table and group level summary data were distributed for the lower grade. Next, the impact data associated with the panelists' median recommendations for the lower-grade were presented for discussion. As a group, panelists were given the opportunity to discuss and react to the recommendations and impact associated with the lower grade level. They were then presented with this same information and data for the upper grade level. After the results for each grade were reviewed separately, the facilitator presented the total group impact data for the two grades side by side. Panelists were asked to think about whether the observed impact made sense in light of the ALDs, the test taking population, and the requirements of the assessment.

Table leaders were reminded to take notes throughout the impact discussions so that they could accurately represent the impressions of their committee at the vertical articulation meeting. After group discussion panelists were asked to make their final, Round 3 recommendations using the Reasoned Judgment methodology. Panelists were reminded that they must be able to defend any changes from a content perspective and should not arbitrarily change their rating in the hope to affect impact. After Round 3 panelists were asked to check in their materials and complete the meeting evaluation. This was the end of the regular by grade-level standard setting activities. Complete details on the standard setting process followed can be found in the *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified.* Table 5.25 shows the participant-recommended cut scores, as taken from participants' Round 3 judgment. Cut scores are shown on the raw score metric. Table 5.26 shows the impact data associated with the cut scores shown in Table 5.25.

Table 5.25. Participant-Recommended Cut Scores (Round 3) for Mathematics MTAS

Content Area	Grade	Cut Scores: Partially Meets	Cut Scores: Meets	Cut Scores: Exceeds
Mathematics	3	13	17	24
	4	14	17	24
	5	12	19	25
	6	11	17	24
	7	12	18	21
	8	12	16	21

Table 5.26. Impact Data Associated with Participant-Recommended Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	15	13	38	34
	4	14	8	52	26
	5	12	31	45	12
	6	15	24	51	11
	7	15	30	28	27
	8	18	12	37	33

Vertical Articulation

Articulation panelists are stakeholders in the results of the assessment system from a broad range of perspectives. Members of an articulation panel include representatives from teacher and administrator professional education organizations, business, higher education, the Minnesota state legislature, parent organizations and the community at large. The role of the articulation panel is to review the recommendations of the content experts and make further recommendations based on the effect that the results would have on the educational system and its members. A subset of the panelists who participated in standard setting, as well as other stakeholders, participated in the vertical articulation.

For the stakeholders who did not participate in the grade-level standard setting activities, an orientation was provided by Minnesota's testing contractor staff. Standard setting method, process and relevant materials were provided so that stakeholders could get an overview of the work that had been completed. Next, stakeholders joined the table leaders in the respective committees for the vertical articulation process.

The steps in the vertical articulation process were as follows:

1. Panelists reviewed the ALDs associated with all grades.
2. Panelists reviewed historical or relevant impact for the assessment.
3. As a group, the panelists discussed their expectations for impact across the grade levels in light of the ALDs and content assessed in each grade.
4. The group reviewed the impact associated with the Round 3 recommended cut scores across all grades and then discussed the extent to which the data mirrored their expectations.
5. As a group the committee discussed how/if the cut scores should be adjusted to provide for impact more consistent with their expectations.
6. Panelists were instructed that, after the meeting, their percentages recommendations would be compared to the content recommendations to make sure that the vertical articulation recommendations were within the range of variability from the content recommendations.
7. Panelists made independent recommendations as to the percentage of students testing in 2011 that they believed should fall in each level for each grade. Panelists were reminded that the goal was to make a recommendation that considered both the content-based ratings (from Round 3) and their expectations.
8. Impact recommendations were entered and the median recommended impact percentages associated with each achievement level in a grade were provided for review and discussion.
9. The panelists were asked to discuss whether the median impact percentages appropriately represented expected impact for the test taking population. The result was a final set of impact recommendations for each assessment.
10. Panelists completed evaluations.

After the completion of vertical articulation, the final recommended impact for each grade within an assessment was mapped back to the obtained 2011 frequency distribution to identify the raw scores or IRT theta values that would provide for impact as similar to that recommended as possible. Table 5.27 shows the cut scores from the vertical articulation. Table 5.28 shows the impact data associated with the cut scores shown in Table 5.27.

Table 5.27. **Vertical Articulation Panel's Smoothed Mathematics MTAS Cut Scores**

Content Area		Grade	Cut Scores: Partially Meets	Cut Scores: Meets	Cut Scores: Exceeds
Mathematics		3	13	17	24
		4	14	18	24
		5	12	19	25
		6	11	17	23
		7	12	18	21
		8	12	17	21

Table 5.28. **Impact Data Associated with Articulation Panel's Smoothed Cut Scores**

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	15	13	38	34
	4	14	13	47	26
	5	12	31	45	12
	6	15	24	45	17
	7	15	30	28	27
	8	18	18	32	33

Commissioner-Approved Results

After the standard setting meeting, the Minnesota Commissioner of Education reviewed the recommended cut scores for overall consistency and continuity. The final cut scores approved by the commissioner for the 2011 grades 3–8 Mathematics MTAS administration are given in Table 5.29. Impact data associated with the final cut scores are reported in Table 5.30.

Table 5.29. Commissioner-Approved Mathematics MTAS Cut Scores

Content Area	Grade	Cut Scores Partially Meets Raw Score	Cut Scores Partially Meets Theta	Cut Scores Meets Raw Score	Cut Scores Meets Theta	Cut Scores Exceeds Raw Score	Cut Scores Exceeds Theta
Mathematics	3	13	0.2223	17	0.9200	24	2.3096
	4	14	0.5616	18	1.2686	24	2.6098
	5	12	0.1670	19	1.5449	25	3.1260
	6	11	0.1852	17	1.6021	23	2.7431
	7	12	0.5059	18	1.6167	21	2.1074
	8	12	0.4167	17	1.4165	21	2.1020

Table 5.30. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2011 Percentage of Students in Achievement Level: Does Not Meet (%)	2011 Percentage of Students in Achievement Level: Partially Meets (%)	2011 Percentage of Students in Achievement Level: Meets (%)	2011 Percentage of Students in Achievement Level: Exceeds (%)
Mathematics	3	15	13	38	34
	4	14	13	47	26
	5	12	31	45	12
	6	15	24	45	17
	7	15	30	28	27
	8	18	18	32	33

Chapter 6: Scaling

The Minnesota assessments, such as the Minnesota Comprehensive Assessments-Series III (MCA-III) and the Minnesota Test of Academic Skills (MTAS), may be referred to as standards-based assessments. The tests are constructed to adhere rigorously to content standards defined by the Minnesota Department of Education (MDE) and Minnesota educators. For each subject and grade level, the content standards specify the subject matter the students should know and the skills they should be able to perform. In addition, as described in Chapter 5, performance standards are defined to specify how much of the content standards students need to demonstrate mastery of in order to achieve proficiency. Constructing tests to content standards ensures the tests assess the same constructs from one year to the next. However, although test forms across years may all measure the same content standards, it is inevitable the forms will vary slightly in overall difficulty or in other psychometric properties. Additional procedures are necessary to guarantee the equity of performance standards from one year to the next. These procedures create derived scores through the process of scaling (which is addressed in this chapter) and the equating of test forms (Chapter 7, “Equating and Linking”).

Rationale

Scaling is the process whereby we associate student performance with some ordered value, typically a number. The most common and straightforward way to score a test is to simply use the student’s total number correct. This initial score is called the raw score. Although the raw number-correct score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Consequently, the raw score is typically mathematically transformed (that is, scaled) to another metric on which test forms from different years are equated. Some tests, like the Minnesota Comprehensive Assessments-Series III (MCA-III) Mathematics assessment, do not use the raw score but instead use a model-based score as the initial score. However, tests like the MCA-III also tend to report on a scale score metric for ease of interpretation. Because the Minnesota assessments are standards-based assessments, the end result of the scaling process should be an achievement level that represents the degree to which students meet the performance standards. For accountability assessments, such as the Minnesota Comprehensive Assessments-Series III (MCA-III) and the Minnesota Test of Academic Skills (MTAS), the final scaling results are a designation of *Does Not Meet the Standards, Partially Meets the Standards, Meets the Standards,* or *Exceeds the Standards.*

Measurement Models

Item response theory (IRT) is used to derive the scale scores for all of the Minnesota tests. IRT is a general theoretical framework that models test responses resulting from an interaction between students and test items. The advantage of using IRT models in scaling is that all of the items measuring performance in a particular content area can be placed on the same scale of difficulty. Placing items on the same scale across years facilitates the creation of equivalent forms each year.

IRT encompasses a number of related measurement models. Models under the IRT umbrella include the Rasch Partial Credit (RPC; Masters, 1982), the two-parameter logistic model (2PL; Lord & Novick, 1968), the three-parameter logistic model (3PL; Lord & Novick, 1968), the generalized partial credit model (GPC; Muraki, 1992), as well as many others. A good reference text that describes commonly used IRT models is van der Linden and Hambleton (1997). These models differ in the types of items they can describe. For example, the 3PL model can be used with multiple-choice items but not with Minnesota's constructed-response items. Models designed for use with test items scored as right/wrong are called dichotomous models. These models are used with multiple-choice and gridded-response items. Models designed for use with items that allow multiple scores, such as constructed-response items, are called polytomous models. Both dichotomous and polytomous models are used for Minnesota assessments.

The models used on the Minnesota assessments can be grouped into two families. One family is the Rasch models, which include the dichotomous Rasch model for multiple-choice items and the RPC model for constructed-response items. Although the dichotomous Rasch model is mathematically a special case of the RPC model, for expository purposes the models are treated separately below. The second family of models is labeled 3PL/GPC and includes the GPC model for constructed-response items, the 3PL model for multiple-choice items, and the 2PL model for gridded-response items. Each model is described in the following sections.

Rasch Models

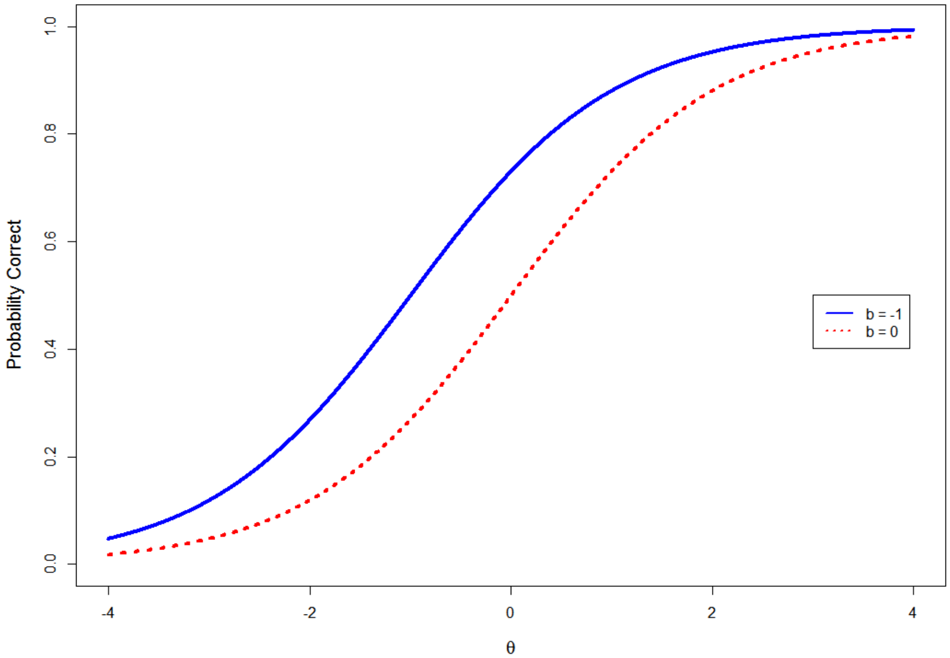
The dichotomous Rasch model can be written as the following mathematical equation, where the probability (P_{ij}) of a correct response for person *i* taking item *j* is given by:

$$P_{ij} = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)} = \frac{1}{1+\exp[-(\theta_i - b_j)]} \tag{6.1}$$

Student ability is represented by the variable θ (theta) and item difficulty by the model parameter *b*. Both θ and *b* are expressed on the same metric, ranging over the real number line, with greater values representing either greater ability or greater item difficulty. This metric is called the θ metric or θ scale. Typically, in Rasch scaling the θ metric is centered with respect to the particular item pool so that a value of zero represents average item difficulty. Often, but not always, the variable θ is assumed to follow a normal distribution in the testing population of interest.

The easiest way to depict the way item response data are represented by the Rasch model is graphically. Figure 6.1 displays the item response functions for two example items. The *x*-axis is the θ scale and the *y*-axis is the probability of a correct answer for the item. The solid curve on the left represents an item with a *b*-value of –1.0, and the dotted curve represents an item with a *b*-value of 0.0. A *b*-value of 0.0 signifies that a student of ability (that is, θ) = 0.0 has a 50% probability of correctly answering the question. The item with a *b*-value of –1.0 is an easier item, as a student with an ability (i.e., θ) of –1.0 has a 50% probability of making a correct answer to it. Students with abilities two or more theta units above the *b*-value for an item have a high probability of getting the item correct, whereas students with abilities two or more theta units below the *b*-value for an item have a low probability of getting the item correct.

Figure 6.1: Rasch Item Response Functions for Two Example Dichotomous Items



The RPC model is a polytomous generalization of the dichotomous Rasch model. The RPC model is defined via the following mathematical measurement model where, for a given item involving *m* score categories, the probability of person *i* scoring *x* on item *j* (where *k* is an index across categories) is given by:

$$P_{ijx} = \frac{exp\sum_{k=0}^{x}(\theta_i - b_{jk})}{\sum_{v=0}^{m_j-1} exp\sum_{k=0}^{v}(\theta_i - b_{jk})} \qquad (6.2)$$

where x = 0, 1, 2, . . . , m_j–1, and

$$\boldsymbol{\sum_{k=0}^{0}(\theta_i - b_{jk}) = 0}. \qquad (6.3)$$

The RPC model provides the probability of a student scoring *x* on task *j* as a function of the student's ability (*θ*) and the category boundaries (b_{jk}) of the m_j –1 steps in task *j*.

The RPC model essentially employs a dichotomous Rasch model for each pair of adjacent score categories. This gives rise to several *b*-parameters (called category boundary parameters) instead of a single *b*-parameter (item difficulty or location) in the dichotomous case. The item difficulty parameter in the dichotomous Rasch model gives a measure of overall item difficulty. In the polytomous model, the category boundary parameters provide a measure of the relationship between the response functions of adjacent score categories.

Figure 6.2 provides an example for a sample four-point polytomous item. The figure graphs the probability that a student at a given ability obtains a score in each of the five score categories. The "zero" curve, for example, plots the probability a student receives a score point zero on the ability scale. The category boundary parameter b_1 (= –1.5) is the value of θ at the crossing point of the "zero" response function and the "1" response function. Similarly, b_2 (= –0.3) is the value of θ at the crossing point of the response functions for score points "1" and "2," b_3 (= 0.5) is the value of θ at the crossing point of the response functions for score points "2" and "3" and b_4 (= 2) is the value of θ at the crossing point of the response functions for score points "3" and "4." The sample item has a fair spread of category boundary parameters, which is an indication of a well-constructed item. Category boundaries that are too close together may indicate the score categories are not distinguishing students in an effective manner.

Figure 6.2: Rasch Partial Credit Model Category Response Functions for Example Polytomous Item with b_1 = –1.5, b_2 = –0.3, b_3 = 0.5 and b_4 = 2

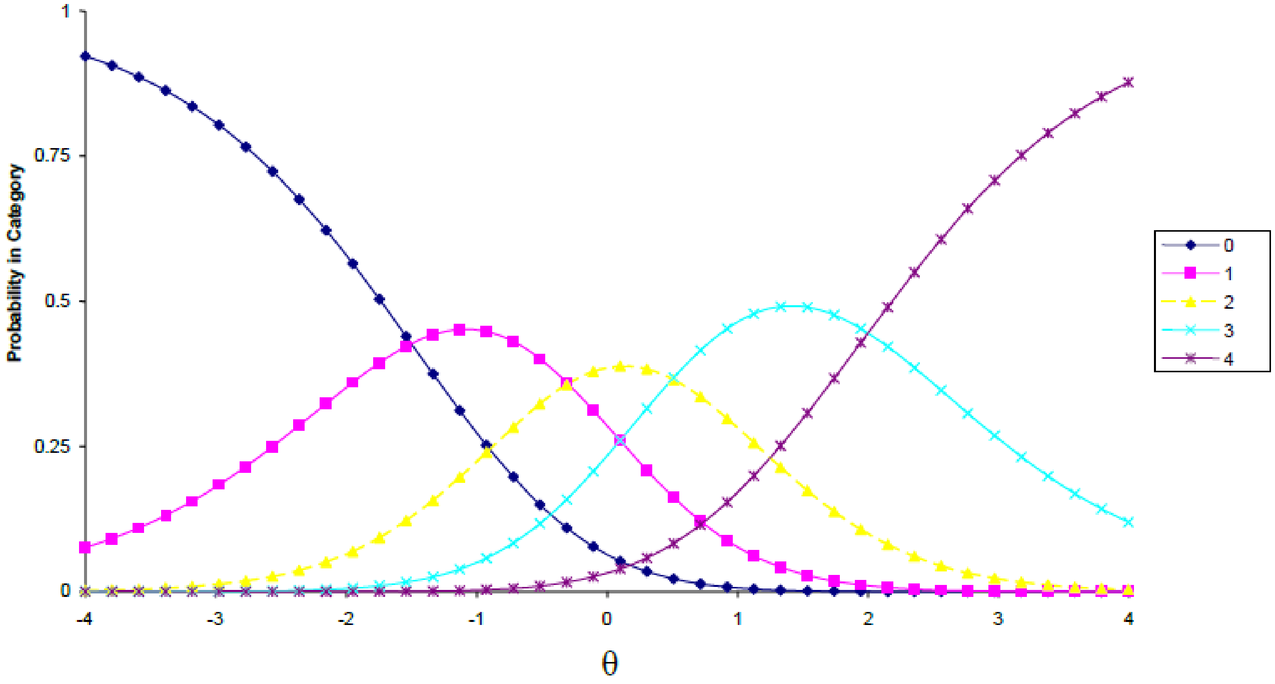
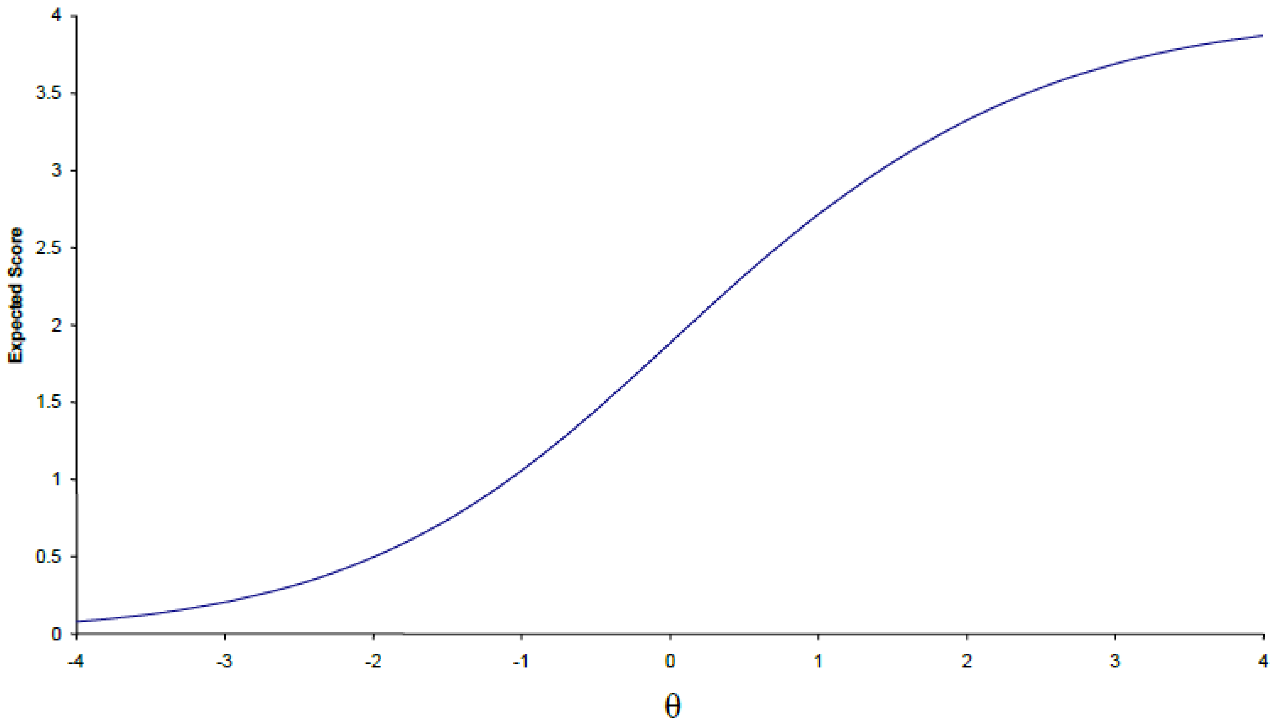


Figure 6.3 displays the average score for every ability value for the sample item given in Figure 6.2. The figure shows that students with ability θ = 0 should, on average, receive a score of "2" on the item, whereas students with ability at about 1 should average about 2.5 points on the item.

Figure 6.3: Rasch Partial Credit Model Item Expected Score Function for an Example Four-Point Item



Calibration of items for the Rasch models is achieved using the computer program WINSTEPS (Linacre, 2006). The program estimates item difficulty for multiple-choice items and category boundary parameters for polytomously scored (for example, constructed-response) items.

3PL/GPC Models

This section discusses three IRT measurement models: the 3PL model, the 2PL model, and the GPC model. The 3PL and 2PL models are used with dichotomous items and are each generalizations of the dichotomous Rasch model. The GPC model can be considered a generalization of the RPC model and the 2PL model.

The 3PL/GPC models differ from the Rasch models in that the former permit variation in the ability of items to distinguish low-performing and high-performing students. This capability is quantified through a model parameter, usually referred to as the *a*-parameter. Traditionally, a measure of an item's ability to separate high-performing from low-performing students has been labeled the "discrimination index" of the item, so the *a*-parameter in IRT models is sometime called the discrimination parameter. Items correlating highly with the total test score best separate the low- and high-performing students.

In addition to the discrimination parameter, the 3PL model also includes a lower asymptote (*c*-parameter) for each item. The lower asymptote represents the minimum expected probability an examinee has of correctly answering a multiple-choice item. For items scored right/wrong that are not multiple-choice, such as gridded-response items, the 2PL model is appropriate. The 2PL model is equivalent to fixing the lower asymptote of the 3PL model to zero.

The 3PL model is mathematically defined as the probability of person *i* correctly answering item *j:*

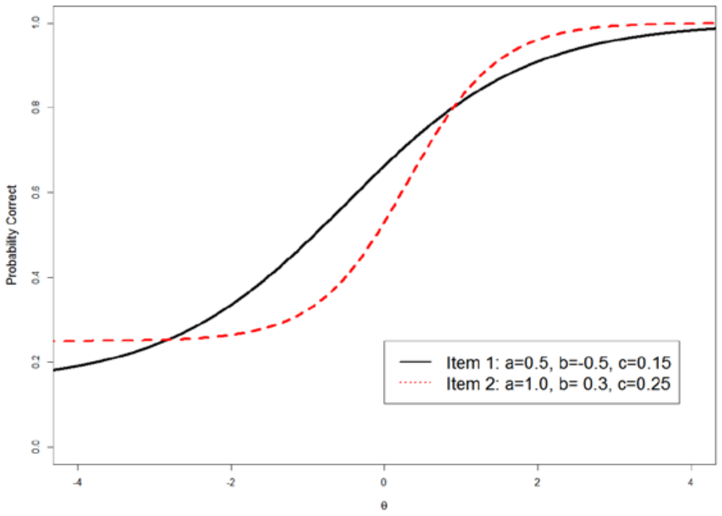
$$P_{ij} = c_j + \frac{1-c_j}{1+\exp\big(-1.7a_j(\theta_i - b_j)\big)} \tag{6.4}$$

where a_j, b_j, c_j are the item's slope (discrimination), location (difficulty), and lower asymptote parameters, and θ_i is the ability parameter for the person (Lord, 1980). The difficulty and ability parameters carry the same general meaning as in the dichotomous Rasch model. As stated before, the 2PL model can be defined by setting the *c*-parameter to zero. The 1.7 term in the expression is an arbitrary scaling factor that has historically been employed because inclusion of this term results in probabilities closely matching another dichotomous IRT model called the normal-ogive model. Equation 6.4 can be reduced to the standard Rasch equation (6.1) by setting $c = 0$, $a = 1$, and removing the 1.7 scaling constant.

Examples of 3PL model item-response functions are presented in Figure 6.4. Several differences from the Figure 6.1 Rasch model curves can be observed. First, a distinguishing characteristic of IRT models whose discrimination parameters allow the slopes of the curves to vary is that the item-response functions of two items may cross. The crossing of item-response functions cannot occur under the Rasch model because it requires that all items in a test have the same slope. Figure 6.4 shows the effect of crossing curves. For students in the central portion of the θ distribution, sample item 2 is expected to be more difficult than sample item 1. However, students with $\theta > 1.0$ or $\theta < –3.0$ have a higher expected probability of getting item 2 correct.

The figure also shows item 2 clearly has a non-zero asymptote ($c = 0.25$). Item 1 also has a non-zero asymptote ($c = 0.15$). However, due to the relatively mild slope of the curve, the asymptote is only reached for extreme negative θ values that are outside the graphed range. Finally, and in contrast to the Rasch or 2PL models, in the 3PL model the *b*-parameter does not indicate the point on the θ scale where the expected probability of a correct response is 0.50. However, in all three models the *b*-parameter specifies the inflection point of the curve and can serve as an overall indicator of item difficulty.

Figure 6.4: 3PL Item Response Functions for Two Sample Dichotomous Items



The polytomous IRT model described in this section is the GPC model. Instead of having a single probability correct, as in the 3PL model, the GPC model has a separate probability for each possible response category. The GPC model is mathematically defined as the probability of person *I* scoring in response category *k* for item *j:*

$$P_{ijk} = \frac{\exp[\sum_{v=1}^{k} 1.7a_j(\theta_i - b_j + d_{jv})]}{\sum_{c=1}^{m} \exp[\sum_{v=1}^{k} 1.7a_j(\theta_i - b_j + d_{jv})]} \tag{6.5}$$

where *m* is the number of response categories for the item and $d_{\mathrm{j}1} = 0$ (Muraki, 1997). The ability parameter is θ_i and the model's item parameters are a_j (slope/discrimination), b_j (location/difficulty), and d_{jk} (threshold parameters representing category boundaries relative to the item location parameter).

Figure 6.5 presents the category response functions for a sample item. The GPC model can be algebraically formulated in more than one fashion (Muraki, 1992). The formulation given above includes the location parameter indicating overall item difficulty. A consequence of having an overall location parameter, though, is the d_{jk} parameters have a different interpretation than the b_{jk} parameters in the RPC model. In the RPC model, the category boundary parameters are simply the θ values at crossing points of adjacent score categories. In the GPC model, the d_{jk} indicates how far the category boundaries are from the location parameter. They could be considered category boundary parameters that have been "offset" by the item's difficulty parameter. In Figure 6.5, for example, d_2 (= 3.7) is the distance on the θ scale that the crossing point for the "zero" and "1" curves is from the location parameter (b = 0.3); the *b*-parameter for this item is 3.7 units greater than the value of θ at the crossing point. As another example, *b* is one half of a unit *less* than the value of θ at the crossing point for the response functions for scores of "2" and "3" (because d_4 is negative). It remains the case for the GPC model that a good spread of the "offset" category boundary parameters indicates a well-functioning item.

Calibration of MCA-III items for the 3PL/GPC models is achieved using the computer program IRTPRO (Cai, Thissen, & du Toit, 2011). IRTPRO estimates parameters simultaneously for dichotomous and polytomous items via a statistical procedure known as marginal maximum likelihood. Simultaneous calibration of these items automatically puts their parameter estimates on the same scale. That scale is created on the assumption that test takers have a mean ability of approximately zero and a standard deviation of approximately one.

Model Selection

Regardless of the particular IRT models used for the items on the test, the relationship between expected performance and student ability is described by a key IRT concept called the test response function. Figure 6.6 displays what a test response function might look like for a reading test on the Minnesota Comprehensive Assessments-Series III (MCA-III). For each level of ability in the range of –4.0 to +4.0, the curve for the overall test score indicates expected performance on the number-correct scale. The graph shows that average ability students (θ = 0.0) can be expected to get a score of around 25 points. For a particular ability, the expected score is called the true score. The use of the test response function is an integral part of the scaling process for all of the Minnesota tests, as will be described in the next section. In addition to the overall test score function, response functions for the two subscores are also graphed in Figure 6.6.

Figure 6.5: Generalized Partial Credit Model Category Response Functions for Sample Polytomous Item with a = 0.4, b = 0.3, $d_1 = 0$, $d_2 = 3.7$, $d_3 = 0.75$, $d_4 = –0.5$, and $d_5 = –3$

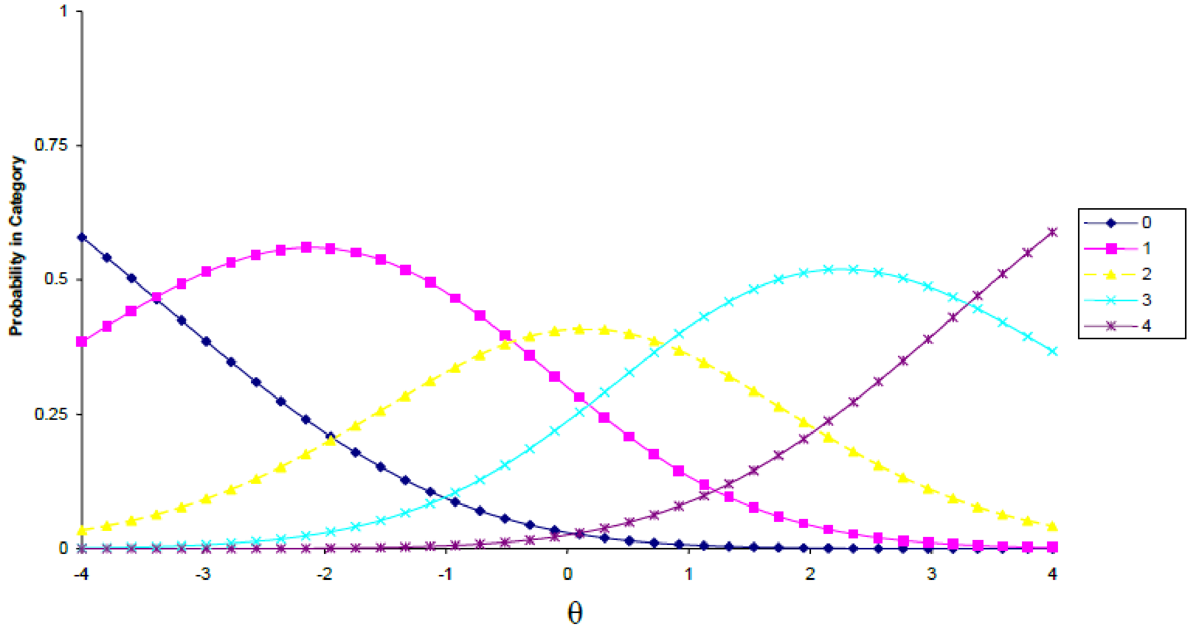
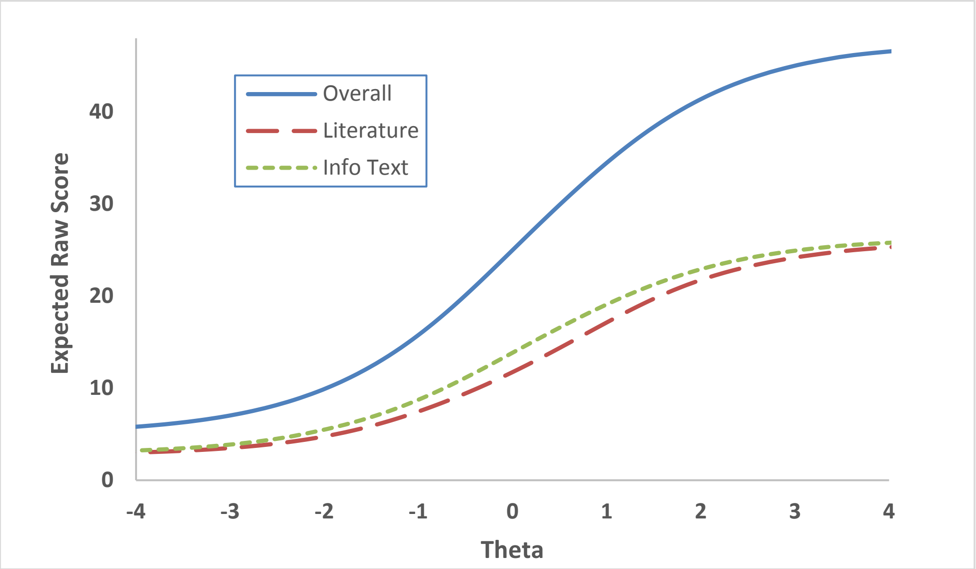


Figure 6.6: Sample Test Response Function for Reading MCA-III



In deciding how to model responses for a particular test, measurement specialists choose from among the developed IRT models based on a number of considerations. Some considerations include the number and type or format of items that comprise the test, expected calibration sample size, and other general measurement theory concerns. The RPC model is well suited to model the performance task-based Minnesota Test of Academic Skills (MTAS). The strengths of the Rasch models include their simplicity and flexibility. The Rasch model was specified for these tests because they are administered to relatively few students. The Rasch model generally performs better than more complex models when sample sizes are small.

Historically, the MCA tests were scaled using the Rasch model. With the advent of the MCA-II, the timing was right to consider using a different measurement model. The planned additional psychometric activities that included creating a vertical scale and linking the scales between the MCA-II and Mathematics Test for English Language Learners (MTELL) suggested a more complex model should be considered. After seeking the advice of the National Technical Advisory Council (TAC), the Minnesota Department of Education (MDE) determined the 3PL and GPC models would be used for the MCA-II. The 3PL model has been continued with the move to Minnesota Comprehensive Assessments-Series III (MCA-III) tests.

Scale Scores

The purpose of the scaled score system is to convey accurate information about student performance from year to year. The scaled score system used for the Minnesota assessments is derived from either the number-correct score or a measurement model–based score. These two initial scores are described below.

Number-Right Scoring

The number-correct score is calculated by summing the number of points the student is awarded for each item. Basing scores on number correct is easy to understand and to explain. However, test forms will undoubtedly vary slightly in difficulty across years, thus a statistical equating process is used to ensure the forms yield scores that are comparable. Because item response theory (IRT) is used in the equating process, in order for scores to be comparable across years, IRT must also play a role in assigning scores. The student’s number-correct score is transformed to an equated ability scale score through true score equating (Kolen & Brennan, 2004, Chapter 6). The true score equating procedure used is described in Chapter 7, “Equating and Linking” (in the “Latent-Trait Estimation” section). The spring 2012 administration is the base year for MCA-III and MTAS Science assessments. In administrations after 2012, the ability score metric is equated back to the spring 2012 base administration. In the case of assessments based on the Rasch measurement model (MTAS), the number right and model-based scoring approaches are mathematically equivalent. The base year for Mathematics MTAS grades 3–8 was 2011. The base year for Reading MTAS was 2013. The base year for Mathematics MTAS grade 11 is 2014.

Measurement Model–Based Scoring

The measurement model used for Minnesota’s assessments—item response theory (IRT)—permits the use of a statistically sophisticated method that is commonly referred to as pattern scoring because the scoring procedure takes the pattern of correct and incorrect responses into account. The Mathematics and Reading MCA-III assessments make use of pattern scoring to determine student scores. Unlike number-correct scoring, where students who get the same number of dichotomously scored questions correct receive the same score, pattern scoring of tests based on the 3PL or GPC model rarely results in students receiving the same scale score even though they have the same number-correct score, because typically they differ in the particular items they answered correctly. Because pattern scoring utilizes information from the entire student response pattern and gives greater weight to more discriminating items, this scoring method theoretically provides greater precision than does number-right scoring. The pattern scoring procedure used is described below in the “Latent-Trait Estimation” section.

Latent-Trait Estimation

For the Minnesota assessments, a measurement model–based score is obtained that represents student proficiency. This is called the latent-trait estimate or the theta score. Different Minnesota assessments obtain the theta score in different ways. The Minnesota Comprehensive Assessments-Series III (MCA-III) Mathematics and Reading assessments use a pattern scoring procedure, described below, to directly obtain the theta score from student responses of individual items. For other Minnesota assessments, a transformation from the raw total correct score to the theta scale is made. After the theta score is obtained, it is then transformed to the reported scale score. The theta-to-reported score transformation is described earlier in this chapter. Pattern scoring and the raw-to-theta transformation are described in this section.

Pattern Scoring

In pattern scoring the entire pattern of correct and incorrect student responses is taken into account. Unlike number-correct scoring, where students who get the same number of dichotomously scored questions correct receive the same score, in pattern scoring students rarely receive the same score, as even students getting the same number correct typically differ in the particular items they got correct or incorrect. Because pattern scoring utilizes information from the entire student response pattern, this type of scoring produces more reliable scores than does number-right scoring.

Students taking the MCA-III Mathematics and Reading assessments are assigned maximum likelihood scores, which are the items the student answers correctly and the difficulty of those items. The Minnesota assessments include multiple item types, much as multiple choice and technology-enhanced items. We can write the likelihood for scoring based on a generalized IRT model based on a mixture of items types as:

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N} \left[c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]}\right]^{x_i} \left[1 - c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]}\right]^{1-x_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N} \frac{\exp \sum_{k=1}^{x_i} Da_i(\theta - \delta_{ki})}{1 + \sum_{j=1}^{m_i} \exp \sum_{k=1}^{j} Da_i(\theta - \delta_{ki})}$$

where c_i is the lower asymptote of the item response curve (i.e., the guessing parameter), a_i is the slope of the item response curve (i.e., the discrimination parameter), b_i is the location parameter, x_i is the observed response to the item, *i* indexes the item, and δ_{ki} is the *k*th step for item *i* with *m* total categories.

By treating the item parameters as fixed, we subsequently find $\arg\max \theta\, L(\theta)$ as the student's theta (i.e., maximum likelihood estimator (MLE)) given the set of items administered to the student.

Raw-to-Theta Transformation

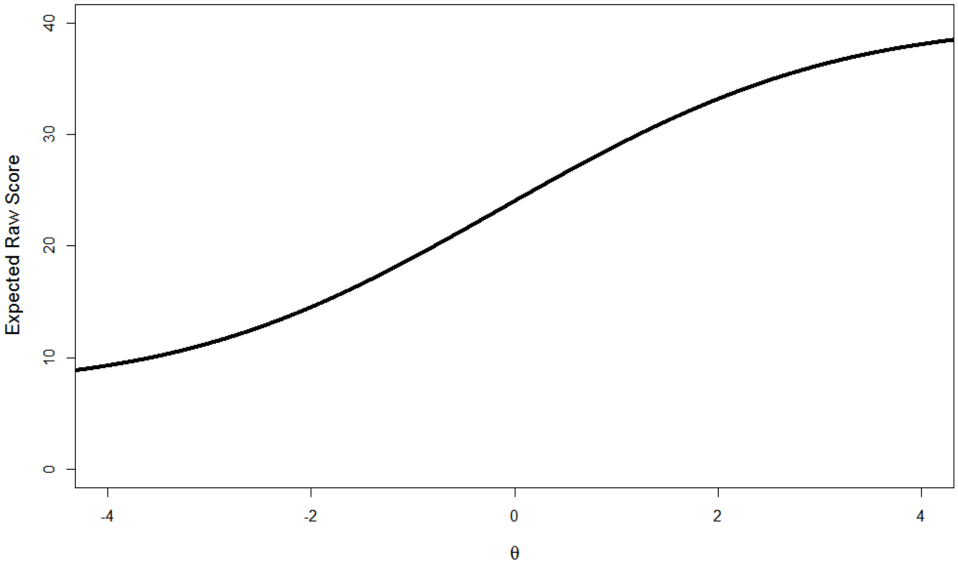
The raw-to-theta transformation can be described as a reverse table lookup on the test characteristic function. The test characteristic function can be defined as

$$\mathrm{TCF}(\theta) = \sum_{j=1}^{N} \sum_{K=0}^{m-1} kP_{ik}(\theta)$$

where *j* is an index of the *N* items on the test, *k* is an index of the *m* score categories for an item and $P_{ik}(\theta)$ is the item response model probability correct for the item. The test characteristic function is the expected raw score given the person proficiency value θ and the item-parameter values of the item response theory (IRT) model. Figure 6.7 presents the test characteristic function for a hypothetical 40-item multiple-choice test. For example, based on Figure 6.7, persons with θ proficiency equal to 2.0 would, on average, have a raw score of 33. Consequently, using reverse table lookup, a raw score of 33 would be assigned an estimated theta score of 1.0.

A variety of estimation procedures can be used to find the theta value that corresponds to a particular raw score. The Newton-Raphson method is a popular choice. For the Minnesota assessments, computer software packages such as WINSTEPS (Linacre, 2006) or POLYEQUATE (Kolen, 2004) are used to find the transformations.

Figure 6.7: Example Test Characteristic Function for 40-Item Test



Minnesota Comprehensive Assessments-Series III Scaling

In order to simplify comparison of student scores across years, the equated student ability estimates are transformed mathematically to a more convenient metric. For the Minnesota Comprehensive Assessments-Series III (MCA-III), the scaled metric ranges from 1 to 99 and is prefixed by the student's grade. For example, grade 5 test scores range from 501 to 599, and grade 8 test scores range from 801 to 899. The passing score to achieve *Meets the Standards* is set to g50, where g is the grade prefix. The cut score to achieve *Partially Meets the Standards* is set to g40. At grade 3, for example, students scoring below 340 are designated *Does Not Meet the Standards*, students with scores from 340 to 349 are designated *Partially Meets the Standards,* and a score of 350 to the next cut score is necessary to achieve *Meets the Standards.* The *Exceeds the Standards* achievement level score is not set to the same value across grades, but it generally ranges from g60 to g65.

Minnesota Comprehensive Assessments-Series III Transformation

The general transformation formula used to obtain scale scores for the Minnesota Comprehensive Assessments-Series III (MCA-III) is the following:

$$Scale\ Score = (\theta - \theta_{Std2}) * Spread + Center + Grade * 100 \tag{6.6}$$

Where θ is the post-equated ability estimate, θ_{Std2} is the ability cut score between *Partially Meets the Standards* and *Meets the Standards, Center* is set to be 50, *Grade* is the grade of the administered test, and *Spread* is a numerical constant unique for each subject-grade combination.

For MCA-III, the transformation formula uses cut scores on the θ scale (see Chapter 5, “Performance Standards”). For Mathematics and Reading MCA-III, the Commissioner of Education approved cut scores that were already on the θ scale. For Science MCA-III, the cut scores on the proficiency scale were obtained by using the test response function to find the θ values that corresponded to the approved raw score cuts.

One goal for the scale transformation was to make the proficiency level scale score cuts as consistent as possible across grades. Using a linear transformation–like equation (6.6) allows two of the three scale cut scores to be fixed. As stated above, the cut score for *Meets the Standards* was desired to be g50, where g is the grade prefix. This was accomplished by setting *Center* = 50. The cut score between *Does Not Meet the Standards* and *Partially Meets the Standards* was desired to equal g40. The *Spread* constant for each grade per subject combination was selected so as to force the first scale cut score to be equal to g40. The formula used to find the *Spread* is

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})} \tag{6.7}$$

where θ_{Std1} is the theta ability cut score between *Does Not Meet the Standards* and *Partially Meets the Standards,* and θ_{Std2} is the theta ability cut score between *Partially Meets the Standards* and *Meets the Standards*. The *Spread* value varies for each grade and subject combination. Because only two of the three scale cut scores can be predetermined using a linear transformation, the scale cut score between *Meets the Standards* and *Exceeds the Standards* was allowed to vary across grades and subjects.

The lowest observable scale score (LOSS) is set to g01 and the highest observable scale score (HOSS) is set to g99, where g is the grade. On grade 4 tests, for example, LOSS = 401 and HOSS = 499. The LOSS and HOSS prevent extreme student scores from being transformed outside the desired range of the scale. Because Science MCA-III uses raw to scale score conversion, some additional scoring rules are necessary. For Science MCA-III, restrictions are placed on the transformation for very high and very low scores. A score of all correct is always assigned the HOSS, regardless of the result of the transformation equation. A score of zero correct is awarded the LOSS. Further restrictions on the transformation are sometimes necessary for very high and very low scores on the Science MCA-III.

For high scores, it is desired that number-right scores less than all correct are given scale scores less than the HOSS. It is possible, however, that the transformation equation could scale number-right scores less than all correct to a value equal to or greater than the HOSS value. For these cases, adjustments are made so non-perfect number-correct scores are assigned a scale score below the HOSS. Usually, this adjusted scale score would be one less than the HOSS. For example, on a grade 5 test the transformation equation could scale the scores of students who get all but one multiple-choice item correct to a scale score equal to or greater than 599 (the HOSS). Because only students who score all correct are awarded a 599, students who get all but one correct would be assigned a score of 598.

For Mathematics and Reading MCA-III, all students are assigned a θ score by the scoring algorithm, so no further manipulation of the score is necessary. However, Science MCA-III scoring is based on raw scores and, when using IRT, special consideration is also necessary for scaling very low number-correct scores. For a test containing multiple-choice items, the expected number-correct score will always be greater than zero, because even a student who is guessing at random is expected to get some questions correct. As a consequence, in IRT expected (true) scores do not exist for raw scores below the chance level raw score; thus, the transformation between the ability metric and number-right scores below the chance level is not defined. On the Science MCA-III, linear interpolation was employed to handle the scaling of number-correct scores below chance level. Boundary points for the interpolation were *x*—the lowest number-correct score *above* chance level—and 0, for a number-correct score of zero correct. The number-correct score *x* was assigned scale score *A*, using the transformation equation, and a number-correct score of zero was assigned the LOSS. For a number-correct score *y* between zero correct and *x*, scale scores were assigned using the following interpolation equation:

$$Scale(y) = LOSS + y * \frac{A - LOSS}{x} \tag{6.8}$$

For MCA-III, non-integer scale values are rounded to the nearest integer value. Because Mathematics and Reading MCA-III θ score estimates are constrained to fall within the range –3 to 3, in some grades the scores of g01 or g99 may not be attainable.

Minnesota Comprehensive Assessments-Series III Progress Score

A vertical or growth scale links tests in the same subject area across grade levels. With a vertical scale, the gain in knowledge from one year to the next can be measured for each student. An accurate measure of student growth is valuable information for users of test scores.

However, the creation of a vertical scale is quite a challenging psychometric enterprise. The main difficulty arises because procedures linking the scores of tests require that the tests to be linked measure the same constructs. It is reasonable to assume this year’s grade 3 form and next year’s grade 3 form measure the same constructs, as long as the tests are constructed to adhere strictly to formally stated test specifications. On the other hand, it may not be reasonable to assume the grade 3 form and the grade 8 form measure the same constructs. Although both tests measure student knowledge of the subject matter, the constructs taught at those two grade levels might be quite different.

Another complication is that linking tests taken by potentially different populations generally requires both populations to take common items. It may be unreasonable to administer the same items to grade 3 students and grade 8 students. Items that would challenge grade 8 students would be far too difficult for grade 3 students, and grade 3 material would be far too easy for grade 8 students. This problem can be mitigated to some degree by using common items in adjacent grades and linking grades in a step-wise fashion.

Beginning in 2012, a vertical scale is reported for Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) in grades 3–8. Beginning in 2014, a vertical scale is reported for the Reading MCA-III. This scale is called the Progress Score. The Progress Score scale is formed by linking across grades using common items in adjacent grades. Underlying the Progress Score scale is an IRT vertical scale. The IRT vertical scale allows a student's scores across time to be compared on the same scale and thus allows student performance on the MCA-III to be tracked as the student progresses from grade to grade. The actual linking process used to form the IRT vertical scale is described in Chapter 7, "Equating and Linking." The following describes how the IRT vertical scale score is obtained and how that score is transformed into the Progress Score to ease interpretation.

For Mathematics and Reading MCA-III the vertical IRT scale score is computed as the linear transformation of the student's post-equated ability estimate:

$$\theta_{\mathrm{VS}} = \mathrm{a} * \ \theta_{\mathrm{reverse}} + \mathrm{b}$$

The constants a and b of the transformation for each grade per subject are provided in Tables 6.1 and 6.2. The theta score $\theta_{\mathrm{reverse}}$ is calculated by reversing the theta to scale conversion using the rounded scale score as the input. That is,

$$\theta_{\mathrm{reverse}} = \ \theta_{\mathrm{cut2}} + \frac{\mathrm{SS-G50}}{\mathrm{Spread}}$$

where SS is the rounded scale score. For the students whose scale score is limited by either the G01 lowest observable scale score or G99 highest observable scale score, the scale score should be recalculated using the theta to scale score conversion to remove the G01 floor or G99 ceiling.

The vertical scale score (VSS) is calculated as $\mathrm{VSS} = 2500 + 100 * \ \theta_{\mathrm{VS}}$. After calculating the vertical scale score, the vertical scale score value is rounded to the nearest integer.

The SEM of the VSS is calculated as $\mathrm{VS}_{\mathrm{SEM}} = 100 * \mathrm{a} * \mathrm{SS}_{\mathrm{SEM}}/\mathrm{Spread}$, where $\mathrm{SS}_{\mathrm{SEM}}$ is the unrounded SEM of the on-grade scale score.

Tables 6.3 and 6.4 shows the minimum and maximum scores on the Progress Score for MCA-III.

Table 6.1: Mathematics MCA-III Vertical Scaling Constants on the Theta Metric

Grade	a	b	Minimum	Maximum
3	0.9960	–1.1584	–4.1165	1.8097
4	0.9700	–0.5665	–3.4590	2.3144
5	1	0	–2.9800	2.9840
6	1.0434	0.4986	–2.6441	3.6225
7	1.0924	0.9202	–2.3330	4.2389
8	1.2028	1.4085	–2.1879	5.0289

Table 6.2: Reading MCA-III Vertical Scaling Constants on the Theta Metric

Grade	a	b	Minimum	Maximum
3	0.962538	–1.073905	–3.9862	1.7884
4	1.019095	–0.563830	–3.6305	2.4793
5	1	0	–2.9785	2.9711
6	0.993229	0.383747	–2.5909	3.3918
7	1.010005	0.771721	–2.2655	3.8072
8	1.051537	1.013116	–2.1706	4.1417

Table 6.3: Mathematics MCA-III Progress Score Minima and Maxima

Grade	Minimum	Maximum
3	2088	2681
4	2154	2731
5	2202	2798
6	2236	2862
7	2267	2924
8	2281	3003

Table 6.4: Reading MCA-III Progress Score Minima and Maxima

Grade	Minimum	Maximum
3	2101	2679
4	2137	2748
5	2202	2797
6	2241	2839
7	2273	2881
8	2283	2914

On the student ISR, the Progress Score is given for grades 3–8 for each year that the student has taken the MCA-III, beginning with the inception of the Mathematics MCA-III in 2011 and the Reading MCA-III in 2013. For example, if a student took the Mathematics MCA-III as a third grader in 2011, as a fourth grader in 2012, and as a fifth grader in 2013, the ISR will report the student's score in each of those years. The progress score is given in both tabular and graphical form. The graph gives both the Progress Score for the student as well as the score on the Progress Score scale that represents *Meets the Standards*. The graph facilitates a comparison of the student's progress across years as well as depicting whether the student's performance in each year met the standards.

Minnesota Test of Academic Skills Scaling

The general transformation formula used to obtain scale scores for the Minnesota Test of Academic Skills (MTAS) is as follows:

$$Scale\ Score = \ (\theta \ - \ \theta_{Std2}) * Spread + Center \tag{6.9}$$

where θ is the post-equated ability estimate, θ_{Std2} is the ability cut score between *Partially Meets the Standards* and *Meets the Standards, Center* is set to be 200, and *Spread* is a numerical constant unique to each test by subject by grade combination. All grades and subjects of the MTAS use the same transformation equation.

Chapter 5, "Performance Standards," describes the process of setting the standards for the MTAS, a procedure culminating in the Commissioner of Education approving the cut scores. The ability cut scores corresponding to the Commissioner of Education approved raw score cuts were used to set the MTAS scales.

As with the MCA-III, it was desired to make the proficiency level scale score cuts as consistent as possible across grades. Using a linear transformation-like equation (6.9) allows two of the three scale cut scores to be fixed. For all grades and subjects of the MTAS, the cut score for *Meets the Standards* was set to 200 by setting *Center* = 200. The cut score between *Does Not Meet the Standards* and *Partially Meets the Standards* was desired to be equal to 190. Note that the 2007 MTAS value was 195, but beginning in 2008, the cut was changed to 190. It was felt that the increase in score points for the revised MTAS justified a corresponding increase in scale score values between the *Partially Meets* and the *Meets* scale score cuts. The *Spread* constant for each grade and subject combination of the MTAS was selected to force the first scale cut score to be equal to 190. The formula used to find the *Spread* is:

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})}, \tag{6.10}$$

where θ_{Std1} is the theta ability cut score between *Does Not Meet the Standards* and *Partially Meets the Standards,* and θ_{Std2} is the theta ability cut score between *Partially Meets the Standards* and *Meets the Standards*. The *Spread* value varies for each grade per subject combination. Because only two of the three scale cut scores can be predetermined using a linear transformation, the scale cut score between *Meets the Standards* and *Exceeds the Standards* was allowed to vary across grades and subjects.

ACCESS for ELLs Scaling

Scaling information about the ACCESS for ELLs is contained in the annual Technical Reports available on the WIDA website at http://www.wida.us/assessment/access/TechReports/index.aspx.

Subscores

The primary goal of each assessment is to provide an indicator of student progress in each subject area. Subject area achievement is reported as the total scale score and achievement level classification. Subject area test scores represent a sample of academic achievement from across a number of content strands. For example, the Mathematics assessments include indicators of achievement in geometry, algebra, number sense, measurement, and probability. It can therefore be useful to break out subject area test scores by content strand to provide more fine-grained analysis of student achievement. This is accomplished through subscale reporting.

The MTAS assessments report subscores as raw score (i.e., number correct) points. As with subject area scores, subscale scores reported as number-correct scores are not as meaningful because number correct ignores information about both the number and difficulty of test items. For example, scoring 15 out of 20 might indicate superior performance on a very difficult test but reflect poor performance on a very easy test. This difficulty is compounded when interpreting performance across subscales, because some subscales may include more easy items, while other subscales comprise more difficult items. For example, if items measuring number sense are easier than items measuring algebra, a higher number-correct score on number sense than algebra might appear to suggest greater achievement in number sense but in reality might indicate greater mastery of algebra than number sense.

To provide subscale scores that can be more meaningfully interpreted, MCA-III assessments report strand level performance on a common scale that reflects relative achievement across the student population. Scale scores are reported for the strands, based on a linear transformation of the estimated strand ability on the theta metric that places scores on a 1 to 9 scale. The linear transformation from the theta ability estimate to scale score for each subscale is:

$$\text{Subscale Score} = 5 + \text{Round}(2\theta)$$

with scores ranging from 1–9. The standard error of the subscale score is calculated as:

$$\text{Subscale SEM} = \text{Round}(2^{*}\text{SEM}(\theta))$$

with values truncated to a minimum of 1 and a maximum of 2. In 2011 and 2012, Mathematics MCA-III strand theta score estimates were obtained using MLE scoring. Beginning in 2013 for Mathematics and Reading MCA-III assessments, expected a posteriori (EAP) scoring is used to obtain theta estimates for strands. For Science MCA-III assessments, EAP sum scoring is used to estimate strand theta values.

Average student performance on each scale will be approximately 5 (in the initial year of the test), with a standard deviation of approximately 2. Assuming a normal score distribution, roughly one-half of students will score in the range of 4–6. Student scores above this range suggest above average performance, while scores below this range suggest below average performance. Thus, if a student scores above the 4–6 range on some strands, those might be considered areas of strength for the student, while scores below the range may indicate areas of weakness. Subscale scores are also reported with a confidence interval to indicate the degree of uncertainty in the student's subscale score. Subscale scores based on more test items will tend to have smaller confidence intervals, while subscale scores based on fewer items will have larger confidence intervals, which indicate a subscale measured with less precision.

Caution is always required when interpreting subscale scores. Because some subscale scores are based on few items, individual subscale scores may not be stable or consistent. As a consequence, differences in student performance across subscales may not be of practical importance. Thus, caution is required when interpreting differences between subscale scores for a student.

Scale Score Interpretations and Limitations

Minnesota Comprehensive Assessments-Series II and Series III

Since the on-grade scale scores associated with the Minnesota Comprehensive Assessments-Series III (MCA-III) are not on a vertical scale, great caution must be exercised in any interpretation of between-grade scale score differences within a subject area. Similar caution should be used in interpreting scale score differences between subject areas within a grade. Even though scale score ranges (g1–g99) and positions of two of the cut scores (g40 and g50) are consistent across grades and subjects, the scale score metrics cannot be presumed equivalent across subject or grade. As indicated by equations (6.6) and (6.9), the scale score difference associated with a theta score difference of 1.0 will depend upon the Spread parameter. As a consequence, scale score differences between two students of, for example, 10 points seen on tests from two subjects or grades can reflect theta score differences of varying size. In general, achievement levels are the best indicators for comparison across grade or subject. The scale scores can be used to direct students needing remediation (that is, students falling below *Meets the Standards*), but scale score gain comparisons between individual students are not appropriate. Progress Scores and the MCA-III vertical scale score, which are based on vertical scales, are intended to provide an appropriate basis for making comparisons across years within a subject area.

For assessments that use raw to scale score conversions to determine scale scores (i.e., MTAS and Science MCA-III), users should be cautioned against over-interpreting differences in scale scores in raw score terms because scale scores and number correct scores are on two distinct score metrics that have a decidedly nonlinear relationship. As a hypothetical example, students near the middle of the scale score distribution might change their scale score values by only 4 points (for example, from 548 to 552) by answering five additional multiple-choice questions correctly. However, students near the top of the scale score distribution may increase their scale score by 20 points with five additional questions answered correctly (for example, from 570 to 590). A similar phenomenon may be observed near the bottom of the score scale. In the case of Mathematics and Reading MCA-III, which utilize pattern scoring and have multiple fixed forms or are administered adaptively, attempts to interpret scale scores in raw score terms are generally inappropriate.

The primary function of the scale score is to be able to determine how far students are from the various proficiency levels without depending upon the changing raw scores. Additionally, schools may use the scale scores in summary fashion for purposes of program evaluation across the years. For example, it is appropriate to compare the average grade 5 scale score in reading for this year to the grade 5 average for last year (if the test series has not changed). Explanations for why the differences exist will depend on factors specific to individual schools.

Finally, it must be stressed that there are substantial differences in test content and scoring metrics between the MCA-III and the MCA-II. These differences should discourage attempts to draw inferences based on score comparisons between students now taking the MCA-III tests in a subject from those who took the MCA-II in past years. Thus, for example, it is not appropriate to compare the grade 5 Reading MCA-III score from 2013 to the grade 5 Reading MCA-II score average from previous years. However, limited and focused linking procedures or prediction analyses may still serve useful purposes.

Conversion Tables, Frequency Distributions, and Descriptive Statistics

The Yearbooks provide tables for converting raw scores to derived scale scores (when applicable) and tables of frequency distributions and summary statistics for scale scores by grade and subject.

Chapter 7: Equating and Linking

Equating and linking are procedures that allow test scores to be compared across years. The procedures are generally thought of as statistical processes applied to the results of a test. Yet, successful equating and linking require attention to comparability throughout the test construction process. This chapter provides some insight into these procedures as they are applied to Minnesota Assessments.

Rationale

In order to maintain the same performance standards across different administrations of a particular test, it is necessary for every administration of the test to be of comparable difficulty. Comparable difficulty should be maintained from administration to administration at the total test level and, as much as possible, at the subscore level. Maintaining test form difficulty across administrations is achieved through a statistical procedure called equating. Equating is used to transform the scores of one administration of a test to the same scale as the scores of a second administration of the test. Although equating is often thought of as a purely statistical process, a prerequisite for successful equating of test forms is that the forms are built to the same content and psychometric specifications. Without strict adherence to test specifications, the constructs measured by different forms of a test may not be the same, thus compromising comparisons of scores across test administrations.

For the Minnesota Assessments, a two-stage statistical process with pre- and post-equating stages has generally been used to maintain comparable difficulty across administrations. This equating design is commonly used in state testing. In the pre-equating stage, item-parameter estimates from prior administrations (either field test or operational) are used to construct a form with a difficulty level similar to that of previous administrations. This is possible because of the embedded field-test design that allows for the linking of the field-test items to the operational form.

In the post-equating stage, all items are recalibrated, and the test is equated to prior forms through embedded linking items. Linking items are items that have previously been operational test items and whose parameters have been equated to the base-year operational test metric. The performance of the linking items is examined for inconsistency with their previous results. If some linking items are found to behave differently, appropriate adjustments are made in the equating process before scale scores are computed.

The Minnesota Department of Education (MDE) strives to use the pre- and post-equating design for all applicable testing programs to ensure the established level for any performance standard on the original test is maintained on all subsequent test forms. The pre- and post-equating design is fully described in the sections that follow. This two-stage equating approach was used with all MCA-II assessments, and also with large scale paper form administrations of the Reading and Mathematics MCA-III. The online Minnesota Comprehensive Assessments-Series III (MCA-III) assessments are the exception to the pre- and post-equating design. The Mathematics, Reading, and Science online MCA-III Assessments are pre-equated in order to permit immediate reporting of score results. Maintenance of performance standards on the online MCA-III is accomplished through careful review of item performance after the close of the test administration window.

In some cases, it may be desired to compare the scores of tests that have been built to different specifications. For example, one may want to compare the reading scores of a group of grade 4 students to their scores on the previous year’s grade 3 reading test. The tests at each grade are designed to measure the specific content expected to be mastered in that grade; consequently, the tests measure different constructs and are built to different specifications. A transformation can be made to place two different forms or tests on the same scale, but when the forms or tests in question are built to different specifications, the process is called linking. The term "linking" is used in place of equating to emphasize the more tenuous relationship created between scores on different tests. Although equating and linking create a relationship between different forms or tests, the strength or quality of the relationship depends on the degree to which the forms or tests measure the same constructs. Discussions on linking are given in Mislevy (1992), Linn (1993) and Kolen and Brennan (2004). The "Linking" section of this chapter describes the Minnesota assessments that are associated through a linking process.

Pre-Equating

The intent of pre-equating is to produce a test that is psychometrically equivalent to those used in prior years. The pre-equating process relies on links (specifically, equated item parameter estimates) between each item on a newly developed test to one or more previously used test forms. In this way, the difficulty level (and other psychometric properties) of the newly developed test can be equated to previously administered tests. For the Minnesota Comprehensive Assessments-Series III (MCA-III), each new assessment is constructed from a pool of items whose parameters have been equated to the base test form (2011 for grades 3–8 mathematics, 2012 for science, and 2013 for reading). Note that 2014 is the base year for grade 11 Mathematics MCA-III.

Test Construction and Review

Test construction begins by selecting the operational (or base) items for an administration. For the MCA-III Science these items are given on every test form for that administration, and they count toward the individual student's score. For the MCA-III Math grade 11 and MCA-III Reading grades 3–8 and 10, there are multiple fixed forms, with different operational items appearing on the forms. Using the items available in the item pool, psychometricians from Minnesota's testing contractor construct new forms by selecting items that meet the content specifications of the subject tested and targeted psychometric properties. Psychometric properties targeted include test difficulty, precision, and reliability. The construction process is an iterative one, involving Minnesota's testing contractor and Minnesota Department of Education (MDE) staff. Since the item response theory (IRT) item parameters for each item in the item bank are on the same scale as the base scale test forms, direct comparisons of test characteristic functions and test information functions can be made in order to ascertain whether the test has similar psychometric properties (e.g., difficulty) to those of the original form.

The newly constructed test is reviewed by psychometricians and content staff to ensure specifications and difficulty levels have been maintained. Although every item on the test has been previously scrutinized by Minnesota educators and curriculum experts for alignment to benchmarks—a match to test specifications' content limits, grade-level appropriateness, developmental appropriateness, and bias—MDE re-examines these factors for each item on the new test. The difficulty level of the new test form—for the entire test and for each objective—is also evaluated, and items are further examined for their statistical quality, range of difficulties, and spread of information. Staff members also review forms to ensure a wide variety of content and situations are represented in the test items, to verify that the test measures a broad sampling of student skills within the content standards, and to minimize "cueing" of an answer based on the content of another item appearing in the test. Additional reviews are designed to verify that keyed answer choices are the only correct answer to an item and that the order of answer choices on the test form varies appropriately.

If any of these procedures uncovers an unsatisfactory item, the item is replaced with a new item and the review process begins again. This process for reviewing each newly constructed test form helps ensure each test will be of the highest possible quality.

Simulations for Adaptive Assessments

The nature of an adaptive test is to construct a test form unique to each student and targeted to the student's level of ability. As a consequence, the test forms will not be statistically parallel—nor should they be. However, scores from the assessment should be comparable, and each test form should measure the same content, albeit with a different set of items.

The adaptive algorithm and a complex blueprint have dozens of adjustable parameters. Examples include balancing the weight given to one strand versus other strands or item type constraints. The optimal values for the algorithm parameters vary depending on the item pool, specifics of the blueprints, and their interaction. Prior to each operational testing window, the testing vendor conducts simulations to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm. Simulations enable key blueprint and configuration parameters to be manipulated to match the blueprint, minimize measurement error and control item exposure.

Simulations begin by generating a sample of examinees from a Normal (μ, σ) ability (theta) distribution for each grade. The parameters for the normal distribution are taken from the previous year's operational administration. Each simulated examinee is then administered a test under the adaptive algorithm. Once simulations are complete a variety of statistical measures are then examined. First, the percentage of simulated test forms that met test specifications is calculated. When all students are administered a test that meets blueprint, the content can be considered equivalent across students. Second, the bias and average standard error of the estimated ability is calculated, as well as the distribution of errors across the true score theta range. When true test scores are adequately recovered the mean of the bias is small and statistically insignificant. If summaries show a failure to meet blueprint specifications or unacceptable levels of error in student ability estimation, algorithm parameters are revised and simulations are rerun. This process continues until requirements are met.

Field-Test Items

Once a newly constructed item has survived committee review and is ready for field-testing, it is embedded in a test form among the operational test items. For example, in a particular grade's Science MCA-III administration there might be 15 different forms containing the same operational test items. However, each form would also contain one or more unique field-test scenarios and corresponding unique field-test items. The field-test items do not count toward an individual student's score. They may be used as equating or linking items to past or future tests, but for the MCA-III the role of linking is usually reserved for items that have been administered operationally in a previous year.

Paper forms are spiraled within testing sites (usually classrooms) across the state so a large representative sample of test takers would respond to each of the field-test items. In online administrations of fixed forms, forms are assigned randomly to students. For example, at grade 10, with a statewide enrollment of approximately 65,000, approximately 4,300 students would respond to each of 15 forms. In online adaptive tests, field test items are assigned at random to students in designated slots during the administration. This spiraling design provides a diverse sample of student performance on each field-test item. In addition, because students do not know which items are field-test items and which items are operational test items, no differential motivation effects are expected. To control for fatigue and start-up effects, all field-test items are placed in similar positions on each test form.

Post-Equating

Item Sampling for Equating

To ensure a successful equating or linking of forms or tests, it is necessary that there exist a solid statistical link between the forms or tests. Typically, this means two forms or tests being equated or linked must have a set of items in common. It is important that the set of linking items be representative of the construct being measured by the tests and have the same approximate difficulty and spread of information as the tests that are being linked.

Before the development of the Minnesota Comprehensive Assessments-Series II (MCA-II), the administrations of the MCA were linked by associating the test's multiple-choice operational item-parameter estimates with estimates for those same items when they were given as field-test items. Operational items typically had been field-tested in the previous year, providing a link for the two administrations. Although this approach results in a fairly large number of linking items (all the multiple-choice operational items), it suffers from the relative instability of field-test item-parameter estimates. Most items are field-tested to a sample of students much smaller than the total number of students who take the test. Consequently, using field-test item-parameter estimates as part of the link between administrations can add errors to the equating process.

With the deployment of the MCA-II and then the MCA-III, a new system of linking items was devised that did not rely on field-test item-parameter estimates. Linking administrations to the base year is achieved by using "internal" and "external" linking items. Internal linking items are multiple-choice or figural-response items that were operational (i.e., counted toward student scores) in a previous administration and also are operational in the current administration. External linking items are multiple-choice or figural-response items that may have been operational in a previous administration, but in the current administration they are given to a random sample of the population (they are placed on a single form as if they were field-test items). Internal linking items count toward a student's score, just as any other operational item. External linking items, however, do not count toward a student's score for the current administration. For each MCA-III administration with post-equating, there are at least eight internal linking items and eight to sixteen external linking items. Linking items are chosen so that the set of linking items gives good coverage of the benchmarks and approximates the overall difficulty and information spread of the operational items. With the transition of MCA-III assessments in all subjects to online administration (except for accommodated paper forms) with immediate score reporting, post-equating of operational MCA-III items is no longer employed prior to scoring.

Some administrations of the Mathematics and Reading MCA-III in grades 3–8 have included off-grade vertical linking items. These are multiple-choice items from an adjacent grade's test that are used to link grades on the Progress Score. This occurred in 2011 for Mathematics MCA-III and in 2013 for Reading MCA-III. Off-grade vertical linking items do not count toward a student's score. The tests for grades 10 and 11 do not use vertical linking items, since no Progress Score is reported for these grades.

Student Sampling for Equating

Because almost all the population for a grade and subject is used for the operational test equating, no sampling procedures are used. Some student data, however, are excluded from the post-equating calibration of items. If the number of items a student attempts does not meet the minimum attemptedness criterion, then data from that student are excluded from the calibration data set. For the MCA-III and other online assessments, students must respond to at least 90% of the questions on the entire test in order to be classified as "attempted." In addition, the responses of home-schooled and private school students are excluded from the calibration data set. Home-schooled and private school students are not required to take the MCA-III and are not to be included in statewide summary statistics or in No Child Left Behind (NCLB) calculations. Their test scores, however, are reported to students, parents, and schools, similar to students at public schools.

Operational Item Post-Equating Procedures

Once the statewide data file has been edited for exclusions, a statistical review of all operational items is conducted before beginning IRT calibration. Items are evaluated for printing or processing problems. A multiple-choice or technology enhanced item is flagged for further review if it has a low mean score, a low item-total correlation, an unusually attractive incorrect option, or a mean score on any one form that differs substantially from all the other forms. Constructed-response items are flagged for unusual score distributions. Any flagged items are reviewed in the published test books to ensure that the item was correctly printed. Flagged items also have keys checked by Minnesota's testing contractor and Minnesota Department of Education (MDE) content staff to certify the key is the correct answer.

For all MCA-III assessments, the commercial software IRTPRO is used for item calibrations. For reading, all MC items are calibrated as three parameter items, and non-MC (technology enhanced) items are calibrated as two parameter items. For mathematics, all "Fill in the Blank" (FIB) items are calibrated as two parameter items; all other items (including all technology enhanced items) are calibrated as three parameter items. These models are described in the "Scaling" section in Chapter 6.

All operational items for a test (regardless of item type) and external linking items are calibrated simultaneously. After obtaining the linking item-parameter estimates on the current administration's operational scale, equating is performed in order to place the current operational scale on the base-year scale. Scaling constants used to transform the current year scale to the base-year scale are obtained by using the Stocking-Lord procedure (Stocking & Lord, 1983).

Once the linking items have been equated to the original scale, a comparison of the item response functions is made to determine whether the linking items are functionally the same across the two administrations. Substantial deviations in the item response functions of an item indicate that students responded differently to the linking item as it appeared in the current form than did students who took the item in a previous operational administration. This could occur, for example, if the sequence order of the linking item is substantially different on the two forms. If the item response function is substantially different for the two administrations, a decision may be made to discard the item from the linking set. The scaling process is then continued with the reduced linking set.

Once a satisfactory linking item set and transformation equation have been determined, the same constants used to transform the linking items to the base scale are applied to all the operational items of the current administration. With the current administration equated, student scores can be placed on the reporting metric, as described in the "Scaling" section in Chapter 6.

Field-Test Item Equating Procedures

Field-test items are brought onto the same scale as the operational test items through post-equating. For MCA-III assessments, all of the operational items that did not show item parameter drift are used as anchor items. To determine the final set of anchor items, drift analysis is conducted to flag items that may have moved from their bank parameters (on the base scale). The purpose of the drift analyses is to identify items that may have shifted in difficulty relative to the bank as a whole. This can occur due to changing emphases in instruction throughout the state from year to year, exposure of the item, or for a host of other reasons. Because that sample sizes are too small on the accommodated paper forms for these analyses to be effective, they are only conducted on the online forms. The drifted items identified through the statistical procedures are discussed between the Minnesota Department of Education, the test vendor, and the Human Resources Research Organization. Upon agreement, a final list of drifted items is excluded from the equating anchor set for each subject and grade. The remaining anchor items are used to establish the equating relationship and bring field-test items on the operational item scale. For post-equated MCA-III non-accommodated paper forms, the Stocking-Lord procedure is used. First, non-drifted operational items and all the field-test items are calibrated together. Secondly, the difference between the bank parameters and the parameter estimates obtained in the current calibration for the anchor items is used to establish the equating relationship, and the resulting constants are applied to the field test items. For the computer-adaptive Mathematics MCA-III in grades 3-8, non-drifted operational items and field-test items are calibrated together in IRTPRO with the non-drifted operational items being anchored to their bank parameters and field-test items being estimated. By anchoring the non-drifted operational items to their reference value, the field-test items are automatically placed on the reference scale. For fixed-form online MCA-III assessments (in reading, science and grade 11 mathematics), field-test item calibration has alternated between use of the Stocking-Lord and fixed operational-item parameter approaches, depending on test vendor, with the Stocking-Lord approach being applied subsequent to the spring 2015 administration.

MTAS Equating

The commercial software package WINSTEPS (Linacre, 2006) is used for the Minnesota Test of Academic Skills (MTAS) performance tasks. As described in Chapter 6, “Scaling,” the IRT model used for calibration is the Rasch Partial Credit model (Masters, 1982). For some MTAS administrations, a combined operational and field-test design is employed. After item or task calibration, MDE staff selects the nine tasks at each grade level to be designated as operational. The base year for grade 11 Mathematics is 2014. The base year for MTAS Science is 2012. For grades 3–8 of Mathematics MTAS, the base year is 2011. For grades 3–8 and 10 of Reading MTAS, the base year is 2013.

Equating to the base year is accomplished using conceptually similar procedures to those used with the Mathematics MCA-III grades 3–8. For MTAS, a simultaneous calibration of operational and field-test tasks is performed by grade and subject. The fit of field-test tasks to the model is scrutinized in order to ensure that a poor-fitting field-test task does not compromise the calibration of the operational tasks. In addition, linearity is checked by plotting linking task IRT difficulty values against those from the base year. Linking tasks are then equated back to the base scale by subtracting the mean of the new IRT difficulty values from the mean of the base-year difficulty values (mean/mean equating). The difference of means is then added to the IRT difficulty values of the linking tasks. The equated IRT parameters are then compared with the base-year values. Differences between equated and base-year values are called displacement values. Displacement values are scrutinized, and tasks with displacements greater than 0.3 may be dropped from the equating. After dropping any linking task that fails the stability check, another

WINSTEPS calibration is performed for all tasks with linking task parameters fixed to their base-year values. The task parameter values from the second calibration are considered the final parameter values for purposes of scale score calculation and item banking.

Development Procedure for Future Forms

Item Pool Maintenance

The next step is to update the item pool with the new statistical information. The post-equated item-parameter estimates for the operational test items are added to the item pool database, as are the item statistics and parameter estimates for the field-test items. In this way, the item pool contains the parameter values from the most recent administration in which the item appeared.

When assessments are scored using pre-equated item parameters, there is no post-administration calibration and equating of operational items. However, items must still be examined for signs of misfit or drift. Following the spring 2014 administration of the MCA-III Mathematics grades 3–8 online, the MCA-III Reading grades 3–8 and 10 online, and MCA-III Science assessments, items were evaluated for parameter drift. The general approach to evaluating goodness-of-fit involves the comparison between observed and model-predicted frequencies for various ability (theta) subgroups using chi-square fit statistics. The item-fit statistics employ a pseudo-observed theta distribution as proposed in Stone (2000). Items with large fit statistics are flagged. Additional graphical analyses were conducted to compare observed and expected item response probabilities in order to identify item misfit and drift. Flagged items may be recalibrated, moved to the Optional Local Purpose Assessment (OLPA) pool, or simply removed from the administration pool.

Linking

When scores are compared between tests that have not been built to the same test specifications, the process of finding the score transformation is called linking. Whereas equating can be used to maintain comparable difficulty and performance standards across administrations of the Minnesota Comprehensive Assessments Series-III (MCA-III), linking is used for two purposes: (1) scaling across grades with the Progress Score (reading and mathematics) and (2) linking the Reading MCA-III to the Lexile reading scale.

Linking Grades 3–8 with the Progress Score

The Mathematics and Reading MCA-III each use a vertical scale called the Progress Score. Vertical scales, such as the Progress Score, are designed to help evaluate how much students improved from one year to the next. Use of vertical scales can assist educators and parents in obtaining information about absolute levels of student growth. Linking for the Progress Score was accomplished by using common items on adjacent grades on the 2011 Mathematics MCA-III administration and the 2013 Reading MCA-III administration. Off-grade items did not count toward a student’s final score. The linking design was such that no student took both upper-grade items and lower-grade items. For example, some fourth-grade students took a linking set of third-grade items, and some fourth graders took a linking set of fifth-grade items. The determination of which students took the linking sets was done by random assignment, in the same manner as the procedure for spiraling field-test items.

After calibration of the operational items was complete, a separate calibration that included the off-grade items was conducted for each grade. The operational items served as linking items to scale the off-grade items to the 2011 operational scale for the Mathematics MCA-III or the 2013 operational scale for the Reading MCA-III. After off-grade items were scaled for each grade, another scaling process was conducted to place the items of grades 3–8 on the fifth-grade scale, which served as the reference scale for vertical scale. IRT linking was conducted sequentially, moving away from the fifth-grade scale. That is, to place the third-grade items on the vertical scale, first the fourth grade items were linked to the fifth-grade scale, and then the third-grade items were linked to the rescaled fourth-grade scale. Likewise, for the upper grades, the sixth-grade items were linked to the fifth-grade items, and then the seventh-grade items were linked to the rescaled sixth-grade items, and finally, the eighth-grade items were linked to the rescaled seventh-grade items. Future administrations of the MCA-III may contain vertical linking items for purposes of maintaining the scale of the Progress Score.

Linking Reading MCA-III to the Lexile® Scale

MetaMetrics typically uses a common-person design to develop linkages between statewide assessments and their proprietary Lexile scale. For example, to link the previous MCA-II Reading scale to the Lexile scale, MetaMetrics administered a stand-alone test to students in a sample of districts following regular administration of the MCA-II Reading assessment in 2010. There are, however, significant disadvantages to the use of the common-person design employed at the time to establish the initial linkage. The independent assessment, although administered nearly concurrently with the accountability assessment, is entirely voluntary and carries no stakes for students or schools. Consequently, motivation for high performance may be diminished. It may be that motivation effects are more pronounced for older students, especially grade 10 students. Younger students may not readily make distinctions between high stakes and low stakes testing situations, and treat the assessments in the same manner. Perhaps more important, this testing design places a substantial additional assessment burden on participating schools and students (as well as increased burden on MDE for recruiting sampled schools), requiring approximately the same amount of testing time as employed for the MCA Reading assessment.

For the MCA-III Reading assessments, MetaMetrics agreed to allow MDE to embed Lexile items in field-test slots of the spring 2013 accountability test administration. Embedding Lexile items in the initial administration of the MCA-III Reading assessment allowed MDE to administer Lexile items under operational testing conditions and confine the burden of field-testing to the standard administration of the accountability assessment, eliminating much of the cost and burden of a stand-alone field test.

Embedded field-test blocks in the Reading assessment were designed to accommodate a Reading passage and associated test items. Lexile items are, by contrast, discrete items that are not passage based. To embed the Lexile items within the MCA-III Reading assessments, the test vendor defined a set of Lexile item blocks for administration in field-test slots, based on the linking sets provided by MetaMetrics. Lexile item blocks were constructed so that test administration times are similar to that required to read a passage and answer associated test items for a typical passage-based set of items. This resulted in Lexile blocks comprising 12 items each. MetaMetrics provided a linking set of 36 items at each grade level that are required to link the MCA-III Reading scale to the Lexile scale, which resulted in administration of three Lexile blocks at each grade level. With Lexile items administered alongside MCA-III Reading passages and items, the items can be calibrated concurrently, allowing all items to be placed on a common scale.

Although there was a sufficient number of linking items to place the MCA-III Reading items on the Lexile scale, individual students were not administered a sufficient number of Lexile items to produce a reliable, independent assessment of Reading ability based solely on those items. MDE therefore used the embedded Lexile items to identify Rasch parameter estimates for MCA-III items linked to the Lexile scale. The result of this Lexile linking was two sets of parameter estimates for each item: a set of 3PL parameter estimates on the MCA-III scale and a set of Rasch parameter estimates on the Lexile scale. The two sets of parameter estimates were used to produce two ability estimates for each student, one on the MCA-III scale and a second on the Lexile scale. Note that both ability estimates were based on the same set of MCA-III operational test items. With the two ability estimates in hand, a mean-sigma linking was employed in order to place MCA-III scores on the Lexile scale. Because item-parameter estimates for the MCA-III Reading assessments are based on the 3PL IRT model, while the Lexile items parameters are estimated using the Rasch model, linking the MCA-III Reading and Lexile scales was accomplished via student ability estimates obtained from the respective models.

Before any linking was performed, an initial analysis was completed with MetaMetrics on only the Lexile items in order to analyze their performance. It was determined that one item administered in grades 8 and 10 was not performing consistently with past experience, and MetaMetrics recommended that it be dropped from further analyses. Thus, at grades 8 and 10, 35 Lexile items were used for calibration. Lexile items were then anchored to their reference scale values, and MCA-III bank items were calibrated using the 1PL model. Student ability estimates were found using the resulting Lexile-scale MCA-III item parameters. Using the same set of items and responses, student ability estimates were found using the MCA-III scale item parameters.

After examining the ability distributions, a single mean-sigma transformation in each grade was used to put the MCA-III scale ability estimates on the Lexile scale. The Lexile-scale ability estimates were then multiplied by the Lexile measure reporting constant in order to obtain the Lexile research measure. Since Lexile scores are reported in values ending in 0 or 5, the Lexile research measures were then rounded to their reported Lexile Measure. Student Lexile measures were reported with a +/–100 Lexile measure upper and lower bounds.

Chapter 8: Reliability

Reliability is the consistency of the results obtained from a measurement. When a score is reported for a student, there is an expectation that if the student had instead taken a different but equivalent version of the test, a similar score would have been achieved. A test that does not meet this expectation (that is, a test that does not measure student ability and knowledge consistently) has little or no value. Furthermore, the ability to measure consistently is a prerequisite to making appropriate interpretations of scores on the measure (i.e., showing evidence of valid use of the results). However, a reliable test score is not necessarily a valid one. And a reliable, valid test score is not valid for every purpose. A measure can be consistent and support certain score interpretations but still not support all the inferences a user of the test wishes to make. The concept of test validity is discussed in Chapter 9, "Validity."

A Mathematical Definition of Reliability

The basis for developing a mathematical definition of reliability can be found by examining the fundamental principle at the heart of classical test theory: all measures consist of an accurate or "true" part and some inaccurate or "error" component. This axiom is commonly written as:

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

Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical world and changes in examinee disposition may work to increase error and decrease reliability. In classical test theory, error is typically assumed to be the result of random, unsystematic influences. If there are systematic influences contributing to the error term, then derived reliability indices are likely to be compromised. For example, if a test is administered under very poor lighting conditions, the results of the test are likely to be biased against the entire group of students taking the test under the adverse conditions.

From equation (8.1), it is apparent that scores from a reliable test generally have little error and vary primarily because of true score differences. One way to operationalize reliability is to define reliability as the proportion of true score variance relative to observed score variance: the variance of the students' true scores divided by the variance of their observed scores (see equation (8.2)).

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2} \tag{8.2}$$

where σ_T^2 is the true score variance, σ_O^2 is the variance of the observed score and σ_E^2 is the error variance. When there is no error, the reliability is the true score variance divided by true score variance, which is unity. However, as more error influences the measure, the error component in the denominator of the ratio increases and the reliability decreases.

Using assumptions from classical test theory (equation (8.1) and random error assumptions), an alternative formulation can be derived. Reliability, the ratio of true variance to observed variance, can be shown to equal the correlation coefficient between observed scores on two *parallel* tests. The term "parallel" has a specific meaning: the two tests meet the standard classical test theory assumptions, as well as yield equivalent true scores and error variances. The proportion of true variance formulation and the parallel test correlation formulation can be used to derive sample reliability estimates.

Estimating Reliability

There are a number of different approaches taken to estimate reliability of test scores. Discussed below are test-retest, alternate forms and internal consistency methods.

Test-Retest Reliability Estimation

Reliability can be estimated by calculating the correlation coefficient between scores from a test given on one occasion with scores from the same test given on another occasion to the same students. Essentially, the test is acting as its own parallel form. Using the test-retest reliability method has potential pitfalls. A long interval between testing sessions likely will result in student growth in knowledge of the subject matter, while a short interval increases the chance students will remember and repeat answers from the first session. In addition, the test-retest approach requires the same students to take a test twice. For these reasons, test-retest reliability estimation is not used on Minnesota assessments.

Alternate Forms Reliability Estimation

Alternate forms reliability is similar to test-retest, except that instead of repeating the identical test, two presumably equivalent forms of the test are administered to the same students. The accuracy of the alternate forms coefficient greatly depends upon the degree to which the two forms are equivalent. Ideally, the forms would be parallel in the sense given previously. For Minnesota assessments, alternate forms reliability estimation is not possible because no student takes more than one form of the test during any test administration.

Internal Consistency Reliability Estimation

Internal consistency methods use a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures requiring multiple tests. Probably the most frequently used internal consistency reliability estimate is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the essentially tau-equivalent measurement model and the formula is:

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} S_{Y_i}^2}{S_X^2}\right) \tag{8.3}$$

where *N* is the number of items on the test, $\mathrm{S}_{\mathrm{Y}_\mathrm{i}}^2$ is the sample variance of the ith item (or component) and S_X^2 is the observed score sample variance for the test. Coefficient alpha is a point estimate of score reliability, and it may be important to consider the precision of that estimate, particularly when it is based on a small number of data points and/or restriction of range. A confidence interval on coefficient alpha (Feldt, Woodruff, & Salih, 1987) may be constructed to support considerations of the precision associated with reliability estimates.

Coefficient alpha is appropriate for use when the items on the test are reasonably homogenous. Evidence for the homogeneity of Minnesota tests is obtained through a dimensionality analysis. Dimensionality analysis results are discussed in Chapter 9, “Validity.” Additionally, alpha is based on the total sample of test takers that take the same items. However, for the MCA-III Mathematics grades 3–8 online adaptive and grade 11 multiple fixed form and MCA-III Reading online multiple fixed form assessments, not all students see the same set of items, and standard measures of reliability are not appropriate. Instead estimates of reliability based on item response theory (IRT) are given. IRT provides a means of estimating reliability that operates on both the individual pattern of responses to items given by examinees and statistical characteristics associated with those items. The IRT analogue to classical reliability is called marginal reliability and is calculated using the variance of the theta (ability) scores and the average of the expected error variance. Similar to the decomposition of an observed score in classical test theory, one can decompose the estimated item response theory ability into the true ability plus error,

$$\hat{\theta} = \theta + \epsilon,$$

where θ is the true ability and ϵ is the error associated with the estimate. The reliability can then be expressed as

$$R_\theta = \frac{var(\theta)}{var(\hat{\theta})} = \frac{var(\hat{\theta}) - var(\epsilon)}{var(\hat{\theta})}.$$

The marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984; Thissen & Wainer, 2001) of the reported scale score can then be expressed as

$$Marginal\ Reliability = \frac{\sigma_{theta}^2 - \overline{\mathrm{SE}_{\mathrm{theta}_i}{}^2},}{\sigma_{theta}^2}$$

where $\sigma_{\mathrm{theta}}^2$ is the variance of ability scores for the population of interest and $\mathrm{SE}_{\mathrm{theta}_i}$is the standard error of the ability estimate of student i. Marginal reliability can be calculated by subtracting the average of the squared CSEM (error variance) for each student from the estimated variance of IRT ability scores and dividing by the estimated variance of IRT ability scores. In the case of MCA-III strand and substrand scores, where expected a posteriori (EAP) methods are used to estimate scores, an alternative formula, described by Bock and Mislevy (1982) is used to estimate score reliability that is based on the assumption the ability distribution is distributed *N*(0,1).

$$\mathrm{EAP}\ Marginal\ Reliability = 1 - \overline{PSD^2},$$

where PSD is the posterior standard deviation of the EAP estimate and

$$\mathrm{PSD} = \mathrm{Var}(\theta|\mathbf{u}) = \frac{\sum_{k=1}^{q}(\mathrm{X}_k - \theta)^2\, L(\mathrm{X}_k)W(\mathrm{X}_k)}{L(\mathrm{X}_k)W(\mathrm{X}_k)}$$

where X_k is one of q quadrature points, $W(\mathrm{X}_k)$ is a weight associated with the quadrature point, and $L(\mathrm{X}_k)$ is the likelihood function conditioned at that quadrature point.

For the MCA-III assessments, the marginal reliability is given for the overall scale score. For these assessments, standardized integer scale scores are reported for the strands, based on a linear transformation of the estimated strand theta (= 5.0 + 2θest), in place of raw scores. For strand scores the marginal reliability is calculated for the estimated theta score. This result is multiplied by the square of the correlation between strand theta estimate and the reported standardized scale score in order to reflect the impact resulting from transformation of the theta score to integer scale score values constrained to a 1-9 range.

Subscore reliability will generally be lower than total score reliability because reliability is influenced by the number of items (as well as their covariation). In some cases, the number of items associated with a subscore is small (10 or fewer). Results involving subscores (and subscore differences in particular) must be interpreted carefully, because these measures have lower reliability associated with them compared to total scores.

Standard Error of Measurement

A reliability coefficient expresses test score consistency in terms of variance ratios. In contrast, the standard error of measurement (SEM) expresses score inconsistency (unreliability) in terms of the reported score metric. The SEM is an estimate of how much error there is likely to be in an individual's observed score, or alternately, how much score variation would be expected if the individual were tested multiple times with equivalent forms of the test. The standard error of measurement is calculated using the following formula:

$$SEM = s_x\sqrt{1-\rho_{xx}}, \tag{8.4}$$

where s_x is the observed score standard deviation for the total test, and ρ_{xx} is the reliability estimate for the set of test scores.

Use of the Standard Error of Measurement

The standard error of measurement (SEM) is used to quantify the precision of a test in the metric on which scores will be reported. The SEM can be helpful for quantifying the extent of errors occurring on a test. A standard error of measurement band placed around the student's scale score would result in a range of values most likely to contain a student's observed score upon replication. For example, if a student has an observed scale score of 350 on a test having score reliability of 0.88 and a standard deviation of 12.1, the SEM would be

$$SEM = 12.1\sqrt{(1-0.88)} = 4.19 \tag{8.5}$$

Placing a one-SEM band around this scale score would result in a score range of 346 to 554 (that is, 350 ± 4.0). Furthermore, in the case of unbiased scores and if measurement error is normally distributed, then the *true scores* for approximately 68% of test takers would fall in the interval band created by adding and subtracting one SEM from their reported score. Thus, the chances are better than 2 out of 3 those students with an observed score of 350 and SEM = 4 would have an estimated true score within the interval 346–354. This interval is called a confidence interval or confidence band. By increasing the range of the confidence interval, one improves the likelihood the confidence interval includes the observed score. For example, an interval of ± 2 SEMs around the scale score is referred to as a 95% confidence interval. It should be noted that the above interpretations of likelihood of having a score within a range is only approximate because the confidence interval is constructed around a point estimate and does not have an associated direct probability statement. While it is common practice to use a frequentist confidence band around the observed score and treat such as a probability statement, only Bayesian methods allow for such an interpretation because the score has a probability distribution due to the use of a the prior distribution. Here, a score based on a Bayesian procedure (such as done with EAP-based strand scores) would have what is denoted as a posterior distribution (e.g., a set of plausible test scores) and *credible intervals* that represent direct probability statements about a score given the observed data.

The overall SEM for each test can be calculated with data provided for in the Yearbooks. However, given the use of IRT for all Minnesota's assessments, the conditional SEM (discussed in the next section) is the primary reporting measure of precision associated with each scale score.

Conditional Standard Error of Measurement

Although the overall SEM is a useful summary indicator of a test's precision, the measurement error on most assessments varies across the score range. This means the measurement accuracy of a test is likely to differ for students depending on their score. To formalize this notion, classical test theory postulates that every student has a true score. This is the score the student would receive on the test if no error were present. The standard error of measurement for a particular true score is defined as the standard deviation of the observed scores of students with that true score. This standard deviation is called the conditional standard error of measurement (CSEM). The reasoning behind the CSEM is as follows: if a group of students all have the same true score, then a measure without error would assign these students the same score (the true score). Any differences in the scores of these students must be due to measurement error. The conditional standard deviation defines the amount of error.

True scores are not observable. Therefore, the CSEM cannot be calculated simply by grouping students by their true score and computing the conditional standard deviation. However, item response theory (IRT) allows for the CSEM to be estimated for any test where the IRT model holds. For assessments scored by a transformation of number correct to scale score, such as Science Minnesota Comprehensive Assessments-Series III (MCA-III) or Minnesota Test of Academic Skills (MTAS), the mathematical statement of CSEM is:

$$CSEM(O_X|\theta) = \sqrt{[\sum_{X=0}^{Max\ X} O_X^2 p(X|\theta)] - [\sum_{X=0}^{Max\ X} O_X p(X|\theta)]^2}, \tag{8.6}$$

where O_X is the observed (scaled) score for a particular number-right score X, θ is the IRT ability scale value conditioned on and $p(\bullet)$ is the probability function. $p(X|\theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud & Williams (1995). Their algorithm is a polytomous generalization of the algorithm for dichotomous items given by Lord and Wingersky (1984). The values of θ used are the values corresponding to each raw score point using a reverse table lookup on the test characteristic function (TCF). The table reverse lookup of the TCF is explained in Chapter 7, "Equating and Linking." For each raw score and score scale pair, the procedure results in a CSEM on the scale score metric.

For the Mathematics and Reading MCA-III, which employ pattern scoring based on the 3PL measurement model, the CSEM of student i's scale score is calculated from the CSEM of the obtained θ_i estimate:

$$CSEM(Scale_i) = Spread * CSEM(\theta_i). \tag{8.7}$$

Under the 3PL model, $CSEM(\theta_i)$ is equal to the inverse of the square root of the test information function at θ_i,

$$CSEM(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{8.8}$$

where $I(\theta_i)$ is the test information, calculated as:

$$I(\theta_i) = \sum_{j=1}^{N} \frac{P'_{ij}(\theta_i)^2}{P_{ij}(\theta_i)(1-P_{ij}(\theta_i)}$$

where *N* is the number of items on the assessment, P_{ij} is the probability of student *i* answering question *j* correctly, and $P'_{ij}(\theta)$ is the first derivative of P_{ij} with respect to θ. Note that the calculation depends on the unique set of items each student answers and their estimate of θ, and different students will have different CSEM values even if they have the same raw score and/or theta estimate.

Additional details on calculation of item and test information functions under the 3PL model can be found in standard IRT texts such as Hambleton and Swaminathan (1985).

The Yearbook gives the conditional standard errors of scale scores in the raw and scale score distribution tables. The conditional standard error values can be used in the same way to form confidence bands as described for the traditional test-level SEM values.

Measurement Error for Groups of Students

As is the case with individual student scores, district, school and classroom averages of scores are also influenced by measurement error. Averages, however, tend to be less affected by error than individual scores. Much of the error due to systematic factors (i.e., bias) can be avoided with a well-designed assessment instrument that is administered under appropriate and standardized conditions. The remaining random error present in any assessment cannot be fully eliminated, but for groups of students random error is apt to cancel out (i.e., average to zero). Some students score a little higher than their true score, while others score a little lower. The larger the number in the group, the more the canceling of errors tends to occur. The degree of confidence in the average score of a group is generally greater than for an individual score.

Standard Error of the Mean

Confidence bands can be created for group averages in much the same manner as for individual scores, but in this case the width of the confidence band varies due to the amount of *sampling error*. Sampling error results from using a sample to infer characteristics of a population, such as the mean. Sampling error will be greater to the degree the sample does not accurately represent the population as a whole. When samples are taken from the population at random, the mean of a larger sample will generally have less sampling error than the mean of a smaller sample.

A confidence band for group averages is formed using the standard error of the mean. This statistic, s_e, is defined as

$$s_e = \frac{s_x}{\sqrt{N}}, \tag{8.9}$$

where s_x is the standard deviation of the group's observed scores and *N* is the number of students in the group.

As an example of how the standard error of the mean might be used, suppose that a particular class of 20 students had an average scale score of 455 with a standard deviation equal to 10. The standard error would equal

$$s_e = \frac{10}{\sqrt{20}} = 2.2 \tag{8.10}$$

A confidence bound around the class average would indicate that one could be 68% confident that the true class average on the test was in the interval 455 ± 2.2 (452.8 to 457.2).

Scoring Reliability for Constructed-Response Items and Written Compositions

Reader Agreement

Minnesota's testing contractor uses several procedures to monitor scoring reliability. One measure of scoring reliability is the between-reader agreement observed in the required second reading of a percentage of student responses. These data are monitored on a daily basis by Minnesota's testing contractor during the scoring process. Reader agreement data show the percent perfect agreement of each reader against all other readers. For all constructed-response items and written compositions, 10% of all responses are given a second reading.

Reader agreement data do not provide a mechanism for monitoring drift from established criteria by all readers at a particular grade level. Thus, an additional set of data, resulting from a procedure known as validity scoring, are collected daily to check for reader drift and reader consistency in scoring to the established criteria.

When scoring supervisors at Minnesota's testing contractor identify ideal student responses (i.e., ones that appear to be exemplars of a particular score value), they route these to the scoring directors for review. Scoring directors examine the responses and choose appropriate papers for validity scoring. Validity responses are usually solid score point responses. The scoring directors confirm the true score and enter the student response into the validity scoring pool. Readers score a validity response approximately once out of every 60 responses for reading and every 90 responses for mathematics. Validity scoring is blind; because image-based scoring is seamless, scorers do not know when they are scoring a validity response. Results of validity scoring are analyzed regularly by Minnesota's testing contractor's scoring directors, and appropriate actions are initiated as needed, including the retraining or termination of scorers.

Tables in the Yearbooks give the score frequency distribution for each essay item. Also presented is the percent agreement among readers. As mentioned above, this check of the consistency of readers of the same composition is one form of inter-rater reliability. Rater agreement is categorized as perfect agreement (no difference between readers), adjacent agreement (one score point difference), non-adjacent agreement (two score point difference) or non-agreement (more than two point score difference). Another index of inter-rater reliability reported in the tables is the correlation of ratings from the first and second reader.

Score Appeals

A district may appeal the score assigned to any student's composition about which a question has been raised. In these instances, Minnesota's testing contractor provides an individual analysis of the composition in question.

Auditing of MTAS Administrations and Task Ratings

Many students taking the Minnesota Test of Academic Skills (MTAS) have unique communication styles that require significant familiarity with the student in order to understand their intended communications. Because of this, the MTAS performance tasks are prepared, administered and scored by educators familiar with the student. In order to evaluate rater agreement for this scoring procedure, the Minnesota Department of Education (MDE) recruited Minnesota educators and administrators (current or retired) to serve as scoring auditors. These auditors were trained in the administration and scoring of the MTAS and visited several, randomly selected schools to observe the administration and scoring of actual assessments. The auditors also interviewed the local teachers to get their opinions on the ease of preparing and administering the test. Results of the audits are provided in the Yearbook.

Classification Consistency

Every test administration will result in some error in classifying examinees. The concept of the standard error of measurement (SEM) provides a mechanism for explaining how measurement error can lead to classification errors when cut scores are used to classify students into different achievement levels. For example, some students may have a true achievement level greater than a cut score. However, due to random variations (measurement error), their observed test score may be below the cut score. As a result, the students may be classified as having a lower achievement level. As discussed in the section on the SEM, a student's observed score is most likely to fall into a standard error band around his or her true score. Thus, the classification of students into different achievement levels can be imperfect, especially for the borderline students whose true scores lie close to achievement level cut scores.

For the Minnesota Comprehensive Assessments-Series III (MCA-III) and the Minnesota Test of Academic Skills (MTAS), the levels of achievement are *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards* and *Exceeds the Standards*. An analysis of the consistency in classification is described below.

True level of achievement, which is based on the student's true score, cannot be observed, and therefore classification accuracy cannot be directly determined. It is possible, however, to estimate classification accuracy based on predictions from the IRT model. The accuracy of the estimate depends upon the degree to which the data are fit by the IRT model.

The method followed is based on the work of Rudner (2005). An assumption is made that for a given (true) ability score θ, the observed score $\hat{\theta}$ is normally distributed with a mean of θ and a standard deviation of SE(θ) (i.e., the CSEM at θ). Using this information, the expected proportion of students with true scores in any particular achievement level (bounded by cut scores c and d) who are classified into an achievement level category (bounded by cut scores a and b) can be obtained by:

$$P(Level_k) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{SE(\theta)}\right) - \phi\left(\frac{a-\theta}{SE(\theta)}\right) \right) f(\theta), \tag{8.11}$$

where a and b are theta scale points representing the score boundaries for the observed level, d and c are the theta scale points representing score boundaries for the true level, ϕ is the normal cumulative distribution function and f(θ) is the density function associated with the true score. Because f(θ) is unknown, the observed probability distribution of student theta estimates is used to estimate f(θ) in our calculations.

More concretely, we are using the observed distribution of theta estimates (and observed achievement levels) to represent the true theta score (and achievement level) distribution. Based on that distribution, we use equation (8.11) to estimate the proportion of students at each achievement level that we would expect the test to assign to each possible achievement level. To compute classification consistency, the percentages are computed for all cells of a True vs. Expected achievement level cross-classification table. The diagonal entries within the table represent agreement between true and expected classifications of examinees. The sum of the diagonal entries represents the decision consistency of classification for the test.

Table 8.1 is an example classification table. The columns represent the true student achievement level, and the rows represent the test-based achievement level assignments expected to be observed, given equation (8.11). The meanings of the achievement level labels are: D = *Does Not Meet the Standards*, P = *Partially Meets the Standards*, M = *Meets the Standards* and E = *Exceeds the Standards*. In this example, total decision consistency is 81.0% (sum of diagonal elements).

Table 8.1: Example Classification Table

Achievement Level	True Category D	True Category P	True Category M	True Category E	Exp %
Expected Category D	9.9	1.3	0.0	0.0	11.2
Expected Category P	2.2	8.7	2.3	0.0	13.2
Expected Category M	0.1	5.4	36.7	3.5	45.6
Expected Category E	0.0	0.0	4.2	25.7	29.9
True %	12.1	15.4	43.3	29.2	

It is useful to consider decision consistency based on a dichotomous classification of *Does Not Meet the Standards* or *Partially Meets the Standards* versus *Meets the Standards* or *Exceeds the Standards* because Minnesota uses *Meets the Standards* and above as proficiency for Adequate Yearly Progress (AYP) decision purposes. To compute decision consistency in this case, the table is dichotomized by combining cells associated with *Does Not Meet the Standards* with *Partially Meets the Standards* and combining *Meets the Standards* with *Exceeds the Standards.* For the example table above, this results in a classification accuracy of 92.2%. The percentage of examinees incorrectly classified as *Partially Meets the Standards* or lower, when their true score indicates *Meets the Standards* or above, is 2.3.

The Yearbook contains a table with the overall classification accuracy for each grade and subject of MCA-III and MTAS.

Chapter 9: Validity

Validation is the process of collecting evidence to support inferences from assessment results. A prime consideration in validating a test is determining whether the test measures what it purports to measure. During the process of evaluating whether the test measures the construct of interest, a number of threats to validity must be considered. For example, the test may be biased against a particular group, test scores may be unreliable, students may not be properly motivated to perform on the test, or the test content may not span the entire range of the construct to be measured. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring the test is measuring what it is supposed to measure, it is also important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Chapter 4, "Reports" (in the "Cautions for Score Use" section), and Chapter 6, "Scaling" (in the "Scale Score Interpretations and Limitations" section).

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought, and debate in the measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result, the field has evolved.

This chapter begins with an overview of the major historical perspectives on validity in measurement. Included in this overview is a presentation of a modern perspective that takes an argument-based approach to validity. Following the overview is the presentation of validity evidence for Minnesota assessments.

Perspectives on Test Validity

The following sections discuss some of the major conceptualizations of validity used in educational measurement.

Criterion Validity

The basis of criterion validity is demonstration of a relationship between the test and an external criterion. If the test is intended to measure mathematical ability, for example, then scores from the test should correlate substantially with other valid measures of mathematical ability. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment tasks and the outcome criterion. In order for the observed relationship between the assessment and the criterion to be a meaningful indicator of criterion validity, the criterion should be relevant to the assessment and reliable. Criterion validity is typically expressed in terms of the product-moment correlation between the scores of the test and the criterion score.

There are two types of criterion-related evidence: *concurrent* and *predictive.* The difference between these types lies in the procedures used for collecting validity evidence. Concurrent evidence is collected from both the assessment and the criterion at the same time. An example might be in relating the scores from a district-wide assessment to the ACT assessment (the criterion). In this example, if the results from the district-wide assessment and the ACT assessment were collected in the same semester of the school year, this would provide concurrent criterion-related evidence. On the other hand, predictive evidence is usually collected at different times; typically the criterion information is obtained subsequent to the administration of the measure. For example, if the ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school, whereas the criterion—college grade point average (GPA)—would not be available until a year or two later.

In ideal situations, the criterion-validity approach can provide convincing evidence of a test's validity. However, there are two important obstacles to implementing the approach. First, a suitable criterion must be found. A standards-based test such as the Minnesota Comprehensive Assessments-Series III (MCA-III) is designed to measure the degree to which students have achieved proficiency on the Minnesota Academic Standards. Finding a criterion representing proficiency on the standards may be hard to do without creating yet another test. It is possible to correlate performance on the MCA-III with other types of assessments, such as the ACT or school assessments. Strong correlations with a variety of other assessments would provide some evidence of validity for the MCA-III, but the evidence would be less compelling if the criterion measures are only indirectly related to the standards.

A second obstacle to the demonstration of criterion validity is that the criterion may need to be validated as well. In some cases, it may be more difficult to demonstrate the validity of the criterion than to validate the test itself. Further, unreliability of the criterion can substantially attenuate the correlation observed between a valid measure and the criterion.

Additional criterion-related validity evidence on the Minnesota assessments will be collected and reported in an ongoing manner. These data are most likely to come from districts conducting program evaluation research, university researchers and special interest groups researching topics of local interest, as well as the data-collection efforts of MDE.

Content and Curricular Validity

Content validity is a type of test validity addressing whether the test adequately samples the relevant domain of material it purports to cover. If a test is made up of a series of tasks that form a representative sample of a particular domain of tasks, then the test is said to have good content validity. For example, a content-valid test of mathematical ability should be composed of tasks that allow students to demonstrate their mathematical ability.

Evaluating content validity is a subjective process that is based on rational arguments. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity speaks only to the validity of the test itself, not to decisions made based on the test scores. For example, a poor score on a content-valid mathematics test indicates that the student did not *demonstrate* mathematical ability. But from this alone, one cannot conclusively conclude the student has low mathematical ability. This conclusion can only be reached if it can be shown or argued that the student put forth his or her best effort, the student was not distracted during the test, and the test did not contain a bias that prevented the student from scoring well.

Generally, achievement tests such as the Minnesota assessments are constructed in a way to ensure they have strong content validity. As documented by this manual, tremendous effort is expended by MDE, the contractors, and educator committees to ensure Minnesota assessments are content-valid. Although content validity has limitations and cannot serve as the only evidence for validation, it is an important piece of evidence for the validation of Minnesota assessments.

Construct Validity

The term construct validity refers to the degree to which the test score is a measure of the characteristic (i.e., construct) of interest. A construct is an individual characteristic assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic is inferred from an assessment result, a generalization or interpretation in terms of a construct is being made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are "good problem solvers" implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate this is a reasonable and valid use of the results.

Construct-related validity evidence can come from many sources. The fourth edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & the National Council on Measurement in Education, 1985) provides the following list of possible sources:

- High intercorrelations among assessment items or tasks that attest the items are measuring the same trait, such as a content objective, subdomain, or construct
- Substantial relationships between the assessment results and other measures of the same defined construct
- Little or no relationship between the assessment results and other measures that are clearly not of the defined construct
- Substantial relationships between different methods of measurement regarding the same defined construct
- Relationships to non-assessment measures of the same defined construct

Messick (1988) describes construct validity as a "unifying force" in that inferences based on criterion evidence or content evidence can also be framed by the theory of the underlying construct. From this point of view, validating a test is essentially the equivalent of validating a scientific theory. As Cronbach and Meehl (1955) first argued, conducting construct validation requires a theoretical network of relationships involving the test score. Validation not only requires evidence supporting the notion that the test measures the theoretical construct, but it further requires evidence be presented that discredits every plausible alternative hypothesis as well. Because theories can only be supported or falsified but never proved, validating a test becomes a never-ending process.

Kane (2006) states that construct validity is now widely viewed as a general and all-encompassing approach to accessing test validity. However, in Kane's view there are limitations of the construct-validity approach, including the need for strong measurement theories and the general lack of guidance on how to conduct a validity assessment.

Argument-Based Approach to Validity

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & the National Council on Measurement in Education, 1999, 2014) recommends establishing the validity of a test through the use of a *validity argument.* According to the 2014 version of the *Standards for Educational and Psychological Testing* this term is defined as "an explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended uses."

Kane (2006), following the work of Cronbach (1988), presents an argument-based approach to validity that seeks to address the shortcomings of previous approaches to test validation. The argument-based approach creates a coherent framework (or theory) that clearly lays out theoretical relationships to be examined during test validation.

The argument-based approach given by Kane (2006) delineates two kinds of arguments. An *interpretative argument* specifies all the inferences and assumptions made in the process of assigning scores to individuals and the interpretations made of those scores. The interpretative argument provides a step-by-step description of the reasoning (if-then statements) allowing one to interpret test scores for a particular purpose. Justification of that reasoning is the purpose of the *validity argument.* The validity argument is a presentation of all the evidence supporting the interpretative argument.

The interpretative argument is usually laid out logically in a sequence of stages. For achievement tests such as the Minnesota assessments, the stages can be broken out as *scoring*, *generalization, extrapolation* and *implication.* Descriptions of each stage are given below along with examples of the validity arguments within each stage.

Scoring

The scoring part of the interpretative argument deals with the processes and assumptions involved in translating the observed responses of students into observed student scores. Critical to these processes are the quality of the scoring rubrics, the selection, training and quality control of scorers, and the appropriateness of the statistical models used to equate and scale test scores. Empirical evidence that can support validity arguments for scoring includes inter-rater reliability of constructed-response items and item-fit measures of the statistical models used for equating and scaling. Because Minnesota assessments use item response theory (IRT) models, it is also important to verify the assumptions underlying these models.

Generalization

The second stage of the interpretative argument involves the inferences about the *universe score* made from the observed score. Any test contains only a sample of all the items that could potentially appear on the test. The universe score is the hypothetical score a student would be expected to receive if the entire universe of test questions could be administered. Two major requirements for validity at the generalization stage are: (1) the sample of items administered on the test is representative of the universe of possible items and (2) the number of items on the test is large enough to control for random measurement error. The first requirement entails a major commitment during the test-development process to ensure content validity is upheld and test specifications are met. For the second requirement, estimates of test reliability and the standard error of measurement are key components to demonstrating that random measurement error is controlled.

Extrapolation

The third stage of the interpretative argument involves inferences from the universe score to the *target score.* Although the universe of possible test questions is likely to be quite large, inferences from test scores are typically made to an even larger domain. In the case of the Minnesota Comprehensive Assessments-Series III (MCA-III), for example, not every standard and benchmark is assessed by the test. Some standards and benchmarks are assessed only at the classroom level because they are impractical or impossible to measure with a standardized assessment. It is through the classroom teacher these standards and benchmarks are assessed. However, the MCA-III is used for assessment of proficiency with respect to all standards. This is appropriate only if interpretations of the scores on the test can be validly extrapolated to apply to the larger domain of student achievement. This domain of interest is called the *target domain*, and the hypothetical student score on the target domain is called the *target score*. Validity evidence in this stage must justify extrapolating the universe score to the target score. Systematic measurement error could compromise extrapolation to the target score.

The validity argument for extrapolation can use either analytic evidence or empirical evidence. Analytic evidence largely stems from expert judgment. A credible extrapolation argument is easier to make to the degree the universe of test questions largely spans the target domain. Empirical evidence of extrapolation validity can be provided by criterion validity when a suitable criterion exists.

Implication

The implication stage of the interpretative argument involves inferences from the target score to the decision implications of the testing program. For example, a college-admissions test may be an excellent measure of student achievement as well as a predictor of college GPA. However, an administrator’s decision of how to use a particular test for admissions has implications that go beyond the selection of students who are likely to achieve a high GPA. No test is perfect in its predictions, and basing admissions decisions solely on test results may exclude students who would excel if given the opportunity.

Although much of this manual describes evidence for the validity of individual student scores for making inferences about student proficiency, the ultimate implications for the MCA-III involve school accountability and the impact the school has on improving student scores. Even if the testing program is successful in increasing student achievement on the standards, other unintended implications of the program must be addressed. Kane (2006) lists some potential negative effects on schools, such as increased dropout rates and narrowing of the curriculum. In the coming years, studies will need to be conducted to validate the intended positive effects of the testing program as well as to investigate possible unintended negative effects.

Validity Argument Evidence for the Minnesota Assessments

The following sections present a summary of the validity argument evidence for each of the four parts of the interpretive argument: scoring, generalization, extrapolation, and implication. Much of this evidence is presented in greater detail in other chapters in this manual. In fact, the majority of this manual can be considered validity evidence for the Minnesota assessments (for example, item development, performance standards, scaling, equating, reliability, performance item scoring, and quality control). Relevant chapters are cited as part of the validity evidence given below.

Scoring-Validity Evidence

Scoring-validity evidence can be divided into two sections: (1) the evidence for the scoring of performance items and (2) the evidence for the fit of items to the model.

Scoring of Performance Items

The scoring of constructed-response items and written compositions on Minnesota assessments is a complex process that requires its own chapter in order to be described fully. Chapter 10, "Constructed-Response Items and Written Compositions," gives complete information on the careful attention paid to the scoring of performance items. The chapter's documentation of the processes of rangefinding, rubric review, recruiting and training of scorers, quality control, appeals, and security provides some of the evidence for the validity argument that the scoring rules are appropriate. Further evidence comes from Yearbook tables reporting inter-rater agreement and inter-rater reliabilities. The results in those tables show both of these measures are generally high for Minnesota assessments.

The auditing of the Minnesota Test of Academic Skills (MTAS) administrations and task ratings supplies validity evidence for the scoring of these performance tasks. The auditing procedure is described in Chapter 8, "Reliability," and results of the audits are provided in the Yearbook.

Model Fit and Scaling

Item response theory (IRT) models provide a basis for the Minnesota assessments. IRT models are used for the selection of items to go on the test, the equating procedures, and the scaling procedures. A failure of model fit would make the validity of these procedures suspect. Item fit is examined during test construction. Any item displaying misfit is carefully scrutinized before a decision is made to put it on the test. However, the vast majority of items fit.

Further evidence of the fit for the IRT model comes from dimensionality analyses. IRT models for Minnesota assessments assume the domain being measured by the test is relatively one-dimensional. To test this assumption, a principal-components analysis is performed.

Another check for one-dimensionality can be made at the item level. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called point-biserial correlation) is the correlation between an item and the total test score. Conceptually, if an item has a high item total correlation (i.e., 0.30 or above), it indicates that students who performed well on the test got the item right and students who performed poorly on the test got the item wrong; the item did a good job discriminating between high-ability and low-ability students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate the items on the test require this construct to be answered correctly. The Yearbooks present item-total correlations in the tables of item statistics. For Minnesota assessments, item-total correlations are generally high.

Justification for the scaling procedures used for Minnesota assessments is found in Chapter 6, "Scaling."

While it is important to validate the fit of IRT models and the scaling procedures used for each specific Minnesota assessment, it is also critical to examine factors specific to the administration of the test questions that could invalidate scores. One such factor relevant for the Mathematics and Reading Minnesota Comprehensive Assessments-Series III (MCA-III) assessments is the mode of administration. The MCA-III can be taken either online or on paper, depending upon the choice made by the school district. Thus, it is important to evaluate whether mode effects between the two versions of the test could raise validity concerns for the test scores. In spring 2011, a mode-comparability study was conducted using a matched group study design to compare the results of students taking one of the online operational test forms with the results of student taking a similar form given on paper for the MCA-III

Mathematics grades 3–8. The results of the comparability study suggested that although testing mode was found to impact certain items in common between the online and paper versions, this effect could be mitigated by essentially treating the online and paper versions of the items as distinct items with mode-specific item parameters. The online and paper parameters were scaled to a common metric by using a set of linking items not affected by mode. The complete MCA-III Mathematics grades 3–8 comparability report can be found on the MDE website at http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html.

In the spring 2013, a mode-comparability study was conducted using matched samples to compare student performance on the MCA-III Reading online and paper assessment modes. The results of the comparability study suggested that there was a mode effect but that it could be resolved by applying the results of a Stocking-Lord equating to place the scores on the same scale. The complete MCA-III Reading comparability report is available upon request from MDE.

.

In addition, in the spring 2014, a mode-comparability study was conducted using matched samples to compare student performance on the MCA-III Mathematics grade 11 online and paper assessment modes. The results of the comparability study suggested that there was a mode effect, but that it could be resolved by applying the results of a Stocking-Lord equating to place the scores on the same scale. The complete MCA-III Mathematics grade 11 comparability report is available upon request from MDE.

Generalization-Validity Evidence

There are two major requirements for validity that allow generalization from observed scale scores to universe scores. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement comes from content validity. Content validity is documented through evidence that the test measures the state standards and benchmarks. The second requirement for validity at the generalization stage is that random measurement error on the test is controlled. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Evidence is also presented concerning the use of Minnesota assessments for different student populations. These sources of evidence are reported in the sections that follow.

Evidence of Content Validity

The tests of the Minnesota Assessment System are based on content standards and benchmarks along with extensive content limits that help define what is to be assessed. Committees of educators collaborate with item development experts, assessment experts and Minnesota Department of Education (MDE) staff annually to review new and field-tested items to assure the tests adequately sample the relevant domain of material the test purports to cover. These review committees participate in this process to ensure test-content validity for each test. If a test is a static test, the committees meet only during the years when the test is being developed.

A sequential review process for committees is used by MDE and was outlined in Chapter 2, “Test Development.” In addition to providing information on the difficulty, appropriateness, and fairness of items and performance tasks, committee members provide a check on the alignment between the items and the benchmarks measured. When items are judged to be relevant, that is, representative of the content defined by the standards, this provides evidence to support the validity of inferences made regarding knowledge of this content from the results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (for example, reclassification, rewording) or elect to eliminate the item from the field-test item pool. For example, items approved are later embedded in live Minnesota Comprehensive Assessments-Series III (MCA-III) forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure the items measure appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the test.

Educators are also involved in evidence of content validity in other ways. Many current and former Minnesota educators and some educators from other states work as independent contractors to write items specifically to measure the objectives and specifications of the content standards for the tests. Using a varied source of item writers provides a system of checks and balances for item development and review, reducing single-source bias. Since many different people with different backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. The input and review by these assessment professionals provide further support of the item being an accurate measure of the intended objective.

The Yearbook contains tables showing the number of assessment components, tasks or items matching each content standard. A comprehensive presentation of the test specifications can be found on the MDE website at http://education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html.

Evidence of Control of Measurement Error

Reliability and the standard error of measurement (SEM) are discussed in Chapter 8, "Reliability." The Yearbook has tables reporting the conditional SEM for each scale score point and the coefficient alpha reliabilities for raw scores, broken down by gender and ethnic groups. As discussed in Chapter 8, these measures show Minnesota assessments to be reliable.

Further evidence is needed to show that the IRT model fits well. Item-fit statistics and tests of one-dimensionality apply here, as they did in the section describing evidence argument for scoring. As described above, these measures indicate good fit of the model.

Validity Evidence for Different Student Populations

It can be argued from a content perspective that Minnesota assessments are not more or less valid for use with one subpopulation of students relative to another. Minnesota assessments measure the statewide content standards that are required to be taught to all students. In other words, the tests have the same content validity for all students because what is measured is taught to all students, and all tests are given under standardized conditions to all students.

Great care has been taken to ensure the items in the Minnesota assessments are fair and representative of the content domain expressed in the content standards. Additionally, much scrutiny is applied to the items and their possible impact on the demographic subgroups that make up the population of the state of Minnesota. Every effort is made to eliminate items that may have ethnic or cultural biases. As described in Chapter 2, "Test Development," item writers are trained on how to avoid economic, regional, cultural, and ethnic bias when writing items. After items are written and passage selections are made, committees of Minnesota educators are convened by MDE to examine items for potential subgroup bias. As described in Chapter 7, "Equating and Linking," items are further reviewed for potential bias by committees of educators and MDE after field-test data are collected.

Extrapolation-Validity Evidence

Validity for extrapolation requires evidence that the universe score is applicable to the larger domain of interest. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. These lines of evidence are detailed below.

Analytic Evidence

The standards create a common foundation to be learned by all students and define the domain of interest. As documented in this manual, Minnesota assessments are designed to measure as much of the domain defined by the standards as possible. Although a few benchmarks from the standards can only be assessed by the classroom teacher, the majority of benchmarks are assessed by the tests. Thus, it can be inferred that only a small degree of extrapolation is necessary in order to use test results to make inferences about the domain defined by the standards.

The Minnesota Test of Academic Skills (MTAS) is also tied to the Minnesota Academic Standards. Because the MTAS is designed to measure the extent to which students with significant cognitive disabilities are making progress in the general curriculum, the achievement standards need to be modified to some degree. The MTAS measures student progress on state grade-level content standards but at reduced breadth, depth, and complexity. Chapter 2, “Test Development,” describes in detail the process of aligning the alternate achievement standards to the general standards and serves as validity-evidence documentation for the MTAS.

The use of different item types also increases the validity of Minnesota assessments. The combination of multiple-choice, technology-enhanced, and constructed-response items results in assessments measuring the domain of interest more fully than if only one type of response format was used.

A threat to the validity of the test can arise when the assessment requires competence in a skill unrelated to the construct being measured. The Minnesota assessments allow accommodations for students with vision impairment or other special needs. The use of accommodated forms allows accurate measurement of students who would otherwise be unfairly disadvantaged by taking the standard form. Accommodations are discussed in Chapter 3, “Test Administration.”

Empirical Evidence

Empirical evidence of extrapolation is generally provided by criterion validity when a suitable criterion exists. As discussed before, finding an adequate criterion for a standards-based achievement test can be difficult.

Studies investigating criterion validity have yet to be carried out for the MCA-III. Because no other assessment is likely to be found to measure the standards as well as the MCA-III, the most promising empirical evidence would come from criterion validity studies with convergent evidence. Any test that measures constructs closely related to the standards could serve as a criterion. Although these tests would not measure the standards as well as the MCA-III, they could serve as an external check. If a number of these external tests could be found that are highly correlated with the MCA-III, the converging evidence from them would provide justification for extrapolation.

Implication-Validity Evidence

There are inferences made at different levels based on the Minnesota assessments. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others. For example, the tests of the MCA-III report individual student scores, but some students may feel that few ramifications of the test directly affect them; such students may fail to put forth their full effort. Although this manual documents in detail evidence showing that the MCA-III is a valid measure of student achievement on the standards, individual and school-level scores are not valid if students do not take the test seriously.

One index of student effort is the percentage of blank or "off topic" responses to constructed-response items. Because constructed-response items require more time and cognitive energy, low levels of non-response on these items is evidence of students giving their full effort. The 2009 Yearbook data show non-response rates for Minnesota Assessments to be approximately 6% or less.

One of the most important inferences to be made concerns the student's proficiency level, especially for accountability tests like the MCA-III and the MTAS. Even if the total correct score can be validated as an appropriate measure of the standards, it is still necessary that the scaling and performance-level-designation procedures be validated. Because scaling and standard setting are both critical processes for the success of Minnesota assessments, separate chapters are devoted to them in this manual. Chapter 5, "Performance Standards," discusses the details of setting performance standards, and Chapter 6, "Scaling," discusses scaling. These chapters serve as documentation of the validity argument for these processes.

At the aggregate level (school, district, or statewide), the implication validity of school accountability assessments like the MCA-III can be judged by the impact the testing program has on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage of students classified as proficient. As mentioned before, there exists a potential for negative effects on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects as well.

Summary of Validity Evidence

Validity evidence is described in this chapter as well as other chapters of this manual. In general, validity arguments based on rationale and logic are strongly supported for Minnesota assessments. The empirical-validity evidence for the scoring- and the generalization-validity arguments for Minnesota assessments are also quite strong. Reliability indices, model fit, and dimensionality studies provide consistent results, indicating that the Minnesota assessments are properly scored and that scores can be generalized to the universe score.

Less strong is the empirical evidence for extrapolation and implication. This is due in part to the absence of criterion studies. Because an ideal criterion for a test like the MCA-III probably cannot be found, empirical evidence for the extrapolation argument may need to come from several studies showing convergent validity evidence. Further studies are also needed to verify some implication arguments. This is especially true for the inference that the state's accountability program is making a positive impact on student proficiency and school accountability without causing unintended negative consequences.

Chapter 10: Written Compositions

For the Graduation-Required Assessment for Diploma (GRAD) Written Composition, students are required to write essays based on a given prompt. The procedure for scoring written essays is described in this chapter.

Scoring Process

Outlined below is the scoring process that the Minnesota's testing contractor follows. This procedure is used to score responses to written composition items for the Minnesota assessments.

Rangefinding and Rubric Review

Rangefinding and rubric review took place prior to scoring the operational assessment for the spring 2013 administration of the Written Composition GRAD. Rangefinding was held at the headquarters of the Minnesota Department of Education in Saint Paul, Minnesota, February 26 through March 1, 2013. The task of this rangefinding session was to augment previously developed Holistic and Analytic scorer training materials.

Rangefinding responses representing various levels of student performance were chosen from the field-test responses for the two prompts used in the spring 2013 administration. These were assembled into rangefinding sets and reviewed by a panel of Minnesota educators. Individual copies of each set were produced for each member of the rangefinding panel to review. The panel discussed each response and arrived at a consensus score. This process continued until the rangefinding panel scored a sufficient number of rangefinding responses to augment the anchor, training sets, and qualifying sets.

Rangefinding responses representing various levels of student performance were chosen from the field-test responses for the two prompts used in the spring 2013 administration. These were assembled into rangefinding sets and reviewed by a panel of Minnesota educators. Individual copies of each set were produced for each member of the rangefinding panel to review. The panel discussed each response and arrived at a consensus score. This process continued until the rangefinding panel scored a sufficient number of rangefinding responses to augment the anchor, training sets, and qualifying sets.

Responses scored by the rangefinding panel were incorporated to update existing training materials alongside essays written to previously used prompts. Together, these comprise the anchor, training, and qualifying sets. Prior to use for training, all materials were forwarded to MDE for review and approval as further assurance that panel decisions were accurately enacted.

Steps were taken throughout the preparation of rangefinding materials and during the meetings to ensure security. Materials were stored in locked facilities. The rangefinding rooms were always locked when unoccupied. All rangefinding materials were accounted for at the end of each rangefinding session.

Recruiting and Training Scorers

Minnesota's testing contractor selects scorers who are articulate, concerned with the task at hand, and, most importantly, flexible. Their scorers must have strong content-specific backgrounds: they are educators, writers, editors, and other professionals. They are valued for their experience, but at the same time, they are required to set aside their own biases about student performance and accept the scoring standards of the client's program.

All the scorers have at least a four-year college degree in a relevant field and a demonstrated ability to write. Many of the scorers have years of experience with scoring large-scale writing responses, and most of the scorers for the GRAD Written Composition have prior experience on a writing project.

The testing contractor has a Human Resources Coordinator dedicated solely to recruiting and retaining their scorer staff. Applications for scorer positions are screened by the Project Director, the Human Resources Coordinator, and recruiting staff to create a large pool of potential scorers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, scorer candidates are asked to demonstrate their own proficiency at writing by responding to a writing topic.

Training

Thorough training is vital to the successful completion of any scoring. Scoring Directors follow a series of prescribed steps to ensure training is consistent and of the highest quality.

Team Leaders assist the Scoring Directors with scorer training and monitoring. Comprehensive Team Leader training lasts approximately two days. Team Leader training follows the procedures that are used in the scorer training (detailed below), but it is more comprehensive due to the training and monitoring responsibilities required of the Team Leaders.

The primary goal of training is for scorers to internalize the scoring protocol so that they will accurately apply the rubric to responses. Scorers are better able to comprehend the scoring guidelines in context, so training begins with a room-wide presentation of the Scoring Guide, which includes the rubric in conjunction with the anchor responses. Anchor responses are the primary points of reference for scorers as they internalize the rubric. There are three anchor responses for each score point value, which are annotated with language from the rubric.

After presentation and discussion of the rubric and anchor papers, the scorers are given training sets. Training sets contain responses that are used to help scorers become familiar with applying the rubric. Some responses clearly represent the score point. Others are selected because they represent borderline responses. Use of these training sets provides guidance to scorers in defining the line between score points. Training is a continuous process, and scorers are consistently given feedback as they score. After training, scorers are required to demonstrate scoring proficiency on a qualifying set of prescored student responses. Scorers who are not able to demonstrate sufficient accuracy are removed from the project and do not score any live Minnesota responses.

Quality Control

A variety of reports are produced throughout the scoring process to allow scoring supervisory staff to monitor the progress of the project, the reliability of scores assigned and individual scorers' work. This includes the Scoring Summary Report, which provides the following information:

- *Daily and Cumulative Inter-Rater Reliability.* This details how many times scorers were in exact agreement, assigned adjacent scores, or required resolutions. The reliability is monitored daily and cumulatively for the project.
- *Daily and Cumulative Score Point Distributions.* This shows how often each score point has been assigned by each scorer. The distributions are produced both on a daily basis and

cumulatively for the entire scoring project. This allows the Scoring Directors and Team Leaders to monitor scoring trends.

Additionally, Team Leaders conduct routine read-behinds to observe, in real time, scorers' performance. Team Leaders utilize live, scored responses to provide ongoing feedback and, if necessary, retraining for scorers.

Validity responses are pre-scored responses that are "seeded" to scorers during scoring. Validity reports compare the scorers' scores to the predetermined scores in order to detect possible room drift and individual scorer trends. The validity responses are "blind" to the scorers: scorers cannot distinguish a validity response from any other type of response.

With the help of the quality-control reports, the Scoring Directors and Team Leaders closely monitor each scorer's performance and take corrective measures, such as retraining, when necessary. If necessary, scorers are dismissed when, in the opinion of the Scoring Directors, those scorers have been counseled, retrained, given every reasonable opportunity to improve, and are still performing below the acceptable standard.

Appeals

Once an appeal has been identified, the Writing Content Specialist reviews the score in question. An annotation is prepared where, following review, the scoring director will either justify the score or provide a re-score. In either case, the annotation explains the action taken.

Security

To ensure security is never compromised, the following safeguards are employed:

- Controlled access to the facility, allowing only Minnesota's testing contractor and customer personnel to have access during scoring.
- No materials are permitted to leave the facility during the project without the express permission of a person or persons designated by the Minnesota Department of Education (MDE).
- Scoring personnel must sign a nondisclosure and confidentiality form in which they agree not to use or divulge any information concerning the tests.
- All staff must wear Minnesota's testing contractor's identification badges at all times in the scoring facility.
- No recording or photographic equipment is allowed in the scoring area without the consent of MDE.
- Any contact with the press is handled through MDE.

Chapter 11: Quality-Control Procedures

The Minnesota assessment program and its associated data play an important role in the state accountability system as well as in many local evaluation plans. Therefore, it is vital that quality-control procedures are implemented to ensure the accuracy of student-, school- and district-level data and reports. Minnesota's testing contractor has developed and refined a set of quality procedures to help ensure that all the Minnesota Department of Education's (MDE) testing requirements are met or exceeded. These quality-control procedures are detailed in the paragraphs that follow. In general, Minnesota's testing contractor's commitment to quality is evidenced by initiatives in two major areas:

1. Task-specific quality standards integrated into individual processing functions and services
2. A network of systems and procedures that coordinates quality across processing functions and services

Quality Control for Test Construction

Test construction for the Minnesota assessments follows the legally-sanctioned test-development process used by Minnesota's testing contractor as described in Chapter 2, "Test Development," of this document (Smisko, Twing, & Denny, 2000). Following this process, items are selected and placed on a particular pre-equated test form in order to provide a strictly parallel form both in terms of content and statistics. Item and form statistical characteristics from the baseline test are used as targets when constructing the current test form. Similarly, the baseline raw score-to-scaled score tables are used as the target tables that the pre-equated test form (under construction) should match. Once a set of items has been selected, MDE reviews and may suggest replacement items (for a variety of reasons). Successive changes are made and the process iterates until both Minnesota's testing contractor and MDE agree to a final pre-equated form. This form is provided to Minnesota's testing contractor for form construction and typesetting, as outlined in a subsequent section of this chapter.

Quality Control for Scannable and Non-Scannable Documents

Minnesota's testing vendor follows a meticulous set of internal quality standards to ensure high-quality printed products. Specific areas of responsibility for staff involved in materials production include monitoring all materials-production schedules to meet contract commitments, overseeing the production of scannable test materials, coordinating detailed printing and post-printing specifications, outlining specific quality control requirements for all materials, and conducting print reviews and quality checks. The quality production and printing processes follow:

1. *Information Systems Review and Quality Check.* Quality Assurance, Information Systems, and Programming staff are responsible for certifying that all scannable documents are designed, developed, and printed within specified scanning requirements and tolerances. This technical review ensures error-free processing of scannable documents that prevents delays in the data delivery and reporting schedule.
2. *Printers' Reviews and Quality Checks*. Project Management and Print Procurement staff works closely with the printers during the print production phase. Press proofs are checked to ensure high-quality printing and to verify adherence to printing specifications. The printing staff randomly pulls documents throughout the print run.

Quality Control in Production

Minnesota's testing contractor uses the "batch control" concept for document processing. When documents are received and batched, each batch is assigned an identifying number unique within the facility. This unique identifier assists in locating, retrieving, and tracking documents through each processing step. The batch-identifying number also guards against loss, regardless of batch size.

All Minnesota assessment documents are continually monitored by Minnesota's testing contractor's proprietary computerized Materials Management System (MMS). This mainframe system can be accessed throughout Minnesota's testing contractor's processing facility, enabling Minnesota's testing contractor staff to instantly determine the status of all work in progress. MMS efficiently carries the planning and control function to first-line supervisory personnel so that key decisions can be made properly and rapidly. Since MMS is updated on a continuous basis, new priorities can be established to account for Minnesota assessments documents received after the scheduled due date, late vendor deliveries or any other unexpected events.

Quality Control in Scanning

Minnesota's testing contractor has many high-speed scanners in operation, each with a large per-hour scanning capability. Stringent quality-control procedures and regular preventative maintenance ensure that the scanners are functioning properly at all times. In addition, application programs consistently include quality-assurance checks to verify the accuracy of scanned student responses.

Through many years of scanning experience, Minnesota's testing contractor has developed a refined system of validity checks, editing procedures, error corrections, and other quality controls, ensuring maximum accuracy in the reporting of results. During scanning, Minnesota assessments documents are carefully monitored by a trained scanner operator for a variety of error conditions. These error routines identify faulty documents, torn and crumpled sheets, document misfeeds, and paper jams. In these events, the scanner stops automatically. The operator can easily make corrections in most cases; otherwise, corrections will be made in the editing department.

All image-scanning programs go through quality review before test materials arrive at our facilities. Throughout the scanning process, batches are checked for quality and scanning accuracy. All scanners are regularly calibrated and cleaned to ensure accurate, consistent scoring.

Quality Control in Editing and Data Input

As Minnesota assessment answer documents are scanned, the data are electronically transcribed directly to data files, creating the project's database. After scanning, documents are processed through a computer-based editing program to detect omissions, inconsistencies, gridding errors, and other error-suspect conditions in specified response fields. Marks or omits that do not meet predefined editing standards are flagged and routed for resolution. To produce clean data files, editing staff follow strict quality-control procedures and edit specifications that will be mutually developed by MDE and the testing contractor. Any changes made to scanned values and all items entered the first time are double-keyed for verification. After verification, a quality-control report is generated for post-editing review.

During the post-editing step, the actual number of documents scanned is compared with the number of scannable documents assigned to the box during booklet check-in; any count discrepancies are resolved. Suspect student precodes, district and school numbers, and document IDs are reviewed for additional verification. Editing quality-control reports are reviewed to ensure that changes were processed accurately. Corrections during post-editing are made electronically. A new validation report is generated to confirm that the changes have been processed accurately and the report is clean.

Quality Control in Handscoring

Accurate and consistent results are the backbone of all handscoring activities. The following methods used by the testing contractor guarantee scoring quality:

1. **Anchors** are pre-scored student responses used to define and exemplify the score scale. For each score point, anchors are selected to reflect the entire range of performance represented by that score based on the judgment of the rangefinding team. The anchors, which are included in the scoring guide and training sets, are used to clarify the scoring scale during scorer training.
2. After an **intensive training** session, qualifying rounds are conducted by scoring directors.
3. **Qualifying** responses are similar to training examples in that they have been pre-scored through rangefinding. The responses are divided into sets and scored independently by each scorer trainee. The data from these qualifying rounds are used to determine which scorer trainees qualify for actual scoring.
4. **Recalibration** responses may be used throughout the scoring session. Similar to the training and qualifying materials, the recalibration materials are selected from responses scored through rangefinding. Recalibration sets are used to monitor scoring and to refocus scorers on the scoring standards by comparing the predetermined score with that assigned by the scorer. In addition, these examples may be used by the scoring director or team leaders for a retraining session.
5. **Validity** responses detect possible room drift and individual scorer problems. Validity reports compare scorers' scores with predetermined scores. The validity responses are "blind" to the scorers; scorers cannot distinguish a validity response from any other type of response.
6. Team leaders conduct routine **read-behinds** for all scorers.
7. Another measure of rating scoring quality is **inter-rater reliability and score point distribution reports.** To monitor scorer reliability and maintain an acceptable level of scoring accuracy, the testing contractor closely reviews reports that are produced daily. The reports document individual scorer data, individual scorer number, number of responses scored, individual score point distributions, and exact agreement rates. The testing contractor investigates the issue and resolves any problems those reports identify.

Quality Control for Online Test Delivery Components

Each release of every one of Minnesota's testing contractor's systems goes through a complete testing cycle, including regression testing. Each release and every time the vendor publishes a test, the system goes through User Acceptance Testing (UAT). During UAT, the vendor provides Minnesota with login information to an identical (though smaller-scale) testing environment to which the system has been deployed. The testing vendor provides recommended test scenarios and constant support during the UAT period.

Deployments to the production environment all follow specific, approved deployment plans. Teams working together execute the deployment plan. Each step in the deployment plan is executed by one team member and verified by a second. Each deployment undergoes shakeout testing following the deployment. This careful adherence to deployment procedures ensures that the operational system is identical to the system tested on the testing and staging servers. Upon completion of each deployment project, management at the testing vendor approves the deployment log.

During the course of the year, some changes may be required to the production system. Outside of routine maintenance, no change is made to the production system without approval of the Production Control Board (PCB). The PCB includes the director of the vendor’s Assessment Program or the Chief Operating Officer, the director of Pearson Computer and Statistical Sciences Center (CSSC), and the project director for the Minnesota Assessment Programs. Any request for change to the production system requires the signature of the system’s lead engineer. The PCB reviews risks, test plans, and test results. In the event that any proposed change will affect client functionality or pose a risk to the operation of a client system, the PCB ensures that Minnesota is informed and in agreement with the decision.

Deployments happen during a maintenance window that is agreed upon by the client and the testing vendor. The vendor schedules the deployments at a time that can accommodate full regression testing on the production machines. Any changes to the database or procedures that in any way might affect performance are typically subjected to a load test at this time.

Quality Control for Test-Form Equating

Test-form equating is the process that enables fair and equitable comparisons both across test forms within a single year and between test administrations across years. Minnesota’s testing contractor uses several quality-control procedures to ensure this equating is accurate.

1. Minnesota’s testing contractor performs a “key check” analysis to ensure the appropriate scoring key is being used. For assessments that are scored immediately, item performance is examined at regular intervals throughout the test window. For tests that are scored after the close of the test window, this check is performed on the equating sample, which historically consists of about 80% of all student records.
2. For fixed-form assessments, the following steps are also included:
    a) Once the key is verified, Minnesota’s testing contractor performs statistical analyses (post-equating) to generate comparable item response theory (IRT) item parameters to those used during test construction or pre-equating.
    b) The post-equated and pre-equated values of anchor items are compared and differences beyond expectation are investigated and resolved.
    c) For assessments that are scored using raw-to-scale score conversion tables, new post-equated conversion tables are generated and compared with the pre-equated tables. Any unexpected differences are resolved.
    d) Expected passing rates or rates of classification are generated and compared with previous years.
    e) An equating summary is provided to MDE and the National Technical Advisory Committee (TAC) for review.

Glossary of Terms

The following glossary of terms as used in this document is provided to assist the reader regarding language that may not be familiar.

Adequate Yearly Progress (AYP)

The amount of progress required by schools each year to meet established federal Title 1 goals. The specific progress required is negotiated by the state.

Assessment

The process of collecting information in order to support decisions about students, teachers, programs, and curricula.

Classification Accuracy

The degree to which the assessment accurately classifies examinees into the various levels of achievement. Also referred to as decision consistency.

Coefficient Alpha

An internal consistency reliability estimate that is appropriate for items scored dichotomously or polytomously. Estimates are based on individual item and total score variances.

Consequential Validity

Evidence that using a test for a particular purpose leads to desirable social outcomes.

Construct Validity

Evidence that performance on the assessment tasks and the individual student behavior that is inferred from the assessment shows strong agreement and that this agreement is not attributable to other aspects of the individual or assessment.

Content Standards

Content standards describe the goals for individual student achievement, specify what students should know, and specify what students should be able to do in identified disciplines or subject areas.

Content Validity

Evidence that the test items represent the content domain of interest.

Differential Item Functioning (DIF)

A term applied to investigations of test fairness. Explicitly defined as difference in performance on an item or task between a designated minority and majority group, usually after controlling for differences in group achievement or ability level.

Internal Consistency Reliability Estimate

An estimate of test-score reliability derived from the observed covariation among component parts of the test (for example, individual items or split halves) on a single administration of the test. Cronbach's coefficient alpha and split-half reliability are commonly used examples of the internal consistency approach to reliability estimation.

Limited English Proficiency (LEP)

A designation give to an individual whose primary language is a language other than English.

Modifications

Changes made to the content and performance expectations for students.

No Child Left Behind (NCLB)

Federal law enacted in 2001 that requires school districts to be held accountable in order to receive federal funding. Every state is required to create a plan that involves setting performance targets so that all students are academically proficient by the year 2013–2014.

Parallel Forms

Two tests constructed to measure the same thing from the same table of specifications with the same psychometric and statistical properties. True parallel test forms are not likely to ever be found. Most attempts to construct parallel forms result in alternate test forms.

Performance Standards

Performance standards define what score students must achieve in order to demonstrate proficiency. On the Minnesota Basic Skills Test (BST), they describe what is required to pass. On the Minnesota Comprehensive Assessments-Series II (MCA-II), they describe the level of student achievement. The four levels for the MCA-II are: D—*Does Not Meet the Standards,* P—*Partially Meets the Standards,* M—*Meets the Standards* and E—*Exceeds the Standards.*

P-Value

A classic item-difficulty index that indicates the proportion of all students who answered a question correctly.

Reliability

The consistency of the results obtained from a measurement.

Reliability Coefficient

A mathematical index of consistency of results between two measures, expressed as a ratio of true-score variance to observed-score variance. As reliability increases, this coefficient approaches unity.

Standard Error of Measurement

Statistic that expresses the unreliability of a particular measure in terms of the reporting metric. Often used incorrectly (Dudek, 1979) to place score bands or error bands around individual student scores.

Test-Centered Standard Setting Methods

A type of process used to establish performance standards that focus on the content of the test itself. A more general classification of some judgmental standard setting procedures.

Test-Retest Reliability Estimate

A statistic that represents the correlation between scores obtained from one measure when compared with scores obtained from the same measure on another occasion.

Test Specifications

A detailed description of a test that helps describe the content and process areas to be covered and the number of items addressing each. The test specifications are a helpful tool for developing tests and documenting content-related validity evidence.

True Score

The piece of an observed student score that is not influenced by error of measurement. The true score is used for convenience in explaining the concept of reliability and is unknowable in practice.

Validity

A psychometric concept associated with the use of assessment results and the appropriateness or soundness of the interpretations regarding those results.

Annotated Table of Contents

The Minnesota Department of Education (MDE) is committed to responsibly following generally accepted professional standards when creating, administering, scoring, and reporting test scores. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 2014) is one source of professional standards. As evidence of our dedication to fair testing practices, the table of contents for this manual is annotated below for the *Standards.*

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Joint Technical Committee. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

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

Antalek, E. E. (2005). The relationships between specific learning disability attributes and written language: A study of the performance of learning disabled high school subjects completing the TOWL-3 (Doctoral dissertation, Clark University, 2005). *Dissertation Abstracts International, 65*(11), 4098.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6,* 75–77.

Bennett, R. E., Rock, D. A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21*(3), 9–21.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Browder, D. M., Gibbs, S., Ahlgrim-Delzell, L., Courtade, G., Mraz, M., & Flowers, C. (2009). Literacy for students with significant cognitive disabilities: What should we teach and what should we hope to achieve? *Remedial and Special Education, 30,* 269–282.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows. Lincolnwood, IL: Scientific Software International.

Cizek, G. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Erlbaum.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11,* 93–103.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Hambleton, R., & Plake, B. (1997). *An anchor-based procedure for setting standards on performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93*(2), 113–125.

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36*(1), 39–47.

Hollenbeck, K., Tindal, G., Harniss, M., & Almond, P. (1999). *The effect of using computers as an accommodation in a statewide writing test.* Eugene, OR: University of Oregon Research, Consultation, and Teaching Program.

Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). *Handwritten versus word-processed statewide compositions: Do judges rate them differently?* Eugene, OR: University of Oregon Research, Consultation, and Teaching Program.

Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education, 17*(1), 39–57.

Idstein, B. E. (2003). Dictionary use during reading comprehension tests: An aid or a diversion? (Doctoral dissertation, Indiana University of Pennsylvania, 2003). *Dissertation Abstracts International, 64*(02), 483.

Jaeger, R. M., (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, Winter,* 16–20.

Johnson, E. S., Kimball, K., & Brown, S. O. (2001). American sign language as an accommodation during standards-based assessments. *Assessment for Effective Intervention, 26*(2), 39–47.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum Associates.

Kolen, M. J. (2004). POLYEQUATE [Computer Software]. Iowa City, IA: The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky.* Los Angeles, CA: Center for the Study of Evaluation (CRESST), UCLA.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring.* Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linacre, J. M. (2006). *A user's guide to WINSTEPS.* Chicago, IL: MESA Press.

Linn, R. L. (1993). Linking results in distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement in assessment and teaching* (7th ed.). New Jersey: Prentice-Hill.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

MacArthur, C. A., & Cavalier, A. R. (1999). Dictation and speech recognition technology as accommodations in large-scale assessments for students with learning disabilities. Newark, DE: Delaware Education Research and Development Center, University of Delaware.

MacArthur, C. A., & Cavalier, A. R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43–58.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Messick, S. (1988). The once and future issues in validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer-Verlag.

Ray, S. (1982). Adapting the WISC-R for deaf children. *Diagnostique, 7,* 147–157.

Raymond, M., & Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–158). Mahwah, NJ: Erlbaum.

Reckase, M. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Mahwah, NJ: Erlbaum.

Robinson, G., & Conway, R. (1990). The effects of Irlen colored lenses on students' specific reading skills and their perception of ability: A 12-month validity study. *Journal of Learning Disabilities, 23*(10), 589–596.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13). Retrieved from http://pareonline.net/pdf/v10n13.pdf

Smisko, A., Twing, J. S., & Denny, P. L. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education, 13*(4), 333–342.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58–75.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39–49.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum.

Tindal, G., Heath, B,. Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*(4), 439–450.

Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *The Journal of Special Education, 42,* 241–254.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wetzel, R., & Knowlton, M. (2000). A comparison of print and Braille reading rates on three reading tasks. *Journal of Visual Impairment and Blindness, 94*(3), 1–18.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Zentall, S., Grskovic, J., Javorsky, J., & Hall, A. (2000). Effects of noninformational color on the reading test performance of students with and without attentional deficits. *Diagnostique, 25*(2), 129–146.

Zieky, M. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.