Technical Manual for Minnesota's Title I and Title III Assessments

For the Academic Year 2010–2011

Table of Contents

Purpose

This technical manual provides information about the development and measurement characteristics of the Minnesota Assessment System. It is organized into two parts: (a) chapters providing general information about the measurement process and (b) yearly appendices providing the specific data for a given year. The text outlines 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. Each year a new Yearbook is added.

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

A teacher constructing a test to provide immediate feedback on instruction desires the best and most accurate assessment possible, but typically does not identify the technical measurement properties of the test before or after using it in the classroom. However, a large-scale standardized assessment does require evidence to support the meaningfulness of inferences made from the scores on the assessment (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.

The manual does not include all of the information available regarding the assessment program in Minnesota. Additional information is available on the Minnesota Department of Education (MDE) website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing. Questions may also be directed to the Division of Research and Assessment at MDE by email: mde.testing@state.mn.us.

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, 1999) is one source of professional standards. As evidence of our dedication to 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 (NCLB) Act 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 2005–06, all students must take annual reading and mathematics tests in grades 3–8 and once during high school. By 2007–08, 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 1996–97 school year" (Minn. Stat. §120B.30.7c). The Minnesota Department of Education (MDE) developed a set of test specifications to measure the minimal skills needed to be successful in the workforce: this was the basis for the Minnesota Basic Skills Test (BST), the first statewide diploma test. To meet the requests for 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 measure 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). All Minnesota students in those grades were tested annually with a single statewide test by grade and subject for system accountability beginning in the following year.

1998

MDE developed the Minnesota Comprehensive Assessments (MCAs) to serve 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). 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 grade 3 and 5 students have been tested annually with 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 2005–06, as a 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. The information from these tests is 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 will 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 measures the writing, reading and mathematics proficiency of high school students. By requiring high school graduates to reach a specified level on each of these assessments, Minnesota is making sure its students are on track to have the essential skills and knowledge necessary for graduation in the 21st century. Students in 8th grade in 2005–2006 and after must pass the GRAD Written Composition given in 9th grade, Reading GRAD given in 10th grade and Mathematics GRAD given in 11th grade or retest opportunities. 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 10th grade 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 11th grade 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 on 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.

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 Mathematics MCA-III either on computer or on paper. The computer version included technology-enhanced items. Grades 5 to 8 of Mathematics MCA-Modified were computer delivered. Mathematics MCA-III, grades 5 to 8 of Mathematics MCA-Modified, and grades 3 to 8 of Mathematics MTAS assessed the *2007 Minnesota K–12 Academic Standards in Mathematics*

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

TABLE 1.1. Minnesota Assessment System Chronology

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 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 MCA-III Mathematics items • Field test of Mathematics and Reading Minnesota Comprehensive Assessments-Modified • Lexile linking study

2010-11	• First operational administration of Mathematics MCA-III • Districts given choice of computer or paper delivery of Mathematics MCA-III • First operational administration of Mathematics and Reading MCA-Modified

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 section120B.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. 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 expires on June 30, 2014.
> (Minn. Stat. §120B.365)

Name	Position	Organization
vacant	Superintendent	
vacant	Superintendent	
Jacki McCormack	Parent	Arc of Greater Twin Cities
vacant	Parent	
vacant	Teacher	
vacant	Teacher	

Paul Carney	Higher Education	Fergus Falls Community College
Paul Halverson	Business Community	IBM
Sandra G. Johnson	Higher Education	St. Cloud State University
Vacant	Public	
Vacant	Public	

Human Resources Research Organization (HumRRO)

HumRRO is a separate vendor working with MDE and completes quality assurance checks associated with elements of the Minnesota Assessment System and accountability program. They, in collaboration with MDE and Pearson, Minnesota's testing contractor, conduct quality checks during test calibrations and equating of data for various assessments, including MCA-II and GRAD.

Local Assessment and Accountability Advisory Committee

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

Name	Position	Organization
Sherri Dahl	District Assessment Coordinator, Title I	Red Lake Schools
Matthew Mohs	Director of Title I/Funded Programs	St. Paul Public Schools
Barb Muckenhirn	Director of Teaching and Learning	Princeton Public Schools
Jim Angermeyr	Director of Research, Evaluation and Assessment	Bloomington Public Schools
Johnna Rohmer-Hirt	District Research, Evaluation and Testing Achievement Analyst	Anoka-Hennepin Public Schools
Mary Roden	Coordinator of Assessment and Evaluation	Mounds View Public Schools
Richard Spicuzza	Director of Research, Evaluation and Assessment	South Washington County Public Schools
Justin Treptow	Assistant Principal	Minnesota Virtual Academy High School
Lori Zimmerman	ALC Principal	North Branch Public Schools

Minnesota Department of Education

The Research and Assessment (R&A) Division of MDE 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, MDE R&A staff, in collaboration with an outside vendor, conduct quality control activities for every aspect of the development and administration of the assessment program. R&A 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 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 from these committee meetings. Items are field-tested and Minnesota educator committees are convened to review each item and its associated data for appropriateness for inclusion in the item banks 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 website (http://education.state.mn.us > MDE > Accountability Programs > Assessment and Testing > Advisory > Advisory Panels).

Minnesota's Testing Contractor(s)

Pearson has been an MDE contractor for the statewide assessment program since November 1997. In 2005, Pearson became the primary contractor for all Minnesota assessments. Pearson has the responsibility to distribute and collect all materials as well as to maintain the security for tests. In addition, Pearson produces ancillary testing materials, including test administration manuals and interpretive guides, report folders, scannable identification sheets, packing lists, report samples, report order forms, return shipping labels, freight bills and security forms. Pearson scores all student tests forms, including conducting scoring by humans for written composition and constructed-response items, and prepares and distributes standard and optional reports.

Pearson also performed the standard setting procedure for both Reading GRAD and Mathematics GRAD. The Reading GRAD standard setting took place February 13, 2008. The Mathematics GRAD standard setting took place May 26, 2009.

Pearson conducted standard setting procedures for Mathematics MCA-III and Reading MCA-Modified on June 27 – June 29, 2011. Also, Pearson conducted standard setting procedures for Mathematics MCA-Modified and grades 3 to 8 of Mathematics MTAS on June 29 – June 30, 2011.

National Technical Advisory Committee

The National Technical Advisory Committee (TAC) serves as an advisory body to MDE. They provide recommendations on technical aspects of large-scale assessment, which includes 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.

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. Lloyd Komatsu	Assessment and Evaluation Coordinator	Forest Lake Area Schools

State Assessment 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.

Name	Position	Organization
Andrew Baldwin	Director of Technology	South Washington County Schools
Tina Clasen	District Technology Supervisor	Roseville Public Schools
Joanne Frei	District Tech for Online Testing	Osseo 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
Sharon Mateer	District Assessment Coordinator	Anoka-Hennepin Public Schools
Marcus Milazza	District Technology Coordinator	Prior Lake-Savage Area Schools
Ed Nelson	IT Services	South St. Paul 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
Chip Treen	District Technology Coordinator	North Branch Public Schools
Jim Varian	Technology Director	Big Lake Schools Public Schools
Luke Vethe	Research, Evaluation and Testing Technology Support Technician	Anoka-Hennepin Public Schools
Rennie Zimmer	District Assessment Coordinator	St. Paul Public Schools

Minnesota Assessment System

The Minnesota Department of Education (MDE) provides general information about statewide assessments at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/index.html. This website includes such documentation as:

- testing schedules;
- rubrics and descriptions of students at various levels of mathematics, reading and writing proficiency;
- test specifications and technical manuals;
- information for parents.

The No Child Left Behind Act (NCLB) 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.

Title I Assessments

The 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.

Mathematics

Minnesota Comprehensive Assessments-Series II

The Mathematics Minnesota Comprehensive Assessments-Series II (MCA-II) is a paper-based exam given in grade 11. On the test, students respond to questions involving mathematical problem solving. They answer questions about concepts and skills in four different mathematics strands: (a) number sense; (b) patterns, functions, and algebraic thinking; (c) data, statistics, and probability; and (d) spatial sense, geometry, and measurement. There are 55 scored items on the test. Test questions are in multiple-choice and gridded-response formats. The test can be administered in four separate segments that may be given on different days.

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 is given in grades 3–8 beginning in spring 2011. On the test, students respond to questions involving mathematical problem solving. They answer questions about concepts and skills in four different mathematics strands: (a) number sense; (b) patterns, functions, and algebraic thinking; (c) data, statistics, and probability; and (d) spatial sense, geometry, and measurement. The 2011 administration included 50 scored items. The Mathematics MCA-III exams include technology dependent item types similar to the items given on the on the Science MCA-II test. These item types allow measurement of higher-level thinking and concepts. In 2011, districts were given the choice of administrating an alternate paper version of the test, in lieu of the computer version. The paper version included 50 scored multiple choice and (in grades 5 and above) gridded response items.

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 mathematics problem solving. The Mathematics MTAS has been aligned with the Minnesota Academic Standards. The mathematical content strands are the same as those tested by the Mathematics MCA-III in grades 3-8 and by the Mathematics MCA-II in grade 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.

Minnesota Comprehensive Assessments-Modified

The Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) is an alternate assessment, based on modified achievement standards, that became operational in spring 2011. The MCA-Modified is an assessment designed to provide increased access to grade-level assessment tasks for students with disabilities. MCA-Modified will be used in grades 5-8 and high school. The MCA-Modified is intended to address the needs of students for whom neither the MCA nor the MTAS is an appropriate assessment choice. The Eligibility Requirements for the MCA-Modified list the criteria students must meet before IEP teams can consider the MCA-Modified as an option. Briefly, students who are eligible to take the MCA-Modified:

- Have an IEP that includes goals based on grade-level standards (often referred to as a “standards-based IEP”)
- Have access to grade-level instruction
- Performed in the “Does Not Meet the Standards” achievement level on the MCAs for the two previous consecutive years, or have scored in the proficient range or above on the MTAS
- Are not expected to meet grade-level standards in the year for which the MCA-Modified is identified as the appropriate assessment

Grades 5-8 of MCA-Modified have 35 scored multiple choice test questions delivered via computer. Grade 11 is a paper-based assessment with 40 multiple choice questions.

Reading

Minnesota Comprehensive Assessments-Series II

The Reading MCA-II is a paper-based exam given in grades 3–8 and 10. On the test, students read expository and poetry passages. The students answer questions about concepts and skills in three sub-strands of reading: (a) vocabulary expansion, (b) comprehension and (c) literature. The number of scored items (40–50) varies by grade, with test questions in multiple-choice formats. The exam is administered in four 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 Academic Standards. The reading content strands are the same as tested by the Reading MCA-II 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 and high school by using grade- and age-appropriate vocabulary, subject matter and increases in word count and length. The performance tasks can be administered on different days according to the needs of the student.

Minnesota Comprehensive Assessments-Modified

The Reading Minnesota Comprehensive Assessments-Modified (MCA-Modified) is an alternate assessment, based on modified achievement standards, that became operational in spring 2011. The MCA-Modified is an assessment designed to provide increased access to grade-level assessment tasks for students with disabilities. MCA-Modified will be used in grades 5-8 and high school. The MCA-Modified is intended to address the needs of some students for whom neither the MCA nor the MTAS is an appropriate assessment choice. The Eligibility Requirements for the MCA-Modified list the criteria students must meet before IEP teams can consider the MCA-Modified as an option. Briefly, students who are eligible to take the MCA-Modified:

- Have an IEP that includes goals based on grade-level standards (often referred to as a “standards-based IEP”)
- Have access to grade-level instruction
- Performed in the “Does Not Meet the Standards” achievement level on the MCAs for the two previous consecutive years, or have scored in the proficient range or above on the MTAS
- Are not expected to meet grade-level standards in the year for which the MCA-Modified is identified as the appropriate assessment

MCA-Modified is a paper-based assessment with 35 scored multiple choice test questions.

Science

Minnesota Comprehensive Assessments-Series II

The computer-delivered Science MCA-II test is administered to students in grades 5, 8 and high school. In the grade 5 and 8 tests, students answer questions about concepts and skills in three strands of science: (a) physical, (b) earth and (c) life. The high school test is a life-science-only test given at the end of the school year to students in grades 10–12 who completed a biology course during the academic year. These assessments include multiple-choice, constructed-response and figural-response items. Figural-response items are designed to allow students to respond by selecting one or more points on, or moving objects around within, a graphic image.

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 tested by the Science MCA-II 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.

TABLE 1.2. Title I Accountability Tests in 2010–11

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

Title III Assessments

The assessments are used to evaluate school and district success toward Annual Measurable Achievement Objectives (AMAOs) related to Title III of NCLB. They also serve as demonstration of proficiency for district funding for EL programming by the state. All EL are required to take the Test of Emerging Academic English (TEAE) and Minnesota Student Oral Language Observation Matrix (MN SOLOM) assessments.

Reading and Writing

Test of Emerging Academic English

The TEAE is a developmental test of reading and writing for EL in grades 3–12. Students participating in the TEAE represent a wide range of proficiency levels. The TEAE includes tasks designed to measure

[1] The high school Science MCA-II is given to students in the year they complete their instruction in life science. Students who were in grade 10 in 2007–08 are required to take the high school Science MCA-II before the end of their high school career.

emerging language proficiency as well as tasks designed to challenge students at the upper levels of proficiency. Tasks and items are structured so that there is some material in every test appropriate for students at each level of English Language Proficiency (ELP). For EL in grades K–2, teachers complete a reading and writing checklist.

The tests are administered in four grade-level spans: grades 3–4, grades 5–6, grades 7–8 and grades 9–12. All students within a grade span take the same test form. For example, all students in grades 3 and 4 who are eligible to participate in the TEAE complete one test form and all students in grades 9–12 are administered the same test form for their grade band. The number of Yes-No scored items on the TEAE Reading is 125 for grades 3–8 and 150 for high school, with 25 items per reading prompt. The reading exam is generally administered in one testing session, but it can be administered in segments, if necessary, for students with individualized education programs (IEPs) or 504 plans. The TEAE Writing consists of two writing prompts, one picture prompt and one text prompt, to which students write short narrative responses.

Listening and Speaking

Minnesota Student Oral Language Observation Matrix

The MN SOLOM was adapted from the SOLOM developed by the San Jose Unified School District. Six teacher ratings for each student are recorded, two for listening and four for speaking. Each item is rated using a rubric on which teachers evaluate listening and speaking skills for EL in grades K–12. The MN SOLOM rubric is bundled with TEAE test documents.

Federal Title III legislation requires progress of EL in reading and writing proficiency be reported for all grades.

TABLE 1.3. Title III Accountability Tests in 2010–11

Test	Subject	Grades
TEAE	Reading and Writing	3–4, 5–6, 7–8, 9–12
K–2 checklist	Reading and Writing	K–2
MN SOLOM	Listening and Speaking	K–12

Diploma Assessments

To be eligible for a Minnesota diploma, students must meet local school requirements and receive passing scores on a Minnesota graduation test for mathematics, reading and writing (see table 1.4). Students can retake these tests in order to attain passing scores for diploma eligibility. Currently, Minnesota is phasing out one set of tests, the Basic Skills Tests (BST) and implementing a new set of tests, the Graduation-Required Assessments for Diploma (GRAD). The BST is for students who entered grade 8 in 2004–05 or earlier. Students entering grade 8 in 2005–06 or later take the GRAD. The first administration of the GRAD Written Composition occurred in 2007. Mathematics and Reading GRAD assessments are done simultaneously with the MCA-II, allowing two routes to graduation: proficiency on the MCA-II or a passing score on the GRAD.

Technical information on the GRAD and BST is found in their respective technical manuals.

Mathematics

Minnesota Comprehensive Assessments-Series II

The Mathematics MCA-II is a paper-based exam given in grade 11. On the test, students respond to questions involving mathematical problem solving. They answer questions about concepts and skills in four different mathematics strands: (a) number sense; (b) patterns, functions, and algebraic thinking; (c) data, statistics, and probability; and (d) spatial sense, geometry, and measurement. Test questions are in multiple-choice, constructed-response and gridded-response formats. The test can be administered in four separate segments that may be given on different days.

Students who achieve *Meets the Standards* or above on the Mathematics MCA-II have met their graduation assessment requirement for mathematics.

Graduation-Required Assessment for Diploma

The Class of 2010, the first group of students required to pass the series of GRAD tests, took the grade 11 Mathematics GRAD in 2009. The first administration of Mathematics GRAD is embedded within the grade 11 Mathematics MCA-II.

Minnesota Comprehensive Assessments-Modified

The Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) is an alternate assessment, based on modified achievement standards, that became operational in spring 2011. The MCA-Modified is an assessment designed to provide increased access to grade-level assessment tasks for students with disabilities in grade 11. The MCA-Modified is a paper-based assessment with 40 multiple choice questions and is intended to address the needs of students for whom neither the MCA nor the MTAS is an appropriate assessment choice. The Eligibility Requirements for the MCA-Modified list the criteria students must meet before IEP teams can consider the MCA-Modified as an option. Briefly, students who are eligible to take the MCA-Modified:

- Have an IEP that includes goals based on grade-level standards (often referred to as a “standards-based IEP”)
- Have access to grade-level instruction
- Performed in the “Does Not Meet the Standards” achievement level on the MCAs for the two previous consecutive years, or have scored in the proficient range or above on the MTAS
- Are not expected to meet grade-level standards in the year for which the MCA-Modified is identified as the appropriate assessment

Students who achieve *Meets the Modified Achievement Standards* or above on the Mathematics MCA-Modified have met their graduation assessment requirement for mathematics.

Minnesota Test of Academic Skills

The Mathematics MTAS test is given in grade 11. This test contains a set of nine performance tasks designed to measure mathematics problem solving. Mathematics MTAS has been aligned with the Minnesota Academic Standards. The mathematical content strands are the same as those tested by the Mathematics MCA-II 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.

Students who achieve *Meets the Alternate Achievement Standards* or above on the Mathematics MTAS have met their graduation assessment requirement for mathematics. If established by their IEP team, students may also demonstrate a Pass Individual score on the MTAS to meet their graduation assessment requirement for mathematics.

Basic Skills Test

To be eligible for a diploma from a Minnesota public high school, students who entered grade 8 in 2004–05 or earlier must receive a passing score on the Mathematics Basic Skills Test (BST).

Reading

Minnesota Comprehensive Assessments-Series II

The Reading MCA-II is a paper-based exam given in grade 10. On the test, students read expository and poetry passages. The students answer questions about concepts and skills in three sub-strands of reading: (a) vocabulary expansion, (b) comprehension and (c) literature. The exam is administered in four separate segments that may be given on different days.

Students who achieve *Meets the Standards* or above on the Reading MCA-II have met their graduation assessment requirement for reading.

Graduation-Required Assessment for Diploma

The Class of 2010, the first group of students required to pass the series of GRAD tests, took the grade 10 Reading GRAD in 2008. The first administration of Reading GRAD is embedded within the grade 10 Reading MCA-II.

Minnesota Comprehensive Assessments-Modified

The Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) is an alternate assessment, based on modified achievement standards, that became operational in spring 2011. The MCA-Modified is an assessment designed to provide increased access to grade-level assessment tasks for students with disabilities in grade 10. The MCA-Modified is a paper-based assessment with 35 multiple choice questions and is intended to address the needs of students for whom neither the MCA nor the MTAS is an appropriate assessment choice. The Eligibility Requirements for the MCA-Modified list the criteria students must meet before IEP teams can consider the MCA-Modified as an option. Briefly, students who are eligible to take the MCA-Modified:

- Have an IEP that includes goals based on grade-level standards (often referred to as a "standards-based IEP")
- Have access to grade-level instruction
- Performed in the "Does Not Meet the Standards" achievement level on the MCAs for the two previous consecutive years, or have scored in the proficient range or above on the MTAS
- Are not expected to meet grade-level standards in the year for which the MCA-Modified is identified as the appropriate assessment

Students who achieve *Meets the Modified Achievement Standards* or above on the Reading MCA-Modified have met their graduation assessment requirement for reading.

Minnesota Test of Academic Skills

The Reading MTAS test is given in grade 10. Each test comprises a set of nine performance tasks designed to measure student understanding of text. The Reading MTAS has been aligned with the Minnesota Academic Standards. The reading content strands are the same as those tested by the Reading MCA-II 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 performance tasks can be administered on different days.

Students who achieve *Meets the Alternate Achievement Standards* or above on the Reading MTAS have met their graduation assessment requirement for reading. If established by their IEP team, students may also demonstrate a Pass Individual score on the MTAS to meet their graduation assessment requirement for mathematics.

Basic Skills Test

To be eligible for a diploma from a Minnesota public high school, students who entered grade 8 in 2004–05 or earlier must receive a passing score on the Reading Basic Skills Test (BST).

Writing

Graduation-Required Assessment for Diploma

The Class of 2010, the first group of students required to pass the series of GRAD tests, took the grade 9 GRAD Written Composition in 2007.

Writing Alternate Assessment

The writing alternate assessment from past years will continue to be used in 2010-11 with students in grade 9.

Students who receive a valid score on the writing alternate assessment have met their graduation assessment requirement for writing.

Basic Skills Test

To be eligible for a diploma from a Minnesota public high school, students who entered grade 8 in 2004–05 or earlier must receive a passing score on the Written Composition Basic Skills Test (BST).

TABLE 1.4. Diploma Tests in 2010–11

Test	Subject	Initial Grade	Retest Grade(s)
BST	Reading, Mathematics	—	11–12
BST	Written Composition	—	11–12
GRAD	Written Composition	9	10–12
MCA-II/GRAD	Reading	10	11-12
MCA-II/GRAD	Mathematics	11	12

Chapter 2: Test Development

The test development phase of each Minnesota Assessment includes a number of 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. (www.achieve.org),
- Recommended Standards for Information and Technology Literacy from the Minnesota Educational Media Organization (MEMO -- www.memoweb.org), and
- 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 II and Series III (MCA-II & MCA-III), Minnesota Comprehensive Assessments-Modified (MCA-Modified), Minnesota Test of Academic Skills (MTAS) and Test of Emerging Academic English (TEAE):

- *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

- *Development of Items and Tasks.* Using the *Standards* and test specifications, the Minnesota Department of Education (MDE) Division of Research and Assessment (R&A) staff and Minnesota's testing contractor work with the item development contractor to develop items and tasks.

- *Item Content Review.* All members of the assessment team review the developed items, discuss possible revisions and make changes when necessary.

- *Item 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.
- *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.
- *Data Review.* Committees review the items in light of the field-test data and make recommendations regarding the inclusion of the items in the available item pool.
- *New Form 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 the domain comprises 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, sub-strands 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 (e.g., 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 independent of the item formats and difficulty of the content; this runs counter to the perception of many people that cognitive level and content difficulty are equivalent concepts. For example, a benchmark measured at a high cognitive level could be assessed with any item format: multiple-choice, drag-and-drop, constructed-response or gridded-response.

Similarly, the ranges in number of items and content limits are based on discussion among the expert teachers in the committees about the emphasis a benchmark has in the classroom and type of curriculum content regularly taught to students in the 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 show MDE's respect for teachers' concerns about the time students spend taking tests, and take into account the grade and age of students involved as well as various 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 some details about the development of test specifications for each test in the Minnesota Assessment System.

Title I Assessments

Minnesota Comprehensive Assessments-Series II

To develop the Minnesota Comprehensive Assessments-Series II (MCA-II), MDE held meetings with Minnesota educators to define general test specifications for each grade and subject area during 2004–05. 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, sub-strands and benchmarks of the Minnesota Academic Standards to be measured in the tests. Some strands, sub-strands or benchmarks were not suitable for the large-scale assessments (for example, the requirement to read aloud fluently in the grade 3 reading standard benchmarks). 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 suggestions were incorporated into the final versions of the test specifications. The complete MCA-II test specifications document is available on the MDE website at http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TestSpecs/index.html.

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 suggestions were incorporated into the final versions of the test specifications. The complete MCA-III test specifications document is available on the MDE website at http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TestSpecs/index.html

Minnesota Comprehensive Assessments-Modified

To develop the Minnesota Comprehensive Assessments-Modified (MCA-Modified), MDE used the test specifications developed for the MCA-II and MCA-III. Alternate assessments based on modified achievement standards (AA-MAS) must measure the same academic standards as the assessments designed for the general population. In its test design for the MCA-Modified, MDE replicated to the extent possible the percentage of items distributed across strands, substrands and standards found in the test specifications for the MCA–II and III.

Item formats and cognitive levels for the MCA-Modified were defined, in part, through MDE's participation in a consortium of states funded to develop an AA-MAS through a General Supervision Enhancement Grant. MDE also referred to other states' AA-MAS development work to identify effective ways to render content more accessible to the population of students who qualify for this assessment.

Committees of Minnesota educators reviewed drafts of these specifications and suggestions were incorporated into the final versions of the test specifications. The MCA-Modified test specifications documents are 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 of 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:

- The grade-level benchmark had been selected for the MCA-II, or in the case of grades 3-8 Mathematics MTAS, MCA-III.
- Proficiency on the benchmark will aid future learning in the content area for students with significant cognitive disabilities.
- 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).
- 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 sub-strands, 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. At these meetings, the teachers made changes to a subset of the recommended benchmarks in reading, mathematics and science.

Content limits had been written and approved for the MCA-II or 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, then 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 then 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 this range of students has access to the test items, student communication modalities were considered and accommodations 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 began to write entry points, in essence extending the benchmark for the students taking the MTAS for each of the selected benchmarks included on the MTAS. The process included the following steps:

- A curriculum specialist described the intent or underlying essence of the benchmark
- A general educator described a classroom activity or activities in which the benchmark could be taught
- A special educator then 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 during December 2006. Test specifications for grades 3-8 Mathematics MTAS were published in 2011 in order to update the assessment to align to the *2007 Minnesota K–12 Academic Standards in Mathematics.* 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

Test of Emerging Academic English

Test specifications for the Test of Emerging Academic English (TEAE) Reading and Writing were first developed in 2000 with the aid of a previous testing vendor, MetriTech, Inc. The design of the reading and writing test was based largely on a test created by MetriTech called the Language Proficiency Test Series; the test specifications are similar to this test. Committees of Minnesota teachers refined and adapted the original test to fit Minnesota needs. Test specifications are available on the MDE website at http://education.state.mn.us/mdeprod/groups/Assessment/documents/FAQ/000431.pdf.

Minnesota Student Oral Language Observation Matrix

The Minnesota Student Oral Language Observation Matrix is an observational assessment adapted from the Student Oral Language Observation Matrix (SOLOM) developed by the California State Department of Education and San Jose (California) Unified School District. It is now a document of the public domain. It was adapted by MDE in 2003. The observation protocols and rubrics are available on the MDE website at http://education.state.mn.us/mdeprod/groups/Assessment/documents/Manual/000424.pdf.

Item Development

This section describes the item writing process used during the development of test items 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 an involved, multi-stage process. Figure 2.1 lists the steps that Minnesota's testing contractor follows in bringing an item or task from its initial creation to being ready for placement on an operational exam.

Items and tasks are written and internally reviewed at the testing contractor before submission to MDE for review by content committees. For each subject and grade, MDE receives an item tally sheet displaying the number of test items submitted 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 resulted from committee work, 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 II

The content limits associated with the Minnesota Comprehensive Assessments-Series II (MCA-II) are item-specific specifications. They identify the boundaries of context under which an item may be developed. In mathematics, for example, the magnitude of the numerals used in an item may be restricted so that the item is appropriate to the standards and the grade level of the test. In reading, this could involve further clarification of what background knowledge from outside the text is necessary to make an appropriate inference. In science, this may be a list of the tools that can be used in an item to engage in scientific inquiry.

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 of the grades that follow (i.e., grades 4 through 8).

Minnesota Comprehensive Assessments-Modified

The Minnesota Comprehensive Assessments-Modified follow the same content limits or item specification guidelines as their Minnesota Comprehensive Assessments-Series III (MCA-III) counterparts. The same content is covered in the MCA-Modified but with less difficult questions. To meet the goal to design a test that is accessible yet challenging for the population of students whose disability has prevented them from attaining grade-level proficiency, several design modifications have been made. In addition to the issues outlined and referenced above, the following guidelines are used for items on the MCA-Modified.

1. Items will be written using language simplification principles.
2. Page and item layout will focus on a simplified design.
3. All items are three-option multiple-choice questions.
4. Key words are presented in boldface in some items to help students identify the main task to be completed in the item.

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 in a multi-step solution to a problem. In reading, this could involve a restriction in the number of literary terms assessed within a benchmark. In science, this might address knowledge of only seminal aspects of the water cycle.

Title III Assessments

Test of Emerging Academic English

The TEAE is a static assessment no longer in active development. Its test specifications do not include content limits.

Minnesota Student Oral Language Observation Matrix

The MN SOLOM is an observational performance assessment in listening and speaking. Its test specifications do not include content limits.

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 II and 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 only retained to submit items based on their competency in this training.

Minnesota Comprehensive Assessments-Modified

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 only retained to submit items based on their competency in this training. For the MCA-Modified, the testing contractor ensures that item writers have experience with and clear understanding of the unique population of students who take the MCA-Modified.

Minnesota Test of Academic Skills

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 only retained to submit items based on their competency in this training. For the MTAS, the testing contractor ensures that item writers have experience with and clear understanding of the unique needs of students with significant cognitive disabilities with respect to their ability to interact and 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 and 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

Test of Emerging Academic English

The TEAE is a static assessment no longer in active development. It does not have current item writers.

Minnesota Student Oral Language Observation Matrix

The MN SOLOM is not commercially published. It was originally developed as the Student Oral Language Observation Matrix (SOLOM) by the San Jose Area Bilingual Consortium and has undergone revisions with leadership from the Bilingual Education Office of the California Department of Education. It is within the public domain and can be copied, modified, or adapted to meet local needs. Minnesota uses the assessment as a static, observational performance assessment no longer in active development. It does not have current item writers.

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 II and Series III

Minnesota's testing contractor conducts comprehensive item writer training for all persons selected to submit items for the MCA-II or MCA-III. Training includes an overview of the test development cycle and very 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 Comprehensive Assessments-Modified

Minnesota's testing contractor conducts comprehensive item writer training for all persons selected to submit items for the MCA-Modified. Training includes an overview of the test development cycle and very 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 specific 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 the following areas:

- 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 Pearson;

- 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;
- lend themselves to use with assistive technology and other accommodations.

Title III Assessments

Test of Emerging Academic English

The TEAE is a static assessment no longer in active development. It does not have current item writer training.

Minnesota Student Oral Language Observation Matrix

The MN SOLOM is a static assessment no longer in active development. It does not have current item writer training.

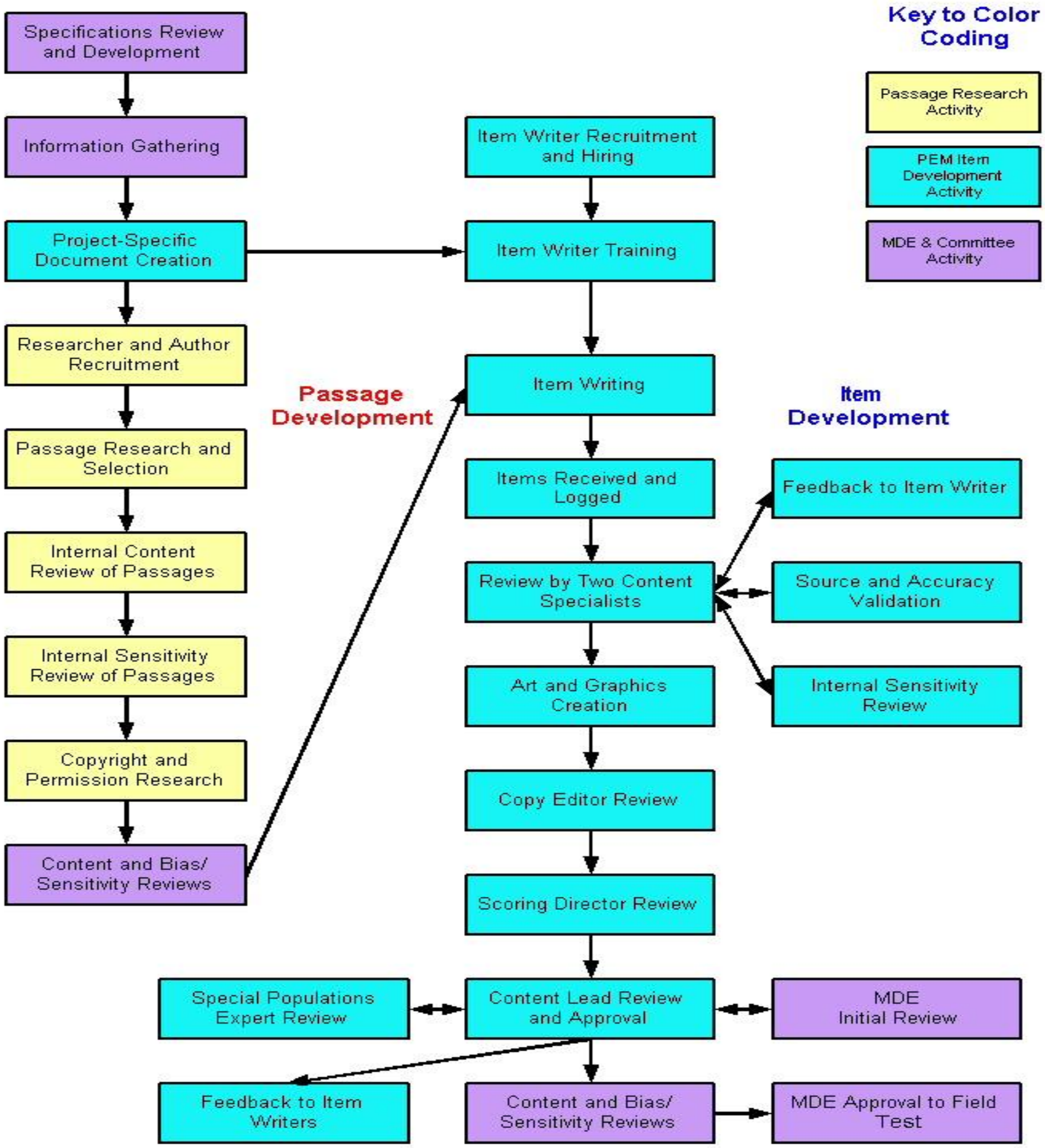


Figure 2.1. Item Development Process

Item Review

Contractor Review

Experienced testing contractor staff members, as well as content experts in the grades and subject areas for which the items 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 of the items, 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 II

Reading passages eligible for placement on the Minnesota Comprehensive Assessments-Series II (MCA-II) are those that adhere to the principles of Universal Design. In accordance with the principles of Universal Design, passages that rely heavily on visual imagery are not considered appropriate. All passages must be able to be Brailled and formatted for large print without compromising important ideas or inhibiting comprehension of the passage. In addition, passages must avoid the use of idioms, regional colloquialisms and other word choices that may be unfamiliar to English Learners (ELs) in order to avoid placing these students at a disadvantage during testing.

Reading passages must be accessible to the widest range of students, thereby allowing all examinees the opportunity to demonstrate their knowledge of the tested content standards. Therefore, reading passages are chosen based on their potential to measure the reading and/or language arts content standards for Minnesota and support the development of quality test items. There are a number of characteristics that define suitable passages. Such passages are written at an appropriate level in terms of content/subject matter, vocabulary and readability for a specified grade level. The passages will be interesting and meaningful to students and reflect the cultural diversity of the state's student population. The passages represent the types of reading that students encounter in their classrooms and in their everyday lives. The passages must be capable of being understood without reliance upon classroom- or teacher-led discussions.

Before a passage, storyboard 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 (strand, sub-strand, standard and benchmark);
- is appropriate in difficulty for the grade level of the examinees;
- has only one correct or best answer (multiple-choice items);
- has an appropriate and complete scoring guideline (for constructed-response, gridded-response and figural-response 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, sub-strand, 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, constructed-response) 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 passage, storyboard and item to identify language or content that might be inappropriate or offensive to students, parents, or community members, and items that contain stereotypical or biased references to gender, ethnicity or culture. The Bias and Fairness Committee reviews each item and accepts, edits or rejects it for use in field tests.

For the computer-based Science assessment, there is an extra panel review. Once the materials have been updated based on the content and bias panel recommendations, Minnesota's testing contractor begins development of the animations needed to support the items and publishes the items in an analog of its online testing engine. An Electronic Review panel is then convened to review these materials. These panels consist of both content reviewers and bias and fairness reviewers. A separate panel is convened for each grade. The purpose of this panel is to review and provide input on the onscreen layout of the materials, including the animations. Following this review, materials are updated and resubmitted to MDE for final approval.

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, sub-strand, standard and benchmark);
- is appropriate in difficulty for the grade level of the examinees;
- has only one correct or best answer (multiple-choice items);
- has an appropriate and complete scoring guideline (for gridded-response and 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, sub-strand, 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, and items that contain stereotypical or biased references to gender, ethnicity or culture. The Bias and Fairness Committee reviews each item and accepts, edits or rejects it for use in field tests.

For the computer-based items, there is an extra panel review. Once the materials have been updated based on the content and bias panel recommendations, Minnesota's testing contractor publishes the items in an analog of its online testing engine. An Electronic Review panel is then convened to review these materials. These panels consist of both content reviewers and bias and fairness reviewers. A separate panel is convened for each grade. The purpose of this panel is to review and provide input on the onscreen layout of the materials. Following this review, materials are updated and resubmitted to MDE for final approval.

Minnesota Comprehensive Assessments-Modified

Reading passages eligible for placement on the Minnesota Comprehensive Assessments-Modified (MCA-Modified) are those that adhere to the principles of Universal Design. In accordance with the principles of Universal Design, passages that rely heavily on visual imagery are not considered appropriate. In addition, passages must avoid the use of idioms, regional colloquialisms and other word choices that may be unfamiliar to English Learners (ELs) in order to avoid placing these students at a disadvantage during testing.

Reading passages must be accessible to the widest range of students, thereby allowing all examinees the opportunity to demonstrate their knowledge of the tested content standards. Therefore, reading passages are chosen based on their potential to measure the reading and/or language arts content standards for Minnesota and support the development of quality test items. There are a number of characteristics that define suitable passages. Such passages are written at an appropriate level in terms of content/subject matter, vocabulary and readability for a specified grade level. The passages will be interesting and meaningful to students and reflect the cultural diversity of the state’s student population. The passages represent the types of reading that students encounter in their classrooms and in their everyday lives. The passages must be capable of being understood without reliance upon classroom- or teacher-led discussions.

Before a passage, storyboard 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 to assure that each item

- is an appropriate measure of the intended content (strand, sub-strand, standard and benchmark);
- is appropriate in difficulty for the grade level of the examinees;
- has only one correct or best answer.

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

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

For the computer-based assessments, there is an extra panel review. Once the materials have been updated based on the content and bias panel recommendations, Minnesota’s testing contractor publishes the items in an analog of its online testing engine. An Electronic Review panel is then convened to review these materials. These panels consist of both content reviewers and bias and fairness reviewers. A separate panel is convened for each grade. The purpose of this panel is to review and provide input on the onscreen layout of the materials. Following this review, materials are updated and resubmitted to MDE for final approval.

Minnesota Test of Academic Skills

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

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, 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-IIs. 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 mirror, to the extent appropriate, those appearing on the MCA-II and 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;
- has only one correct or best answer for each multiple-choice item.

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, sub-strand, standard and benchmark, or 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, community members, or items that contain "stereotypic" or biased references to gender, ethnicity or culture. The Bias and Fairness Committee reviews each item and accepts, edits, or rejects it 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

Test of Emerging Academic English

The Test of Emerging Academic English (TEAE) is a static assessment no longer in active development. It does not have current contractor item review.

Minnesota Student Oral Language Observation Matrix

The Minnesota Student Oral Language Observation Matrix (MN SOLOM) is a static assessment no longer in active development. It does not have current contractor item review.

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 any ethnic, gender, economic or cultural bias.

Title I Assessments

Minnesota Comprehensive Assessments-Series II and Series III

Content assessment staffs 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 Comprehensive Assessments-Modified

Content assessment staffs 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 staffs 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

Test of Emerging Academic English

The TEAE is a static assessment no longer in active development. It does not have current MDE item review.

Minnesota Student Oral Language Observation Matrix

The MN SOLOM is a static assessment no longer in active development. It does not have current MDE item review.

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 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 representing 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 the committee judges 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 II and Series III

Item review committees are composed of content teachers in English language arts, 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 a ten-point checklist (presented below) to ensure alignment to the Standards. Teachers' discussion of the test items is facilitated by MDE and its testing contractor.

<u>Item Review Checklist</u>

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 clang words used which 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 Comprehensive Assessments-Modified

Item review committees are composed of content teachers in English language arts and mathematics and special educators who have experience working with students identified as eligible for the MCA-Modified. Many content educators who served on these committees had also served on item review panels for the MCA-II or the MCA-III, and so were very familiar with the Minnesota Academic Standards.

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-II and so are 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

Test of Emerging Academic English

The TEAE is a static assessment no longer in active development. It does not have current committee item reviews.

Minnesota Student Oral Language Observation Matrix

The MN SOLOM is a static assessment no longer in active development. It does not have current committee item reviews.

Bias and Fairness Review

All items placed on Minnesota assessments are evaluated by a panel of teachers and community experts familiar with the diversity of cultures represented within 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. The Minnesota Department of Education (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 students' 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;
- percent of all students choosing each response;
- students distributed into thirds based upon performance on the overall test and that group of students' distribution choosing each response;
- percent 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;
- 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. Items that the data review committee accepts with reservation are flagged in the item bank for use only under extraordinary circumstances.

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 below, 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 to determine the quality of each item. Two types of differential item functioning (DIF; that is, 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 or not 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 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) grades 3-8 were held in March, 2010. Items reviewed at these meetings were fielded in a stand-alone on-line field test conducted in the fall of 2009. MCA-III data reviews use the procedures described above. 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 Comprehensive Assessments-Series II

The Minnesota Comprehensive Assessments-Series II (MCA-II) data reviews use the procedures described above. 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 Comprehensive Assessments-Modified

The first data review meetings for Minnesota Comprehensive Assessments-Modified (MCA-Modified) were held in July, 2010. Items reviewed at these meetings were fielded in an embedded field test conducted as part of the spring 2010 administration of Minnesota Comprehensive Assessments-Series II (MCA-II). MCA-Modified data reviews use the procedures described above.

Minnesota Test of Academic Skills

The Minnesota Test of Academic Skills (MTAS) data reviews use the procedures described above. Emphasis is placed on inviting panelists who have both 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: Developmentally Cognitively Disabled—Mild, Developmentally Cognitively Disabled—Severe, and Autism Spectrum Disorder.

Title III Assessments

Test of Emerging Academic English

The Test of Emerging Academic English (TEAE) is a static assessment no longer in active development. It does not have current data reviews.

Minnesota Student Oral Language Observation Matrix

The Minnesota Student Oral Language Observation Matrix (MN SOLOM) is a static assessment no longer in active development. It does not have current data reviews.

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) and its testing contractor maintain 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 Sub-Strand is the secondary organizational element (e.g., Patterns and Functions or Vocabulary). Each Sub-Strand 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 appropriately. This is true for other accommodations for items as well (for example, large print). 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), 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 kindergarten through Grade 12 are annually assessed with language proficiency tests.

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series II

General education students—and students in special populations (that is, ELs and students with disabilities (SWDs)) who are able to do so—take the Mathematics Minnesota Comprehensive Assessments-Series II (MCA-II) to fulfill their Title I grade 11 mathematics requirement.

Minnesota Comprehensive Assessments-Series III

General education students—and students in special populations (that is, 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 Title I grades 3-8 mathematics requirement.

Minnesota Comprehensive Assessments-Modified

The Mathematics MCA-Modified (MCA-M) is an alternate assessment based on modified achievement standards. It is designed for a small group of students whose disability has precluded them from achieving grade-level proficiency but who do not qualify to take Minnesota's other alternate assessment, the Minnesota Test of Academic Skills (MTAS). The MCA-Modified differs from the MCA-III in a few key ways, including the following:

- The student must have an IEP. The IEP team is responsible for determining, on an annual basis, how a student with a disability will participate in statewide testing. This decision-making process must start with a consideration of the general education assessment.
- The MCA-Modified may only be administered to a student who currently receives special education services, though participation in the administration is not limited to any particular disability category.
- Students must meet all eligibility requirements for the MCA-Modified before it is selected by the IEP team. Eligibility requirements for the MCA-Modified can be found below.

1. The student demonstrates persistent low performance as defined by performance at the lowest achievement level on the MCA (Does Not Meet the Standards) for the past 2 years. OR The student meets or exceeds the standards on the MTAS and the IEP team determines that the student is most appropriately assessed with the MCA-M.

2. The student has access to instruction on grade-level content standards.

3. The student has an IEP based on grade-level content standards in the content area(s) being assessed by MCA-M.

4. The IEP team determines that the student is highly unlikely to achieve proficiency on the grade-level content standards within the year the test is administered, even with specially designed instruction.

   a. Objective and valid data from multiple measures should be collected over time to confirm that the student is not likely to achieve proficiency on grade-level content standards within the year. Examples of objective and valid measures include state assessments, district-wide assessments, curriculum-based measures and other repeated measures of progress over time.

   b. Appropriate accommodations, such as assistive technology, are provided as needed on evaluations of classroom performance, and the student's accommodation needs are carefully considered before the IEP team makes a determination that the student is not likely to achieve proficiency on grade-level content standards.

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 Title I mathematics requirement.

Reading

Minnesota Comprehensive Assessments-Series II

General education students – and students with disabilities who are able to do so – take the Reading MCA-II to fulfill their Title I reading requirement.

Minnesota Comprehensive Assessments-Modified

The Reading MCA-Modified (MCA-M) is an alternate assessment based on modified achievement standards. It is designed for a small group of students whose disability has precluded them from achieving grade-level proficiency but who do not qualify to take Minnesota's other alternate assessment, the Minnesota Test of Academic Skills (MTAS). The MCA-Modified differs from the MCA-II in a few key ways, including the following:

- The student must have an IEP. The IEP team is responsible for determining, on an annual basis, how a student with a disability will participate in statewide testing. This decision-making process must start with a consideration of the general education assessment.

- The MCA-Modified may only be administered to a student who currently receives special education services, though participation in the administration is not limited to any particular disability category.
- Students must meet all eligibility requirements for the MCA-Modified before it is selected by the IEP team. Eligibility requirements for the MCA-Modified can be found below.

1. The student demonstrates persistent low performance as defined by performance at the lowest achievement level on the MCA (Does Not Meet the Standards) for the past 2 years. OR The student meets or exceeds the standards on the MTAS and the IEP team determines that the student is most appropriately assessed with the MCA-M.

2. The student has access to instruction on grade-level content standards.

3. The student has an IEP based on grade-level content standards in the content area(s) being assessed by MCA-M.

4. The IEP team determines that the student is highly unlikely to achieve proficiency on the grade-level content standards within the year the test is administered, even with specially designed instruction.

    a. Objective and valid data from multiple measures should be collected over time to confirm that the student is not likely to achieve proficiency on grade-level content standards within the year. Examples of objective and valid measures include state assessments, district-wide assessments, curriculum-based measures and other repeated measures of progress over time.

    b. Appropriate accommodations, such as assistive technology, are provided as needed on evaluations of classroom performance, and the student's accommodation needs are carefully considered before the IEP team makes a determination that the student is not likely to achieve proficiency on grade-level content standards.

Minnesota Test of Academic Skills

Students with individualized education programs (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 Title I reading requirement.

Science

Minnesota Comprehensive Assessments-Series II

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

Title III Assessments

Reading and Writing

Test of Emerging Academic English

English Learners in grades 3–12 must participate in the Test of Emerging Academic English (TEAE) and those in grades K–2 must participate in the K-2 Reading and Writing checklist.

Listening and Speaking

Minnesota Student Oral Language Observation Matrix

English Learners in grades K–12 must participate (that is, be observed) in the Minnesota Student Oral Language Observation Matrix (MN SOLOM).

Administration to Students

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series II

The grade 11 Mathematics Minnesota Comprehensive Assessments-Series II (MCA-II) is divided into four segments, so districts have the option of administering the test over two or more days. 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 four segments can be done in a number of ways: all four segments at one time, one segment per day, two segments per day, etc. Some segments of the Mathematics MCA-II do not allow calculators to be used in answering questions.

Minnesota Comprehensive Assessments-Series III

The grades 3-8 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) are offered to districts as a computer option or a paper option. The computer version is divided into eight segments, while the paper version is divided into four segments, so in both cases districts have the option of administering the test over two or more days. 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. Some segments of the Mathematics MCA-III do not allow calculators to be used in answering questions.

Minnesota Comprehensive Assessments-Modified

The grades 5-8 Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) are computer-based exams divided into two segments, and the grade 11 Mathematics MCA-Modified is a paper-based exam divided into four segments. For all exams, districts have the option of administering the test over two or more days. 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. Some segments of the Mathematics MCA-Modified do not allow calculators to be used in answering questions.

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 consider 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 II

The Reading Minnesota Comprehensive Assessment-Series II (MCA-II) is divided into four segments, so districts have the option of administering the test over two or more days. MDE allows district staff to determine how many test segments will be administered during each testing session. Administration of the four segments can be done in a number of ways: all four segments at one time, one segment per day, two segments per day, etc.

Minnesota Comprehensive Assessments-Modified

The Reading Minnesota Comprehensive Assessment-Modified (MCA-Modified) is divided into four segments, so districts have the option of administering the test over two or more days. MDE allows district staff to determine how many test segments will be administered during each testing session. Administration of the four segments can be done in a number of ways: all four segments at one time, one segment per day, two segments per day, etc.

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 MDE training (when available), 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 PowerPoint picture book, a Boardmaker representation or other accommodations appropriate for student's 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 and determine performance levels in the spring of 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-II 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 et al. 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. 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 in press article for the journal *Remedial and Special Education*, Browder et al. 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. notes, "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 consider 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 II

The Science MCA-II is a computer-delivered assessment administered in two segments. MDE allows district staff to determine how many test segments will be administered during each 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 MDE training (when available), attending a 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

Reading and Writing

Test of Emerging Academic English

The Test of Emerging Academic English (TEAE) Reading has four segments in grades 3–8 and five segments in grades 9–12. The TEAE Writing is divided into two segments. Each segment of the TEAE is timed. Districts may administer the entire reading portion of the TEAE in one day and all of the writing in another day. However, MDE recommends that the reading portion be administered over two days, completing two to three sections during each day's testing period, and that the writing portion be administered over two days, completing one section during each day's testing period.

K–2 Reading and Writing Checklist

The K–2 Reading and Writing Checklist is a checklist that teachers fill out during the testing window for students in grades K–2. This checklist assesses reading and writing language proficiency.

Listening and Speaking

Minnesota Student Oral Language Observation Matrix

The Minnesota Student Oral Language Observation Matrix (MN SOLOM) is a rubric for evaluating listening and speaking proficiency for students in grades K–12 completed during the testing window. There are two components in the listening domain and four components in the speaking domain.

Test Security

The recovery of testing materials after each administration is critical for two reasons. First, scannable student testing materials must be sent in for scoring in order to provide student reports. Second, 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 seven-digit bar-coded security numbers. School districts complete answer document packing lists to assist Minnesota's testing contractor in determining whether there are missing student answer documents. Minnesota's testing contractor compares bar-code 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*; see the Yearbook for a copy of the current manual.

Title I Assessments

Mathematics

Minnesota Comprehensive Assessments-Series II

The secure test materials for the grade 11 Mathematics Minnesota Comprehensive Assessments-Series II (MCA-II) are the regular student test and answer books and any accommodated materials, including large print test books (18- and 24-point) and answer books, Braille test books and scripts and CDs. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed. All used and unused test books and accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Comprehensive Assessments-Series III

The grades 3-8 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) are delivered either as a computer option or a paper option. For the computer-delivered assessments, there are no secure materials to return For districts chosing the paper-based option, secure test materials include the regular student test and answer books. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed, with the exception of grade 3, which is a combined test/answer book that must be returned with the other secure test books. For students taking accommodated forms, which are paper-based, secure materials include large print test books (18- and 24-point) and answer books, Braille test books and scripts and CDs. All used and unused accommodated materials must be returned to Minnesota's testing contractor. All used and unused test books and accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Comprehensive Assessments-Modified

The grades 5-8 Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) are computer-delivered exams, and as such, there are no secure materials. The grade 11 Mathematics Minnesota Comprehensive Assessments-Modified (MCA-Modified) is paper-based. For this exam, secure test materials include the regular student test and answer books. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed.

Minnesota Test of Academic Skills

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

Reading

Minnesota Comprehensive Assessments-Series II

Secure test materials for the Reading MCA-II include the regular student test and answer books and any accommodated materials, including large print test books (18- and 24-point) and answer books and Braille test books. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed, with the exception of grade 3, which is a combined test/answer book that must be returned with the other secure test books. All used and unused test books and accommodated materials must be returned to Minnesota's testing contractor.

Minnesota Comprehensive Assessments-Modified

The Reading Minnesota Comprehensive Assessments-Modified (MCA-Modified) is paper-based. For this exam, secure test materials include the regular student test and answer books. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed.

Minnesota Test of Academic Skills

Secure test materials for the Reading MTAS include the Task Administration Manuals and Response Option Cards shipped to the district and the student presentation pages available on the online system SchoolSuccess. Following administration, all used and unused Task Administration Manuals must be returned to Minnesota's testing contractor. All Response Option Cards and student presentation pages must be securely destroyed at the district.

Science

Minnesota Comprehensive Assessments-Series II

Since the Science MCA-II is a computer-delivered assessment, the only secure test materials for the Science MCA-II are accommodated materials, including large print test books (18- and 24-point) 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 and Response Option Cards shipped to the district and the student presentation pages available on the online system SchoolSuccess. Following administration, all used and unused Task Administration Manuals must be returned to Minnesota's testing contractor. All Response Option Cards and student presentation pages must be securely destroyed at the district.

Title III Assessments

Reading and Writing

Test of Emerging Academic English

Secure test materials for the Test of Emerging Academic English (TEAE) include the regular student test and answer books and any accommodated materials, including large print test books (18- and 24-point) and answer books, Braille test books and Test Monitor Directions. Districts return all used student answer books to Minnesota's testing contractor. Unused student answer books must be securely destroyed with the exception of the grades 3–4 book, which is a combined test/answer book that must be returned with the other secure test books. All used and unused test books, accommodated materials and Test Monitor Directions must be returned to Minnesota's testing contractor.

Listening and Speaking

Minnesota Student Oral Language Observation Matrix

Secure test materials for the Minnesota Student Oral Language Observation Matrix (MN SOLOM) for grades K–2 include MN SOLOM and Reading and Writing Checklist answer documents and for grades 3–12 include MN SOLOM answer documents. All used K–12 MN SOLOM answer documents must be returned to Minnesota's testing contractor. Unused answer documents must be securely destroyed.

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. See the Yearbook for a copy of the current manual.

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, multi-sensory, 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
- Mathematics MCA-Modified

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:

> 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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

Handheld Calculator for Online Assessment

Description:

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

Rationale:

Allowable Assessments:

- Mathematics MCA-Modified

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. If the student writes responses directly in the test book/document, then the transfer of answers into an answer document must be documented (including the names of school personnel involved) on the *Test Administration Report*.

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

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 scannable answer book.

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
- TEAE

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

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, 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-Modified
- 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.

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

Allowable Assessments:

- Mathematics MCA
- Mathematics MCA-Modified
- 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
- Mathematics MCA-Modified
- 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.

Rationale:

Allowable Assessments:

- Mathematics MCA
- Mathematics MCA-Modified
- Reading MCA
- Science MCA
- TEAE

Templates to Reduce Visual Print, 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
- TEAE

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA

- TEAE

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.

Rationale:

Allowable Assessments:

- Mathematics MCA
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

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:

Brenda Eleanor 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
- Mathematics MCA-Modified
- 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 TEAE, which is a timed test. 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 TEAE is available to EL students who have an IEP. Other EL students must finish the segment(s) on the day they are scheduled.

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:

- TEAE

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.

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
- Reading MCA-Modified
- Science MCA

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

Assistive Technology

Description:

Assistive technology refers to technology that is used to maintain, increase or improve the functional response capabilities of students with 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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

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 scannable answer book.

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

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.

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
- Reading MCA-Modified
- Science MCA

Scratch Paper or Graph Paper (Always Allowed for online assessments and TEAE)

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. 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 TEAE.

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA

- TEAE

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
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

Translation Scribes

Description:

> Translation scribes may be provided to EL students who write a constructed response in a language other than English for the Science MCA. Scribes must be impartial and experienced in translation, and their translation must be transcribed onto the student's scannable answer document immediately below the native writing. This means the student should be told to leave

sufficient room for the scribe to write the translation. The translation must be documented (including the names of school personnel involved) on the Test Administration Report.

Rationale:

According to 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:

- Science MCA

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, then when using speech recognition software. Both adapted conditions were better in quality than handwritten essays.

Allowable Assessments:

- Mathematics MCA
- Mathematics MCA-Modified
- Reading MCA
- Science MCA
- TEAE

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 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.

Rationale:

> According to Hollenbeck, Tindal, Harniss and Almond P (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.
>
> **MDE continues to evaluate the efficacy of this accommodation for future administrations.**

Allowable Assessments:

- Mathematics MCA
- Mathematics MCA-Modified
- Reading MCA
- Reading MCA-Modified
- Science MCA
- TEAE

Other Accommodations Not Listed

If an IEP or 504 Team desires to use an accommodation not on the approved list, they 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 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 over-use 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 assessment scored without accommodations. MDE continually reviews these numbers both in overall percentage and in percent 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

1. verify that accommodations used on state assessments are documented in the IEP;
2. 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

Appropriate Uses for Scores and Reports

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 located on the Minnesota Department of Education (MDE) website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/index.html.

Sample reports for the Minnesota Comprehensive Assessments-Series II (MCA-II), the Minnesota Comprehensive Assessments-Series III (MCA-III), the Minnesota Comprehensive Assessments-Modified (MCA-Modified) Test of Emerging Academic English (TEAE) and Minnesota Student Oral Language Observation Matrix (MN SOLOM) and Minnesota Test of Academic Skills (MTAS) can be found on the MDE website. For the MCA-II and the MCA-III, go to http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/Reports/index.html.
For the TEAE and MN SOLOM, go to http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/ELL_Tests/Reports/index.html.
For the MTAS and the MCA-Modified, go to http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/Alternate/Alternate_Reports_Schools_Parents/index.html.

Individual Student Reports

Minnesota Assessment individual 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-II and the MCA-III, these reports provide scaled scores as well as achievement-level designations associated with the student's performance level. Sub-scores at the strand or sub-strand 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.

Summary Reports for Schools, Districts and the State

Reports summarizing test results at the school, district and state level are used to provide information to school and district educators and administrators. The data is reported for all students tested and all students enrolled on October 1 in the school or district. 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.

Although individual student scores are confidential by law, reports of group (aggregated) scores are considered public information and are reported on the MDE website under Data Downloads, located at http://education.state.mn.us/mde/Data/Data_Downloads/index.html. Student confidentiality on public documents is filtered; if any specific group (for example, English Learners) consists of fewer than ten students, mean scores and the percent of students who are proficient are not included in data files posted to the MDE website.

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-II or 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 or Minnesota English Language Proficiency Standards.

After each test administration, a number of reports are provided for each of the testing programs. The reports include individual student reports and labels, summary reports with data aggregated across all students for each test or disaggregated by demographic variables identified by MDE (for example, gender, ethnicity) and an electronic District Student Results (DSR) file in a tab-delimited format containing individual student records with demographics and multiple scores used to prepare all other reports. 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.

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 sub-strand 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 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 MCA-III.

The TEAE Writing score is the total points earned on two writing prompts; no scale scores are calculated. The scores range from 0 to 44.

The Minnesota Student Oral Language Observation Matrix (MN SOLOM) uses a form of raw scores: teacher ratings. The sum of six ratings on a listening and speaking rubric are used to determine student proficiency. The scores range from 6 to 30.

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. 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 because scale scores adjust for different form difficulties. 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 II (MCA-II) and Minnesota Comprehensive Assessments-Series III (MCA-III) subject and grade range from X01 to X99, where X is the grade tested. 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 MCA-II 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.

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 and the MCA-II, MTAS scale scores from different grades and subjects are not directly comparable.

Like the MTAS scale, the range of observed scale scores for the Minnesota Comprehensive Assessments-Modified (MCA-Modified) also is designed to vary somewhat from year to year. And like the MTAS, all subjects and grade levels of MCA-Modified use the scale score 190 as the cut score for *Partially Meets* and the scale score 200 for the cut score for *Meets*. Although the scales of the MCA-Modified and the MTAS have some properties in common, the scales are distinct, and comparisons between the two scales are not appropriate. Like the MCA-III, the MCA-II, and the MTAS scales, the

MCA-Modified scales are designed to allow comparisons within a grade for a particular subject across years. Scores from different grades and subjects are not directly comparable.

The scale scores for the Test of Emerging Academic English (TEAE) Reading test are expressed on a continuous across grades metric, and range from 1 to 450. The TEAE Reading is a developmental test allowing comparison across grades and years. In scaling the test, only one scale score is assigned to each raw score point.

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, MCA-II, MCA-Modified, MTAS	**Mathematics Reading, Science**	Does Not Meet the Standards	No
		Partially Meets the Standards	No
		Meets the Standards	Yes
		Exceeds the Standards	Yes
TEAE	**Reading**	Level 1	No
		Level 2	No
		Level 3	No
		Level 4	Yes
	Writing	Level 1	No
		Level 2	No
		Level 3	No
		Level 4	No
		Level 5	Yes
MN SOLOM	**Listening & Speaking**	Level 1	No
		Level 2	No
		Level 3	No
		Level 4	Yes
		Level 5	Yes

Description of Reports

The available reports are listed in Table 4.2. Sample reports can be found in MDE's *Interpretive Guide* at http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=017824&RevisionSelectionMethod=latest&Rendition=primary. The summary reports distributed to schools and districts are not for public release; all student data is reported. The public data are available through the School Report Card Interactive website, available on the MDE website at http://education.state.mn.us/ReportCard2005/.

TABLE 4.2. Test Reports

File or Report Name	Report Format	Applies to	
		MCA-III MCA-II, MCA-Modified, MTAS	TEAE, MN SOLOM
Individual Student Report (Home Copy)	Paper	X	X
Individual Student Report (School Copy)	PDF	X	X
Student Labels	Paper	X	X
School Alpha Roster	PDF	X	X
District Student Results File (DSR)	Electronic	X	X

Individual Student Reports

The Individual Student Report (ISR) is a document sent home and maintained by schools 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) and the Minnesota Comprehensive Assessments-Series II (MCA-II) ISRs provide all three averages; the Test of Emerging Academic English (TEAE) and Minnesota Student Oral Language Observation Matrix (MN SOLOM), the Minnesota Test of Academic Skills (MTAS) and the Minnesota Comprehensive Assessments-Modified (MCA-Modified) ISRs 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 TEAE 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 ten. For tests such as the TEAE, MN SOLOM, MCA-Modified or MTAS, the number of students is frequently quite small for school or district populations.

The inside or back of each ISR presents further information about student performance, including presentation of sub-scores. For the MCA-III, the provided sub-score information includes student strand scale scores, along with scale score range and a tolerance band for each strand score representing score precision. For the MCA-II, MCA-Modified, MTAS, TEAE, and MN SOLOM, student raw scores, maximum possible scores and state mean raw scores are reported for each strand (mathematics), sub-strand (reading) or defined strand in test specifications (TEAE and MN SOLOM). MCA-II ISRs also display school and district mean raw scores, while the TEAE and MN SOLOM display the district mean raw score. A proficiency indicator is provided for each subject tested along with the achievement level descriptors for the earned achievement level.

For grades 3-8, a Progress Score is also provided for MCA-II Reading when longitudinal data is available for a student. 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.

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.

School Alpha Roster Report

The school alpha roster contains student-level information, including total raw and scale scores, achievement level and the student’s earned raw points for each strand and/or sub-strand for each subject assessed on a single test. No summary statistics are reported. The reports are designed for teachers and school administrators to reference the information reported to parents and students. The reports can be used to provide hard-copy access to the school’s test results for educational staff who do not have access to the electronic student results file. The report is provided to the district in a print-ready PDF read-only electronic file format.

Overall Proficiency at a Glance

The report includes a series of bar charts showing the percentage of students who are classified as proficient for each subject for the school, district and state. It is the first of three reports providing summary information for the school and district. For the MCA-III and the MCA-II, the data are reported for all students with valid scores enrolled during the test administration as well as students enrolled in the school or district on October 1. For the TEAE, the data are reported in two different formats for all students with valid scores: all students and by number of years enrolled in Minnesota schools. These reports provide a quick graphic representation of proficiency that school and district administrators may share with the public.

Summary Reports

Summary reports are prepared that aggregate scores at the school, district and state level. A separate report is prepared for each grade. This is the second of three reports providing summary data from schools and districts. The top half of the report provides a summary of average raw points on each strand and/or sub-strand. The bottom half provides a summary of scale scores including mean, standard deviation, standard error of the mean and 95 percent confidence interval for all students tested. The school- and district-level reports are provided as PDFs.

Subgroup Reports

The Subgroup Reports contain disaggregated summary data for the school or the district. This is the third of three reports providing summary data. The purpose of this report is to provide a basis for comparing the comparing the achievement of various groups of students. It includes the total number of students, mean scale score and the number and percentage of students at each achievement level. Two major categories of students are reported: all students and students enrolled in the school or the district on October 1. The subgroups reported include the remaining thirteen subgroups reported for No Child Left Behind (NCLB) Adequate Yearly Progress (AYP) results: male/female, ethnic groups (American Indian, Asian Pacific Islander, Hispanic, Black and White Non-Hispanic), special-education-identified students, EL-identified students, students eligible for free- and reduced-priced lunch programs and migrant students. In each case, the data for students whose demographic status is unknown are also reported.

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 sub-strands 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.

Sub-scores provide information about student performance in more narrowly-defined academic content areas. For example, individual scores on strands and/or sub-strands 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 Reading Minnesota Comprehensive Assessments-Series II (MCA-II) in 2011 can be compared to any of the 2006-2010 MCA-II populations. 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-II now could take MCA-Modified instead.. Comparisons to the 2006 administration of MCA-II Reading must also take into account a potential population shift. In the 2006 administration, EL students could take Test of Emerging Academic English (TEAE) Reading instead of the MCA-II. Beginning in 2007, however, EL students were required to take the MCA-II, regardless of whether they also took TEAE Reading or not. An additional change in the testing program took place in 2007 when the Minnesota Test of Academic Skills (MTAS) was first given operationally and as a result students who previously may have been required to take the MCA-II now could take the MTAS instead. For grade 11 Mathematics MCA-II, 2011 saw two important changes to the testing program. First, this year was the first administration of the MCA-Modified. Second, in 2011 the Mathematics Test for English Language Learners (MTELL), which had been offered to EL students since 2007, was no longer made available. Consequently, in making comparisons with past administrations it is important to consider that the population taking this test has changed over time.

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 2011 administration of 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 in 2011 and cannot be directly compared to the grades 3-8 Mathematics MTAS from prior years.

The percentages of students in each achievement level can also be compared across administrations within the same grade and 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 Reading MCA-II in 2011 can be compared to any of the 2006-2010 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 percent 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 sub-strand 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 sub-strand 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 sub-strands 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 only based on 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:"Reliablity" describes measures that provide evidence indicating 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 Reading MCA-II is 49 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 ten 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 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 sub-strand 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 sub-strand 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 sub-strand, 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 sub–strand-level performance should be made cautiously. Significant effort is made to approximate the overall difficulty of the strands or sub-strands from year to year during the test construction process, but fluctuations in difficulty do occur across administrations. Observing trends in strand- and/or sub–strand-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/sub-strand 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, sub-strand 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 15.4 in the *Standards for Educational and Psychological Testing*, "In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of the test result." 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 Comprehensive Assessments-Series II (MCA-II) is reported using the same achievement levels. The Minnesota Comprehensive Assessments-Modified use the same achievement levels and the same content stardards as the MCA-III and MCA-II; however, the cut scores and corresponding achievement-level descriptions for the MCA-Modified are independent of those for the MCA-III or MCA-II. 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 and MCA-II; however, for the MTAS, this performance is related to the Alternate Achievement Standards. The remaining tests for English learner (EL) students use numbered levels. The Test of Emerging Academic English (TEAE) Reading has four numbered levels (1–4), TEAE Writing has five numbered levels (1–5) and the Minnesota Student Oral Language Observation Matrix (MN SOLOM) has five numbered levels (1–5).

Standard setting for grades 3-8 Mathematics MCA-III was conducted in June, 2011. Standard setting for grade 11 Mathematics and Reading MCA-II was conducted from March to July 2006. Standard setting for grades 3-8 Mathematics MTAS aligned to 2007 academic standards was conducted in June, 2011. Standard setting for grade 11 Mathematics and Reading MTAS was conducted in May 2007, and a standards validation workshop was held in May of 2008 due to the administration and rubric revisions of the test. For MCA-II Science and MTAS Science, standard setting was held in July 2008. Standard setting for Mathematics and Reading MCA-Modified was conducted in June, 2011. 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. The 2006 MCA-II standard setting technical report can be found on the Minnesota Department of Education (MDE) website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MCA_II/MCA_II_Technical_Reports/index.html.
The 2008 MTAS standards validation report can be found on the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTA

S/MTAS_Technical_Reports/index.html. The 2008 MCA-II Science and MTAS Science standard setting technical report can be found at the MDE website at http://education.state.mn.us/mdeprod/groups/Assessment/documents/Publication/034904.pdf.

The 2011 MCA-III Mathematics, MTAS Mathematics, and MCA-Modified Mathematics and Reading standard setting technical report can be found at the MDE website at

Standard setting for the TEAE occurred in summer 2002 and in winter 2003 using a procedure known as item mapping. Complete details of the TEAE standard setting have already been published in a document titled "Test of Emerging Academic English (TEAE) Standards Setting Report: Summer 2002–Winter 2003." This report is available on the MDE website at: http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/ELL_Tests/ELL_Technical_Reports/index.html.

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 (Zeiky, 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 Grades 3-8 Mathematics Minnesota Comprehensive Assessments-Series III

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 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 at the MDE website.

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*. The report can be found at the MDE website.

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 the 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 report can be found at the MDE website.

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. Round 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:

- 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.

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*. The report can be found at the MDE website.

Table 5.1 shows the participant-recommended cut scores, as taken from participants' Round 3 bookmark placements. Cut scores are shown on the theta metric. Table 5.2 shows the impact data associated with the cut scores shown in table 5.1.

TABLE 5.1. Participant-Recommended Cut Scores (Round 3) for Mathematics MCA-III

Content Area	Grade	Cut Scores (Theta Metric)		
		Partially Meets	Meets	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 5.2. 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 the 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 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 5.3 shows the cut scores from the vertical articulation. Table 5.4 shows the impact data associated with the cut scores shown in table 5.3.

TABLE 5.3. Vertical Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Cut Scores (Theta Metric)		
		Partially Meets	Meets	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 5.4. 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 5.5. Impact data associated with the final cut scores are reported in Table 5.6.

TABLE 5.5. Commissioner-Approved Cut Scores

Content Area	Grade	Cut Scores (Theta Metric)		
		Partially Meets	Meets	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 5.6. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2006 Percentage of Students in Achievement Level			
		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

Standard Setting for Mathematics and Reading Minnesota Comprehensive Assessments-Series II

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was implemented for the Minnesota Comprehensive Assessments-Series II (MCA-II) standard setting held in St. Paul, Minnesota, on June 27–30, 2006. Achievement-level cut scores were established for reading and mathematics in grades 3, 5, 8 and high school. For grades 4, 6 and 7, achievement-level cut scores were statistically interpolated from the cut scores in grades 3, 5 and 8. The BSSP consisted of training, orientation and three rounds of judgments. The MCA-II standard setting lasted four days, with the first day devoted to table leader training and three days for standard setting and description writing. The outcomes of the conference are reported in this summary.

The CTB/McGraw-Hill (CTB) Standard Setting Team worked with staff from MDE to design and implement the MCA-II standard setting. The CTB Standard Setting Team is comprised of CTB psychometricians and standard setting specialists.

Note that starting in 2011, students in grades 3-8 Mathematics no longer take the MCA-II; instead they take the Minnesota Comprehensive Assessments-Series III (MCA-III). For a description of the Mathematics MCA-III standard setting, see the previous section.

Table Leaders

Group leaders administer the standard setting for the portions of the process during which participants are divided into groups, often called “tables.” In each grade per content area, the group leader is in charge of security, data management and time management. Group leaders collect the bookmark data from participants and communicate with the CTB Research staff and Minnesota Department of Education (MDE) staff. Group leaders also keep the tables on approximately the same schedule and lead cross-table discussions in Round 3. The group leaders were all from the CTB Content Development department.

Participants

MDE invited approximately 10–12 participants from across Minnesota to set cut scores in each grade per content area. The details of the credentials and demographics of the participants can be found in the *MCA-II Bookmark Standard Setting Technical Report* on the MDE website at: http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=032785&RevisionSelectionMethod=latest&Rendition=primary.

Articulation Panelists

Articulation panelists are stakeholders in the results of the assessment system from a broad range of perspectives. Members of the 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.

Articulation commenced with a presentation of the process used by the previous committee, and the results. Also included were the interpolated cut scores for grades 4, 6 and 7. The panel discussed the policy and educational implications of the cut scores, and how those implications might change if the cut score values changed. After this discussion, there were two rounds of smoothing. In Round 1, panelists individually smoothed the scores. In Round 2, panelists were asked to come to a consensus about where the final cut score recommendations should be.

Finally, the full report of the standard setting was provided to the National TAC, an independent committee of nationally recognized technical advisors in educational measurement. They reviewed the procedures and outcomes, providing advice to the Commissioner on the adoption of final cut scores.

Bookmark Materials

Ordered Item Booklets

The ordered item booklets (OIB) contained the operational items from the 2006 MCA-IIs arrayed by item difficulty. The details of the OIB construction can be found in the *MCA-II Bookmark Standard Setting Technical Report* on the MDE website at:

http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=032785&RevisionSelectionMethod=latest&Rendition=primary.

Item Maps

The item maps summarize the material in the OIB. For each item, the item maps give the difficulty, location in the test, test segment, item number, item type (multiple-choice, constructed-response or gridded-response), score key (correct response for a multiple-choice item and score points for a constructed-response item) and content strand or substrand measured by the item. Participants responded to two questions for each item as they studied the OIB. The first question asks, “What does this item measure in regard to the Minnesota Academic Standards? That is, what do you know about a student who can respond successfully to this item and score point?” The second question asks, “Why is this item more difficult than the preceding items?”

Training for Table Leaders and Articulation Panelists

Table leaders and articulation panelists were trained on Day 1 of the MCA-II Standard Setting. Articulation panelists participated in training activities for the BSSP but did not participate in the discussions or bookmark ratings. Training lasted about four hours for articulation panelists and about five-and-one-half hours for table leaders. Training included an overview of the reasons for standard setting and training on the BSSP. The group also participated in a mock BSSP using a sample OIB. Training materials are available in Section D of the MCA-II Standard Setting Technical Report.

Target Students

A target student is a student whose performance is equivalent to the minimum score required for entry into a particular achievement level. After training in the BSSP, the table leaders discussed the *Partially Meets the Standards*, *Meets the Standards* and *Exceeds the Standards* target students. The table leaders were directed to use the Minnesota Academic Standards and the Achievement Level Descriptors developed during initial standard-setting activities in March 2006. These definitions served as a basis for establishing a common understanding of the types of students who should be considered as *Partially Meets the Standards*, *Meets the Standards*, and *Exceeds the Standards* on the Reading and Mathematics MCA-II for grades 3, 5, 8 and high school.

The Standard Setting Meeting

Before beginning the standard-setting activities, MDE and CTB 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 to MDE. 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 MDE. The participants were given an overview of standard setting and were introduced to the BSSP. The standard setting meeting was conducted in a series of three rounds of setting bookmarks, as described below. The details of the workshop can be found in the MCA-II *Bookmark Standard Setting Technical Report* on the MDE website at: http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=032785&RevisionSelectionMethod=latest&Rendition=primary.

Round 1

After the introduction, participants spent approximately one hour taking the operational MCA-II for their grade and content area. Table leaders reviewed target student descriptors and the Achievement Level Descriptors with the participants. Participants studied each item in the OIB in terms of what each item measures and why each item is more difficult than the items preceding it.

Once all tables had completed their study of the OIB, the participants were trained on how to place their bookmarks. Participants were given training materials and three explanations of bookmark placement. The training materials "Bookmark Placement" and "Frequently Asked Questions about Bookmark Placement" were read aloud. The first explanation demonstrated the mechanics of bookmark placement; participants were instructed that all items preceding the bookmark define the knowledge, skills and abilities that, for example, a *Meets the Standards* student is expected to know. The second explanation was conceptual; participants were instructed to examine each item in terms of its content. They were instructed to make a judgment about content that a student would need to know in order to be considered just *Meets the Standards*. The final explanation addressed the relationship between the placement of the bookmark and the scale score.

Participants were tested on their knowledge of bookmark placement with a brief check set. They were given the correct answers for the check set as well as explanations of those answers. The answers to the check set questions were reviewed with the participants, and they were given another opportunity to ask questions about bookmark placement. Once participants had demonstrated that they understood bookmark placement, they were directed to individually place 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. Participants were instructed to discuss those items for which there was disagreement at their table; thus, they discussed the range of items between the lowest and highest bookmark placements for each achievement level. After the discussion, participants again placed their bookmarks. Participants were reminded that bookmark placement is an individual activity.

Round 3

At the beginning of Round 3, a member of the CTB Standard Setting Team, working with an MDE representative, presented participants with impact data based on their Round 2 bookmark placements. Impact data, based on the most recent test administration, are the percentages of students who would be classified in each achievement level if the groups' recommendations were adopted at that point. Impact data were shown for the entire population of test-takers, as well as for seven disaggregations (subgroups): male and female students and five ethnic groups of students (American Indian, Asian, Hispanic, Black Non-Hispanic and White). The CTB team member answered process-related questions, and MDE representatives answered all policy-related questions concerning the impact data. It was emphasized to the participants that the impact data were being presented as a "reality check." The group leaders facilitated discussion among the participants on their bookmark placements. After the discussion, participants again placed their bookmarks independently.

After the Round 3 bookmark placement, the two table leaders from each grade per content area convened with an independent group of articulation panelists to review the cut score recommendations. Together, the table leaders and articulation panelists reviewed the participant-recommended cut scores and resulting impact data, and they suggested changes to promote better cross-grade articulation.

Table 5.7 shows the participant-recommended cut scores, as taken from participants' Round 3 bookmark placements for grades 3, 5, 8 and high school. Cut scores shown for grades 4, 6 and 7 were statistically interpolated from the cut scores in grades 3, 5 and 8. Cut scores are shown on the theta metric, subsequently transformed by MDE onto an operational test scale. Table 5.8 shows the impact data associated with the cut scores shown in table 5.7.

TABLE 5.7. MCA-II Participant-Recommended Cut Scores (Round 3) for Grades 3, 5, 8 and High School, and Statistically Interpolated Cut Scores for Grades 4, 6 and 7

Content Area	Grade	Cut Scores (Theta Metric)		
		Partially Meets	Meets	Exceeds
Mathematics	3	−1.46	−0.76	0.59
	4	−1.04	−0.40	0.75
	5	−0.80	−0.16	0.84
	6	−0.78	−0.14	0.97
	7	−0.80	−0.09	1.00
	8	−0.76	−0.02	1.00
	11	0.00	0.54	1.32
Reading	3	−1.22	−0.82	−0.08
	4	−1.19	−0.75	0.32
	5	−1.24	−0.77	0.46
	6	−1.18	−0.60	0.48
	7	−1.10	−0.33	0.49
	8	−1.10	−0.17	0.54
	10	−1.08	−0.48	0.40

TABLE 5.8. Impact Data Associated with Participant-Recommended and Interpolated Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	8.9	13.3	46.5	31.3
	4	14.8	16.2	42.3	26.7

	5	20.0	21.1	36.8	22.1
	6	20.4	20.4	40.5	18.7
	7	20.2	22.3	40.0	17.5
	8	20.3	23.0	39.9	16.8
	11	49.2	21.0	20.9	8.9
Reading	3	12.2	8.2	24.5	55.1
	4	12.0	8.9	36.9	42.3
	5	11.6	9.0	44.0	35.4
	6	12.3	13.2	39.1	35.4
	7	13.4	20.0	30.8	35.8
	8	14.5	24.2	28.9	32.5
	10	15.5	14.4	34.3	35.8

Table 5.9 shows the articulation panel's smoothed cut scores. These cut scores were smoothed by the articulation panelists and the table leaders after the Round 3 of standard setting.

TABLE 5.9. Articulation Panel's Smoothed Cut Scores

Content Area	**Grade**	**Cut Scores (Theta Metric)**		
		Partially Meets	**Meets**	**Exceeds**
Mathematics	3	−1.46	−0.76	0.59
	4	−1.04	−0.40	0.75
	5	−0.80	−0.16	0.84
	6	−0.78	−0.14	0.97
	7	−0.80	−0.09	1.00
	8	−0.76	−0.02	1.00
	11	0.00	0.54	1.32
Reading	3	−1.22	−0.82	0.02
	4	−1.19	−0.65	0.32
	5	−1.24	−0.64	0.46
	6	−1.18	−0.51	0.48
	7	−1.10	−0.33	0.49
	8	−1.10	−0.28	0.54
	10	−1.08	−0.32	0.60

Table 5.10 shows the impact data associated with these smoothed cut scores. Note that the articulation panel made smoothing recommendations for Reading but made the recommendation not to smooth Mathematics.

TABLE 5.10. Impact Data Associated with Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	8.9	13.3	46.5	31.3
	4	14.8	16.2	42.3	26.7
	5	20.0	21.1	36.8	22.1
	6	20.4	20.4	40.5	18.7
	7	20.2	22.3	40.0	17.5
	8	20.3	23.0	39.9	16.8
	11	49.2	21.0	20.9	8.9
Reading	3	12.2	8.3	30.4	49.2
	4	12.0	11.4	34.5	42.3
	5	11.6	14.2	38.8	35.4
	6	12.3	16.2	36.2	35.4
	7	13.4	20.0	30.8	35.8
	8	14.5	21.0	32.1	32.5
	10	15.5	19.3	37.4	27.9

Quality Control Procedures

The CTB Standard Setting Team adhered to many quality control procedures to ensure the accuracy of the materials used and the consistency of the presentation of results during the standard setting. Before the workshop, the Standard Setting Team verified the ordering of items in the OIB, the accuracy of the information in the Item Maps, the accuracy of the Microsoft Excel macros and Bookmark Pro software used to generate results and impact data and the completeness of the anchor papers and scoring guides. During the workshop, all data entry and reporting was monitored by a two-person team who first scanned the data and checked them for accuracy. Any results that appeared to be questionable were further investigated by the Standard Setting Team and CTB Research staff.

Effectiveness of Training

An indication of the effectiveness of training may be found in the participants' understanding of the training as reported in their evaluations of the MCA-II Standard Setting. The majority of participants reported that they understood how to place a bookmark, the bookmark training made the task of bookmark placement clear, the training materials were helpful, the BSSP was well described and the goals of the process were clear to them. Complete results of participant evaluations. The details of participant evaluations can be found in Section F of the MCA-II Standard Setting Technical Report.

Perceived Validity

Another indication of the success of the standard setting may be found in the participants' perception of the validity of the BSSP itself. The majority of participants reported that the BSSP produced valid cut scores, and they were satisfied with their group's final bookmarks. The details of participant perceived validity evaluations can be found in Section F of the MCA-II Standard Setting Technical Report.

Commissioner-Approved Results

After the standard setting meeting, the MDE commissioner reviewed the recommended cut scores for overall consistency and continuity. Slight adjustments were made to a few of the cut scores. The final cut scores approved by the commissioner for the 2006 MCA-II administration are given in table 5.11. Impact data associated with the final cut scores are reported in table 5.12.

TABLE 5.11. Commissioner-Approved MCA-II Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	3	21	−1.69	31	−0.72	42	0.59
	4	24	−1.23	33	−0.40	43	0.75
	5	25	−0.87	34	−0.13	44	0.86
	6	27	−0.86	37	−0.14	50	0.97
	7	26	−0.80	36	−0.09	51	1.00
	8	25	−0.73	35	−0.01	48	0.98
	11	27	0.00	37	0.55	50	1.30
Reading	3	22	−1.40	29	−0.84	37	−0.01
	4	22	−1.36	30	−0.65	38	0.32
	5	23	−1.45	32	−0.67	42	0.50
	6	23	−1.37	32	−0.51	41	0.48
	7	31	−1.10	40	−0.33	47	0.49
	8	29	−1.06	39	−0.26	47	0.58
	10	27	−1.04	37	−0.29	47	0.59

TABLE 5.12. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2006 Percentage of Students in Achievement Level			
		Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	7.6	14.5	46.5	31.3
	4	14.8	16.2	42.3	26.7
	5	20.0	21.1	36.8	22.2
	6	20.4	20.4	40.5	18.7
	7	20.2	22.3	40.0	17.5
	8	20.3	23.0	37.3	19.4
	11	49.2	21.0	19.7	10.1
Reading	3	9.0	9.4	26.6	55.1
	4	9.5	13.8	34.4	42.3
	5	8.0	15.0	41.5	35.3
	6	9.6	18.9	36.2	35.4
	7	13.4	20.0	30.8	35.8
	8	14.5	20.9	32.1	32.5
	10	15.5	19.3	33.4	31.8

Method to Assign Observed Scores to Levels

The cut scores approved by the Commissioner of Education were applied to the raw (observed total correct) scores for the 2006 MCA-II administration. For the MCA-II administrations in 2007 and beyond, scores will be equated, through the theta metric, to the 2006 administration. Thus, cut scores on the theta scale remain the same, but raw score cuts may change from year to year. To find the cut scores on the raw score scale for a given year, the raw score to theta score transformation is found as described in chapter 6, "Scaling." The raw score whose corresponding theta score is closest to the theta cut score becomes the raw score cut for that administration.

Standard Setting for Science Minnesota Comprehensive Assessments-Series II

Standard setting for MCA-II Science was held during 15-16 July 2008, with the consequential validity meeting occurring on 18 July 2008. The activities of the meeting are documented in a paper titled *Minnesota Comprehensive Assessment – II (MCA-II) Minnesota Test of Academic Skills (MTAS) Report on Science Standard Setting*. The report can be found at the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/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's testing contractor facilitated the standard setting under the supervision of MDE.

In order to match what was done in the Mathematics and Reading standard setting, an item mapping procedure was followed. The primary variation in procedure from the previous standard setting meeting was that participants were not divided into tables within grade level. Instead, participants within grade levels worked as a single group. Details of the meeting can be found in the *Minnesota Comprehensive Assessment – II (MCA-II) Minnesota Test of Academic Skills (MTAS) Report on Science Standard Setting* on the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

Participants

MDE invited 10–20 participants from across Minnesota to set cut scores in each grade. The details of participant credentials and demographics can be found in *Minnesota Comprehensive Assessment – II (MCA-II) Minnesota Test of Academic Skills (MTAS) Report on Science Standard Setting*. The report can be found at the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

Standard Setting Meeting

Round 1

When back in grade-level groups the actual ordered-item books were distributed and the facilitators walked through the item-mapping process again. The facilitators asked questions to verify that panelists understand the procedure. The following features of the ordered-item books were carefully covered:

- Constructed-response items appear once for each non-zero score level. Each page for CRs represents the probability of scoring at that level or higher for the CR item.
- Pages DO NOT correspond to raw scores.

After panelists indicated that they understood their task, they were asked to document their readiness on their judgment form. The facilitator emphasized that each person was to make independent judgments. Once all panelists had done so, the group was allowed to complete the task.

Round 2

Panelists were first asked to discuss their understanding of the process, and ask any questions needed to clarify the task. Following a short discussion, feedback in the form of a graphical presentation of the page numbers of each panelist was provided. The graphs were presented such that persons were anonymous (judge "names" were numerals). The facilitator explained how to read the graph and called out features of the distributions, such as minimum and maximum. The median of the cut score judgments was included as an additional mark on the graph. The facilitator then led discussion about the similarities/differences of the results. Following this discussion the panelists were asked to discuss the knowledge, skills, and abilities (KSA) required to respond correctly (or at that score level or higher for CR) to each item beginning from the earliest page number judged as a cut score and ending with the last page number thusly marked. Panelists were asked to relate these KSA to the achievement level descriptors.

At the conclusion of this discussion, panelists were given an opportunity to refine their judgments. Before doing such, panelists filled out the readiness portion of their judgment sheet. Panelists were asked to continue to do independent work.

Round 3

Feedback and timing of feedback in Round 3 was:

0 Minutes: judge minimum, maximum, median and range of pages

5 Minutes: raw score & percent correct minimum, maximum, median cutscores

25 Minutes: impact data

After 5 minutes the raw score cuts produced by the panelists' page number judgments were presented and explained. The relationship between the raw score and page number (or lack of relationship) was discussed. After panelists indicated understanding of this data, the ordered-item map was passed out. The facilitator explained how to read the ordered-item map by attending to particular features: places where several pages had the same scale score values, places where there were large gaps in scale scores between pages, and distance from one score level to the next for the same constructed-response item. The implications of these features were discussed, including how placing markers on pages close to one another in the ordered-item book may result in very similar judgments from a scaling – and presumably construct – perspective.

At 25 minute mark the impact data was presented on the projector screen. The facilitator led discussion on whether the impact data met expectations of the panelists.

Panelists indicated their readiness and adjusted cut scores using the ordered-item books again.

Round 4

All grade-level groups were brought together to discuss the results. A final opportunity to refine cutscores was provided, and panelists were asked to respond to surveys about their experience.

Training

Before each round, panelists were trained on the process to be used in the upcoming round and given the opportunity to ask questions. Before any round of judgments was entered, panelists were asked to indicate in writing that they were ready to begin. If they were not ready, Minnesota's testing contractor staff re-taught the process or answered questions until everyone was ready to move forward.

Panelists reported that the training was effective and that they understood the procedures. At the end of the standard setting, panelists provided evaluations of the activity. The majority of panelists indicated they were satisfied with the process and the final outcomes.

Consequential Validity Panel

A group of educators and community members convened to form a combined consequential validity panel for MCA-II Science and MTAS Science. Panelists included one member from each grade group attending the previous meetings, plus additional panelists who were selected to represent the broader community. The panel commenced with a presentation of the process used by the previous committee, and the results. The cut scores and associated impact data prior to smoothing is provided in tables 5.13 and 5.14.

The facilitator led discussion about the consequences of the educator panel recommendations, and continued to explain why it was reasonable to consider alternative outcomes based on policy implications. The standard error bands were explained: their meaning and potential usage. Panelists were provided an opportunity to see the impact of changing some cut scores both in terms of the percent correct and the percentage of students classified in each level after changes were made.

Finally, panelists were provided a judgment form and given an opportunity to offer recommendations for the cutscores. The judgment form reported the educator panel recommendations and standard error bands. Panelists were instructed that they could make recommendations within the standard error bands. The smoothed cut scores and associated impact data are provided in tables 5.15 and 5.16.

Finally, the full report of the standard setting was provided to the National TAC, an independent committee of nationally recognized technical advisors in educational measurement. They reviewed the procedures and outcomes, providing advice to the commissioner on the adoption of final cut scores. The commissioner-approved cut scores and associated impact data are provided in tables 5.17 and 5.18.

TABLE 5.13. Panelist-Recommended Cut Scores (Round 4)

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	24	–0.3941	29	0.4173	35	1.8230
8	24	–0.6279	32	0.4302	39	1.4966
High School	20	–1.3231	35	0.0560	50	1.4242

TABLE 5.14. Impact Data Associated with Panelist-Recommended Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	32	29	32	7
8	24	38	29	9
High School	11	36	45	8

TABLE 5.15. Consequential Validity Panel's Smoothed Cuts Scores

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	24	–0.3941	29	0.4173	35	1.8230
8	24	–0.6279	32	0.4302	39	1.4966
High School	26	–0.7087	38	0.3025	52	1.6654

TABLE 5.16. Impact Data Associated with Consequential Validity Panel's Smoothed Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	32	29	32	7
8	24	38	29	9
High School	22	35	38	5

TABLE 5.17. Commissioner-Approved Cut Scores

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	24	–0.3941	29	0.4173	35	1.8230
8	24	–0.6279	32	0.4302	39	1.4966
High School	26	–0.7087	38	0.3025	52	1.6654

TABLE 5.18. Impact Data Associated with Commissioner-Approved Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	32	29	32	7
8	24	38	29	9
High School	22	35	38	5

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, Sweeeny, and 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.

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*. The report can be found at the MDE website.

Table Leaders

During the standard setting, participants were divided into groups, called "tables." Each table had one table leader that that had been previously selected by the 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 Modfied Angoff and the third round using Reasoned Judgment. Round 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:

- Read each task in the book thinking about the knowledge, skills and abilities required to answer the item correctly.
- Record comments or notes about competencies required to address a given task in the book.
- 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, 2008-2010 MTAS 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*. The report can be found at the MDE website.

Table 5.19 shows the participant-recommended cut scores, as taken from participants' Round 3 judgment. Cut scores are shown on the raw score 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 MTAS

Content Area	Grade	Cut Scores		
		Partially Meets	Meets	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 5.20. 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 5.21

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

Content Area	Grade	Cut Scores		
		Partially Meets	Meets	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 5.22. 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 5.23. Impact data associated with the final cut scores are reported in Table 5.24.

TABLE 5.23. Commissioner-Approved Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	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 5.24. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2011 Percentage of Students in Achievement Level			
		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

Standard Setting for Mathematics and Reading Minnesota Test of Academic Skills (MTAS)

To enhance the coherence of the Minnesota assessment system, procedures similar to those used for the Minnesota Comprehensive Assessments-Series II (MCA-II) standard setting were designed for the Minnesota Test of Academic Skills (MTAS). Because the MTAS is composed of a small number of observations of student achievement, the test design does not lend itself to the exclusive use of the Bookmark Standard Setting Procedure (BSSP) method already described in this chapter. A combination of score estimation and item mapping strategies was used, with a similar approach to committee work as was used for the MCA-II.

Minnesota's testing contractor collaborated with the Minnesota Department of Education (MDE) and MDE's National Technical Advisory Committee (TAC) in designing the standard setting activities. Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

In May of 2008, MDE and its testing contractor conducted a standards validation workshop to review the achievement levels set in 2007 because of the revisions to the administration and rubric of the MTAS. The details of this workshop can be found in the May 2008 MTAS *Standards Validation Technical Report* on the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

Note that starting in 2011, students taking the MTAS in grades 3-8 mathematics are assessed under the *2007 Minnesota K–12 Academic Standards in Mathematics*. For a description of the current grades 3-8 Mathematics MTAS assessment standard setting, see the previous section.

Participants

MDE invited 10–20 participants from across Minnesota to set cut scores in each grade per content area. The invitation approach differed from that of the MCA-II in that approximately half of the invited participants were educators involved in special education either through academic specialty or classroom experience. The details of participant credentials and demographics can be found in Appendix A of the May 2008 MTAS *Standards Validation Technical Report* on the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html

Process Theory

The 2007 standard setting activity began with Modified Angoff, a form of score estimation, and finished with an item mapping approach. The details are provided in the descriptions of rounds below. To be consistent with the MCA-II, cut scores were set at four of the seven grade levels: 3, 5, 8 and high school. The remaining grades were set by the articulation panel.

The 2008 standards validation workshop did not use the Modified Angoff procedure described below in Round 1, mainly because the workshop was designed to validate the 2007 cut scores. These 2007 cut scores were pre-seeded in the OIBs that participants used in providing their judgment.

Round 1—Modified Angoff

Panelists made judgments about the percentage of students from each achievement level who are expected to score in each rubric level for each operational task. These percentages were to round to 0 or 5 in the ones place, with 0 and 100 as possible values. For each task, the ratings for a score level (within an achievement level) were required to sum to 100. Therefore, ratings such as 0, 10, 40 and 50 (score levels 0, 1, 2 and 3 respectively) were acceptable, while ratings such as 0, 13, 37 and 50 were not.

To find a panelist's cut score, each rating was multiplied by the associated score level, and the products were summed over all tasks and score levels. The final sum was divided by 100, returning a value in the interval [0,18] with decimals possible (for example, 12.4). Finally, this value was rounded to the nearest integer. This value represented a cut score on the raw score scale.

Round 2—Item Mapping

An item mapping approach was used to provide a tool for panelists to fine-tune their cut score decisions. The item map and ordered-task book (OTB) were created using a pool of up to 12 tasks. The strategy was to pre-insert marks in the OTBs associated with the raw score cuts produced by panelists in Round 1.

The pool of tasks was analyzed using the Rasch Partial Credit Model. The WINSTEPS calibration software provides 50 percent response probability (RP) measure estimates for non-zero rubric steps (that is, three step values per task). These values were used to create OTBs, with each page representing a task and score combination and the pages ordered from the least to the greatest measure. There were between 27 (9 tasks) and 36 (12 tasks) pages in each book.

Prior to providing the OTBs to panelists, their individual judgments from Round 1 for the *Partially Meets*, *Meets*, and *Exceeds* cut scores were pre-inserted into the books. The 50 percent RP theory was explained to panelists, who were asked to use the theory to determine if their cut scores should be modified.

Round 3

Panelists were given an opportunity to adjust their cut scores in the OTBs after reviewing impact data.

Round 4

Panelists were given an opportunity to adjust their cut scores in the OTBs after reviewing the cut scores recommended by the other committee in the same subject area.

Round 5

Panelists were given an opportunity to adjust their cut scores in the OTBs after reviewing the cut scores recommended by the other subject committees.

Training

Before each round, panelists were trained on the process to be used in the upcoming round and given the opportunity to ask questions. Before Rounds 1 and 2, Minnesota's testing contractor staff provided training on the specific cut score setting procedure. For the Modified Angoff procedure, a practice test was provided, and panelists were trained on how the raw score cut values were obtained. Before any round of judgments was entered, panelists were asked to indicate in writing that they were ready to begin. If they were not ready, Minnesota's testing contractor staff re-taught the process or answered questions until everyone was ready to move forward.

Panelists reported that the training was effective and that they understood the procedures. At the end of the standard setting, panelists provided evaluations of the activity. The majority of panelists indicated they were satisfied with the process and the final outcomes.

Articulation

The articulation panel was held the following day. Panelists included one member from each subject per grade group attending the previous meeting, plus three additional panelists who were selected to represent the broader community. Articulation commenced with a presentation of the process used by the previous committee, and the results. Also included were the interpolated cut scores for grades 4, 6 and 7. The cut scores and associated impact data prior to smoothing is provided in tables 5.25 and 5.26.

The panel discussed the policy and educational implications of the cut scores, and how those implications might change if the cut score values changed. After this discussion, there were two rounds of smoothing. In Round 1, panelists individually smoothed the scores. In Round 2, panelists were asked to come to a consensus about where the final cut score recommendations should be. The smoothed cut scores and associated impact data are provided in tables 5.27 and 5.28.

Finally, the full report of the standard setting was provided to the National TAC, an independent committee of nationally recognized technical advisors in educational measurement. They reviewed the procedures and outcomes, providing advice to the commissioner on the adoption of final cut scores. The commissioner-approved cut scores and associated impact data are provided in tables 5.29 and 5.30.

TABLE 5.25. Panelist-Recommended Cut Scores (Round 5) for Grades 3, 5, 8 and High School, and Statistically Interpolated Cut Scores for Grades 4, 6, and 7

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	3	9	0.4296	12	1.2427	16	2.2246
	4	9	0.2384	12	0.8618	16	1.8941
	5	8	0.4323	11	1.1093	15	1.9540
	6	8	0.2045	11	0.9785	16	2.2397
	7	9	0.4967	11	0.9024	16	2.0234
	8	10	0.6651	12	1.0486	18	3.2536
	11	8	0.2433	12	1.3005	14	1.7603
Reading	3	10	0.6242	13	1.2037	15	1.6476
	4	10	0.7319	13	1.3351	15	1.8271
	5	10	0.7745	12	1.1486	15	1.7761
	6	10	0.7929	12	1.1585	15	1.7614
	7	9	0.5849	10	0.8016	14	1.5892
	8	10	0.6289	11	0.8203	14	1.3959
	10	9	0.5984	11	1.0411	14	1.6627

TABLE 5.26. Impact Data Associated with Panelist-Recommended and Interpolated Cut Scores

Content Area	Grade	2007 Percentage of Students in Achievement Level			
		Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	19	14	26	41
	4	16	12	34	38
	5	12	21	37	30
	6	12	26	37	25
	7	22	19	40	19
	8	24	16	50	10
	11	16	32	21	30
Reading	3	18	22	14	46
	4	19	17	16	48
	5	21	6	20	52
	6	17	8	17	58
	7	15	4	15	67
	8	13	5	11	71
	10	14	6	9	71

TABLE 5.27. Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	3	9	0.4296	12	1.2427	16	2.2246
	4	9	0.2384	12	0.8618	16	1.8941
	5	8	0.4323	11	1.1093	15	1.9540
	6	8	0.2045	11	0.9785	16	2.2397
	7	9	0.4967	11	0.9024	16	2.0234
	8	10	0.6651	12	1.0486	18	3.2536
	11	8	0.2433	12	1.3005	14	1.7603
Reading	3	10	0.6242	13	1.2037	15	1.6476
	4	10	0.7319	13	1.3351	15	1.8271
	5	10	0.7745	12	1.1486	15	1.7761
	6	10	0.7929	12	1.1585	15	1.7614
	7	9	0.5849	11	0.9992	15	1.8262
	8	9	0.4205	11	0.8203	15	1.6284
	10	9	0.5984	11	1.0411	15	1.9102

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

Content Area	Grade	2007 Percentage of Students in Achievement Level			
		Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	3	19	14	26	41
	4	16	12	34	38
	5	12	21	37	30
	6	12	26	37	25
	7	22	19	40	19
	8	24	16	50	10
	11	16	32	21	30
Reading	3	18	22	14	46
	4	19	17	16	48
	5	21	6	20	52
	6	17	8	17	58
	7	15	10	16	60
	8	11	7	18	64
	10	14	6	15	65

TABLE 5.29. Commissioner-Approved Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	3	9	0.4296	13	1.1759	17	2.3889
	4	9	0.2384	13	1.0704	17	2.4136
	5	8	0.4323	12	1.3033	16	2.2619
	6	8	0.2045	12	1.2009	17	2.7451
	7	9	0.4967	12	1.0897	17	2.5031
	8	10	0.6651	13	1.2394	18	3.2536
	11	8	0.2433	13	1.5254	15	2.0238
Reading	3	10	0.6242	14	1.4108	16	1.9474
	4	10	0.7319	14	1.5612	16	2.1719
	5	10	0.7745	13	1.3378	16	2.0745
	6	10	0.7929	13	1.3412	16	2.0475
	7	9	0.5849	12	1.1895	16	2.1282
	8	9	0.4205	12	1.0056	16	1.9256
	10	9	0.5984	12	1.2427	16	2.2246

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

Content Area	Grade	2007 Percentage of Students in Achievement Level			
		Does Not Meet	Partially Meets	Meets	Exceeds
Mathematics	3	19	23	30	28
	4	16	24	41	19
	5	12	26	37	25
	6	12	31	43	14
	7	22	23	47	8
	8	24	30	37	10
	11	16	48	15	20
Reading	3	18	25	14	43
	4	19	21	15	44
	5	21	13	17	49
	6	17	15	13	55
	7	15	11	16	57
	8	11	9	19	61
	10	14	7	16	63

The 2007 administration was the inaugural year for the MTAS and performance standards were established and published in May of that year. Following federal peer review of the MTAS program several changes were recommended and incorporated into the 2008 version of the tests. These changes were not intended to change the construct of measurement, but instead were designed to broaden content coverage, increase standardization of the administration and clarify scoring issues encountered in 2007. The changes included addition of three tasks to each test, more detailed scripting of the administration of each task and embedding the rubric in the scripts so that test administrators better understood how to award scores.

2008 Standards Validation Study

In May of 2008, MDE and its testing contractor conducted a standards validation workshop to review the achievement levels set in 2007 because of the revisions to the administration and rubric of the MTAS. The outcomes of the validation study are presented below, and full details of this workshop can be found in the May 2008 MTAS *Standards Validation Technical Report* on the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

The standards validation was composed of two partly overlapping events: a cut score review, and an articulation. The final cut scores were determined after Minnesota Department of Education (MDE) collected and reviewed information from the workshop.

The cut score review component of the standards validation meeting was held during May 29-30, 2008, and was attended by 34 educators. The articulation meeting was held on May 30, 2008, and was attended by nine stakeholders along with educators who participated in the standards validation meeting.

The MTAS standards validation activities mirrored the standard setting activities held for the MTAS in 2007. Two components of the standards validation were used in both the standard setting and the validation study: cut score location at select—but not all—grades, and cut score articulation. Small alterations to each of the two components were made in keeping with the goal of standards validation. Noteworthy are the following:

- Cut score location
  - Rounds were dropped from 5 to 3
  - 2007 cut scores were pre-identified in the ordered-task book
  - Only the item mapping procedure was used
  - Panelists were instructed to provide rationale for judgments
- Articulation
  - 2008 results were compared to 2007
  - Judgments were solicited in a consensus manner
  - The articulation panel made a final recommendation after reviewing the work of the standards validation content judges

At the conclusion of the panel meetings, an evaluation was conducted on the collective recommendations of the educator and articulation panels with respect to the linked scores from 2007, in order to determine the reasonableness of the linked 2007 cut scores given the changes made to the assessment. G-theory variability of judgments was used to create standard error bands around the panelist judgments. In cases where the linked 2007 cut scores fell within the standard error band, it was concluded that the judgments of the 2008 panel were consistent with retention of the linked 2007 cut scores. This was the case for 75% of the cut score decisions. For cases in which the bands did not contain the 2007 linked value, MDE content staff reviewed the rationales provided by panelists for moving the cut scores. When the rationales for moving the cut scores were compelling, the cut scores from the 2008 panels were used; otherwise the 2007 linked values were retained.

The full report of the validation activity was provided to the National TAC, an independent committee of nationally recognized technical advisors in educational measurement. They reviewed the procedures and outcomes, providing advice to the commissioner on the adoption of final cut scores. The commissioner-approved cut scores and associated impact data are provided in tables 5.31 and 5.32.

TABLE 5.31. 2008 Commissioner-Approved Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	3	14	0.4247	20	1.4305	25	2.8242
	4	14	0.4124	19	1.2769	26	3.3581
	5	10	-0.1747	16	1.1929	24	2.7802
	6	11	-0.1594	19	1.2880	25	2.7098
	7	13	0.6233	17	1.3283	22	2.1732
	8	15	0.8073	21	1.7795	25	2.8110
	11	10	-0.4758	20	1.5792	23	2.1922
Reading	3	14	0.6365	17	1.0918	23	2.1133

	4	15	0.9079	18	1.3609	23	2.2430
	5	16	1.0255	20	1.6337	24	2.4544
	6	15	0.8714	19	1.4893	22	2.0059
	7	15	0.7742	20	1.5142	24	2.3137
	8	13	0.5536	17	1.2149	24	2.4870
	10	14	0.6669	17	1.1151	26	3.3567

TABLE 5.32. 2008 Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2008 Percentage of Students in Achievement Level			
		Does Not Meet	Partially Meets	Meets	Exceeds
Mathematics	3	19	22	32	27
	4	15	23	47	15
	5	12	18	46	24
	6	12	27	40	21
	7	21	22	36	21
	8	24	27	30	19
	11	17	46	18	19
Reading	3	19	9	29	43
	4	18	12	27	43
	5	22	12	19	47
	6	18	13	16	53
	7	15	11	15	59
	8	11	9	21	59
	10	14	6	45	35

Standard Setting for Science Minnesota Test of Academic Skills (MTAS)

Standard setting for MTAS Science was held during 16-17 July 2008, with the consequential validity meeting occurring on 18 July 2008. The activities of the meeting are documented in a paper titled *Minnesota Comprehensive Assessment – II (MCA-II) Minnesota Test of Academic Skills (MTAS) Report on Science Standard Setting*. The report can be found at the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

Minnesota's testing contractor collaborated with the Minnesota Department of Education (MDE) and MDE's National Technical Advisory Committee (TAC) in designing the standard setting activities. Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE invited 10–20 participants from across Minnesota to set cut scores in each grade. The invitation approach differed from that of the MCA-II in that approximately half of the invited participants were educators involved in special education either through academic specialty or classroom experience. The details of participant credentials and demographics can be found in *Minnesota Comprehensive Assessment – II (MCA-II) Minnesota Test of Academic Skills (MTAS) Report on Science Standard*

Setting. The report can be found at the MDE website at http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/Assessments/MTAS/MTAS_Technical_Reports/index.html.

Standard Setting Meeting

Round 1

A Modified Angoff procedure was used. For each task panelists rated expected mean scores by achievement level. Specifically, panelists were asked to consider 100 students who are just barely Partially Meets the Standard, then provide a mean score, in quarter point increments, for those 100 students on the task. They then conduct the same rating procedure for Meets the Standard, and then Exceeds the Standard. After all three ratings were made for a task, they were to go on to the next task. The rating sheet is designed to provide spatial/graphical feedback during the process. Each achievement level is a row in a table. Within each row, lower average scores are on the left of the sheet and upper average scores are on the right. The rows are stacked with Partially on the top, then Meets, and finally Exceeds. Logically, ratings should result in a pattern with Partially marks to the left, Meets to the right of Partially, and Exceeds to the right of both.

Round 2

Agreement feedback was provided in the form of minimum, maximum, and median of the raw score cuts for each level of achievement. Panelists shared rationale for their ratings of each task. The facilitator asked questions to help panelists relate their judgments to the content on the tasks.

After 30 minutes panelists were given an opportunity to provide another set of ratings. Again, panelists were directed to consider the achievement level descriptors and the administration and scoring of the MTAS in making this judgment.

Round 3

Agreement feedback was provided again, with a short discussion. Total score cuts were provided and discussed.

After this discussion, the facilitator provided and discussed patterns of responding. These patterns were based on item scores, but not specific items. If a student received a 1 on a single item, and a 0 on the remaining 8, regardless of which item they responded correctly to, the pattern was “100000000”. Three patterns were specifically discussed by the facilitator: “111111111”, “333111111”, and “333222111”. The first pattern results in a score of 9, and can only be interpreted as an “effort” score. That is, the student was engaged in the test, but did not answer a single question correctly. Such a score cannot be interpreted as demonstrating any content mastery. The other two patterns are possible guessing patterns. The MTAS tasks are 3-option questions. Although the proctor is instructed to repeat the administration of a task if they feel the student chose the correct answer by guessing, it is still possible that a student could guess and be awarded credit. If they guessed correctly without the need for scripted support, they would receive a 3. If guessing correctly with scripted support, they would receive a 2. It is possible that students with no content mastery could achieve these patterns and the associated raw score levels.

Panelists discussed the patterns, and used this information, along with impact data tables, as information in refining, if desired, their cut scores at the end of this round. The impact data included disability subgroups in addition to gender and ethnicity.

Round 4

All grade-level groups were brought together to discuss the results. An opportunity to refine cut-scores was provided a final time, and panelists were asked to respond to surveys about their experience.

Training

Before each round, panelists were trained on the process to be used in the upcoming round and given the opportunity to ask questions. As part of the training for the Modified Angoff procedure, a practice test was provided, and panelists were trained on how the raw score cut values were obtained. Before any round of judgments was entered, panelists were asked to indicate in writing that they were ready to begin. If they were not ready, Minnesota's testing contractor staff re-taught the process or answered questions until everyone was ready to move forward.

Panelists reported that the training was effective and that they understood the procedures. At the end of the standard setting, panelists provided evaluations of the activity. The majority of panelists indicated they were satisfied with the process and the final outcomes.

Stakeholder Impact Panel

A group of educators and community members convened to form a combined Stakeholder Impact panel for MTAS Science and MCA-II Science. Panelists included one member from each grade group attending the previous meetings, plus additional panelists who were selected to represent the broader community. The panel commenced with a presentation of the process used by the previous committee, and the results. The cut scores and associated impact data prior to smoothing is provided in tables 5.33 and 5.34.

The facilitator led discussion about the consequences of the educator panel recommendations, and continued to explain why it was reasonable to consider alternative outcomes based on policy implications. The standard error bands were explained, including their meaning and potential usage. Panelists were provided an opportunity to see the impact of changing some cut scores both in terms of the percent correct and the percentage of students classified in each level after changes were made.

Finally, panelists were provided a judgment form and given an opportunity to offer recommendations for the cut scores. The judgment form reported the educator panel recommendations and standard error bands. Panelists were instructed that they could make recommendations within the standard error bands. The smoothed cut scores and associated impact data are provided in tables 5.35 and 5.36.

Finally, the full report of the standard setting was provided to the National TAC, an independent committee of nationally recognized technical advisors in educational measurement. They reviewed the procedures and outcomes, providing advice to the commissioner on the adoption of final cut scores. The commissioner-approved cut scores and associated impact data are provided in tables 5.37 and 5.38.

TABLE 5.33. Panelist-Recommended Cut Scores (Round 4)

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	14	0.7512	18	1.3867	21	1.8749
8	16	1.0189	21	1.8672	24	2.5870
High School	14	0.6713	20	1.6693	25	2.8875

TABLE 5.34. Impact Data Associated with Panelist-Recommended Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	15	10	13	62
8	19	15	23	43
High School	18	26	33	23

TABLE 5.35. Consequential Validity Panel's Smoothed Cuts Scores

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	14	0.7512	18	1.3867	24	2.5505
8	16	1.0189	21	1.8672	25	2.9647
High School	14	0.6713	20	1.6693	26	3.4853

TABLE 5.36. Impact Data Associated with Consequential Validity Panel's Smoothed Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	15	10	36	39
8	19	15	35	31
High School	18	26	40	16

TABLE 5.37. Commissioner-Approved Cut Scores

Grade	Cut Scores					
	Partially Meets		Meets		Exceeds	
	Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
5	14	0.7512	18	1.3867	24	2.5505
8	16	1.0189	21	1.8672	25	2.9647
High School	14	0.6713	20	1.6693	26	3.4853

TABLE 5.38. Impact Data Associated with Commissioner-Approved Cut Scores

Grade	2008 Percentage of Students in Achievement Level			
	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
5	15	10	36	39
8	19	15	35	31
High School	18	26	40	16

Standard Setting for Mathematics and Reading Minnesota Comprehensive Assessments-Modified (MCA-Modified)

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was implemented for the Minnesota Comprehensive Assessments-Modified (MCA-Modified) standard setting held in Roseville, Minnesota, on June 27–30, 2011. The standard setting for Reading MCA-Modified took place on June 27-29 and Mathematics MCA-Modified took place on June 29-30. Achievement-level cut scores were established for all grades in mathematics and reading. 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.

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 Grades 3-8 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III). Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

Participants

MDE invited approximately 12–15 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, except for Grade 11 Mathematics and Grade 10 Reading, which each had their own committee. 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*. The report can be found at the MDE website.

Table Leaders

During the standard setting, participants were divided into groups, often called "tables." Each table had one table leader that that had been previously selected by the 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

For Reading MCA-Modified, the ordered item booklets (OIB) contained 35 operational items along with 8-10 field test items in order to avoid gaps on the proficiency continuum. For Mathematics MCA-Modified, the OIB contained 35-40 operational items along with 10 field test items. 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 report can be found at the MDE website.

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. Round 1 and 2 recommendations were first completed for the lower grade, followed by Rounds 1 and 2 for the upper grade, except for the high school committees which only set standards on a single 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, or their single grade in the case of the high school committees. 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.

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 for the non-high school committees.

Round 3

At the beginning of Round 3, historical impact or relevant impact data were presented to the panelists as external reference. For Reading MCA-Modified, 2006-2011 MCA-II impact data were presented. For Mathematics MCA-Modified, 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. 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*. The report can be found at the MDE website.

Table 5.39 shows the participant-recommended cut scores, as taken from participants' Round 3 bookmark placements. Cut scores are shown on the raw score. Table 5.40 shows the impact data associated with the cut scores shown in table 5.39.

TABLE 5.39. Participant-Recommended Cut Scores (Round 3) for MCA-Modified

Content Area	Grade	Cut Scores		
		Partially Meets	Meets	Exceeds
Mathematics	5	19	22	25
	6	16	20	24
	7	17	24	26
	8	17	22	23
	11	18	23	28
Reading	5	18	25	27
	6	19	23	26
	7	20	26	28
	8	15	25	27
	10	16	23	27

TABLE 5.40. Impact Data Associated with Participant-Recommended Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	5	78	11	6	5
	6	66	24	7	3

	7	73	25	1	1
	8	66	30	1	3
	11	59	31	8	2
Reading	5	37	42	10	12
	6	43	26	15	16
	7	33	38	11	17
	8	18	61	10	11
	10	14	30	25	31

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 5.41 shows the cut scores from the vertical articulation. Table 5.42 shows the impact data associated with the cut scores shown in table 5.41.

.

TABLE 5.41. Vertical Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Cut Scores		
		Partially Meets	Meets	Exceeds
Mathematics	5	16	22	25
	6	15	20	24
	7	14	21	23
	8	15	21	23
	11	17	23	28
Reading	5	18	24	27
	6	18	23	26
	7	20	25	28
	8	16	23	27
	10	16	23	27

TABLE 5.42. Impact Data Associated with Articulation Panel's Smoothed Cut Scores

Content Area	Grade	Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	5	56	33	6	5
	6	58	32	7	3
	7	46	46	5	3
	8	45	48	4	3
	11	51	39	8	2
Reading	5	37	37	14	12
	6	37	32	15	16
	7	33	31	18	17

	8	22	44	23	11
	10	14	30	25	31

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 Mathematics and Reading MCA-Modified administration are given in Table 5.43. Impact data associated with the final cut scores are reported in Table 5.44.

TABLE 5.43. Commissioner-Approved Cut Scores

Content Area	Grade	Cut Scores					
		Partially Meets		Meets		Exceeds	
		Raw Score	Theta	Raw Score	Theta	Raw Score	Theta
Mathematics	5	17	-0.2807	22	0.3828	25	0.8230
	6	15	-0.3950	20	0.2563	24	0.8092
	7	15	-0.3835	22	0.5253	24	0.8058
	8	15	-0.4219	21	0.3773	23	0.6579
	11	17	-0.3297	23	0.3645	28	0.9847
Reading	5	18	-0.0932	24	0.7123	27	1.1934
	6	18	0.1477	23	0.7861	26	1.2237
	7	20	0.1434	25	0.8218	28	1.3216
	8	16	-0.2278	23	0.6921	26	1.1431
	10	16	-0.1709	23	0.7453	28	1.5514

TABLE 5.44. Impact Data Associated with Commissioner-Approved Cut Scores

Content Area	Grade	2011 Percentage of Students in Achievement Level			
		Does Not Meet (%)	Partially Meets (%)	Meets (%)	Exceeds (%)
Mathematics	5	64	25	6	5
	6	58	32	7	3
	7	55	40	3	2
	8	45	48	4	3
	11	51	39	8	2
Reading	5	37	37	14	12
	6	37	32	15	16
	7	33	31	18	17
	8	22	44	18	16
	10	14	30	31	25

Chapter 6: Scaling

The Minnesota assessments, such as the Minnesota Comprehensive Assessments-Series III (MCA-III), the Minnesota Comprehensive Assessments-Series II (MCA-II), Minnesota Comprehensive Assessments-Modified (MCA-Modifed), the Minnesota Test of Academic Skills (MTAS) and the Test of Emerging Academic English (TEAE), 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), 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 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 MCA-III, the Minnesota Comprehensive Assessments-Series II (MCA-II), Minnesota Comprehensive Assessments-Modified (MCA-Modified) 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

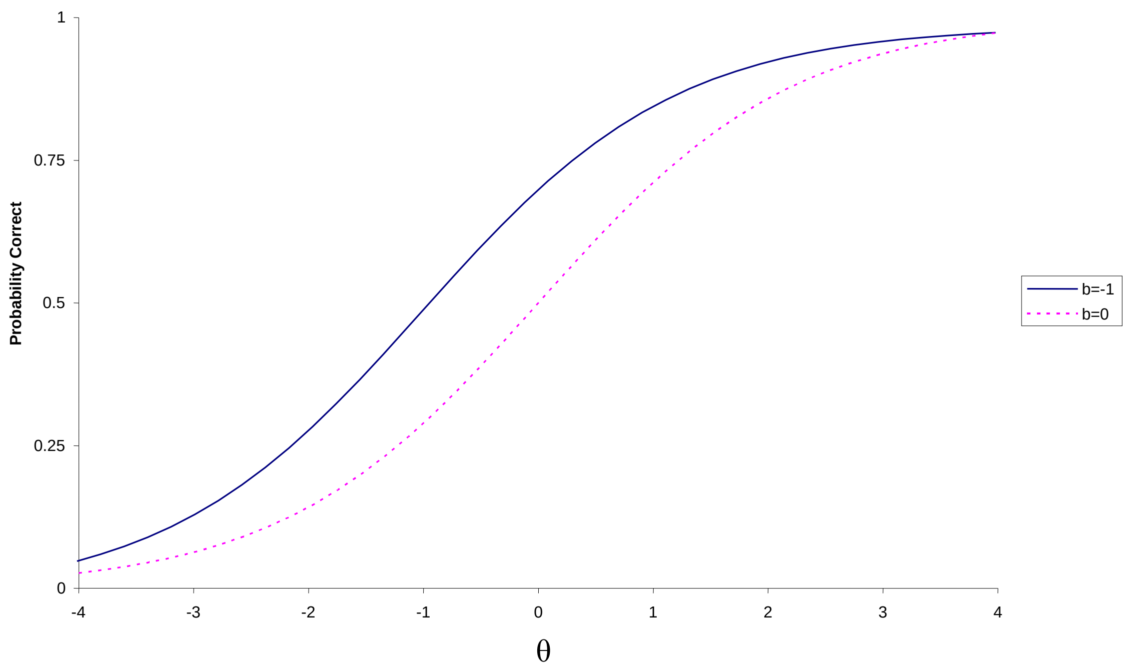
The dichotomous Rasch model can be written as the following mathematical equation, where the probability (P_{ij}) of a correct response for person i taking item j is given by:

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

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 percent 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 percent 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:

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

where $x = 0, 1, 2, ..., m_j - 1$, and,

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

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 4-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 (= –.3) is the value of θ at the crossing point of the response functions for score points "1" and "2", b_3 (= .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
b_1=-1.5; b_2=-.3; b_3=.5; b_4=2

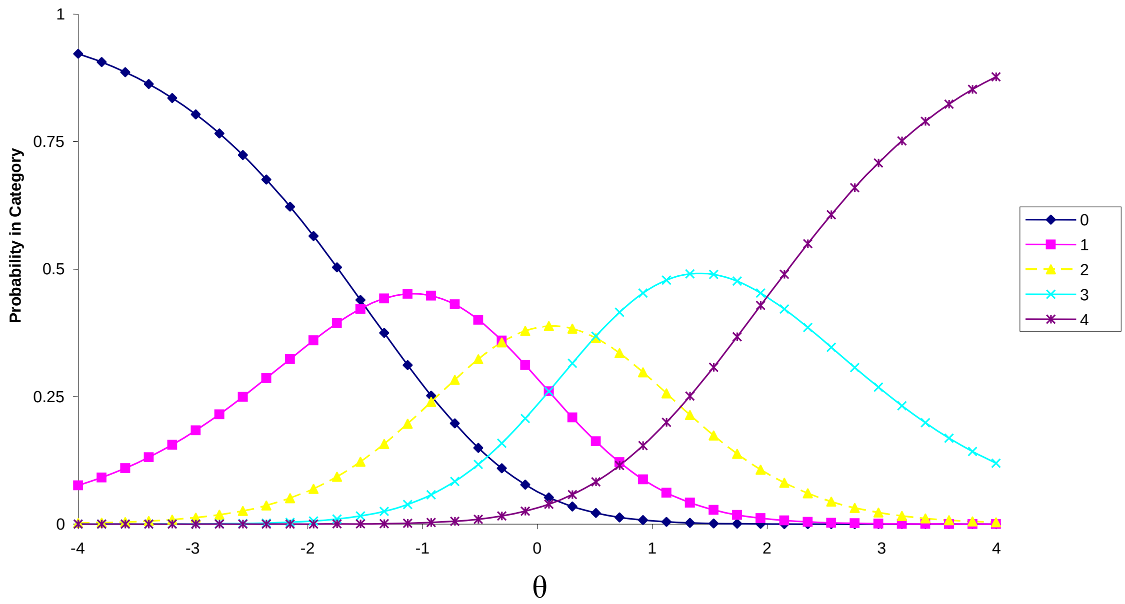
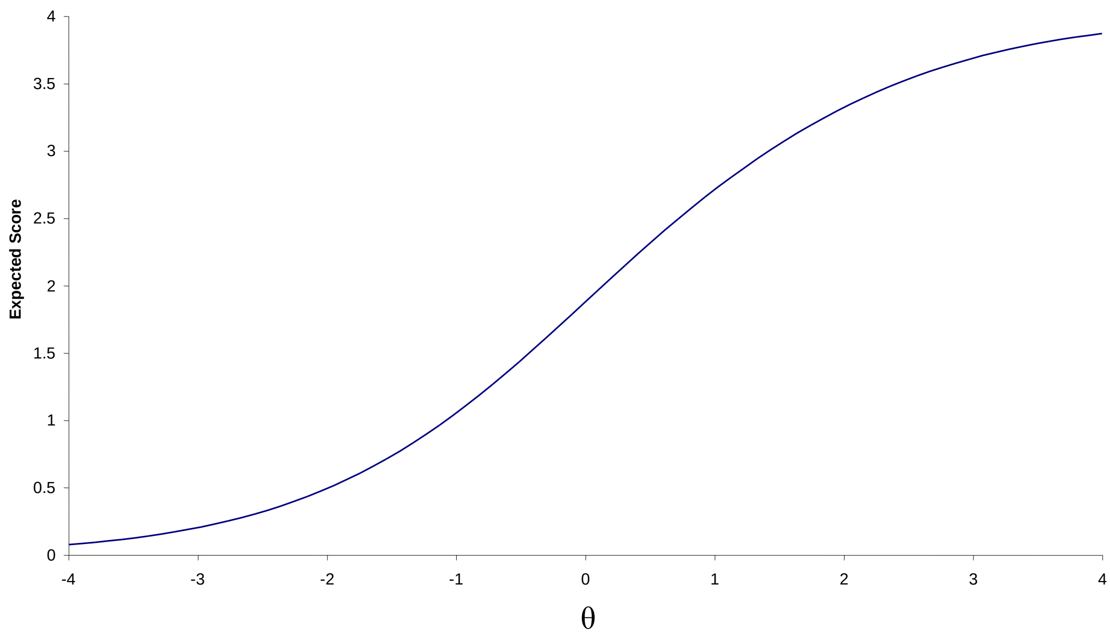


Figure 6.3 displays the average score for every ability value for the sample item given in Figure 6.2. The figure shows that students with ability θ = 0 should, on average, receive a score of "2" on the item, whereas students with ability at about 1 should average about 2.5 points on the items.

Figure 6.3. Rasch Partial Credit Model Item Expected Score Function for an Example 4-Point Polytomous 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.

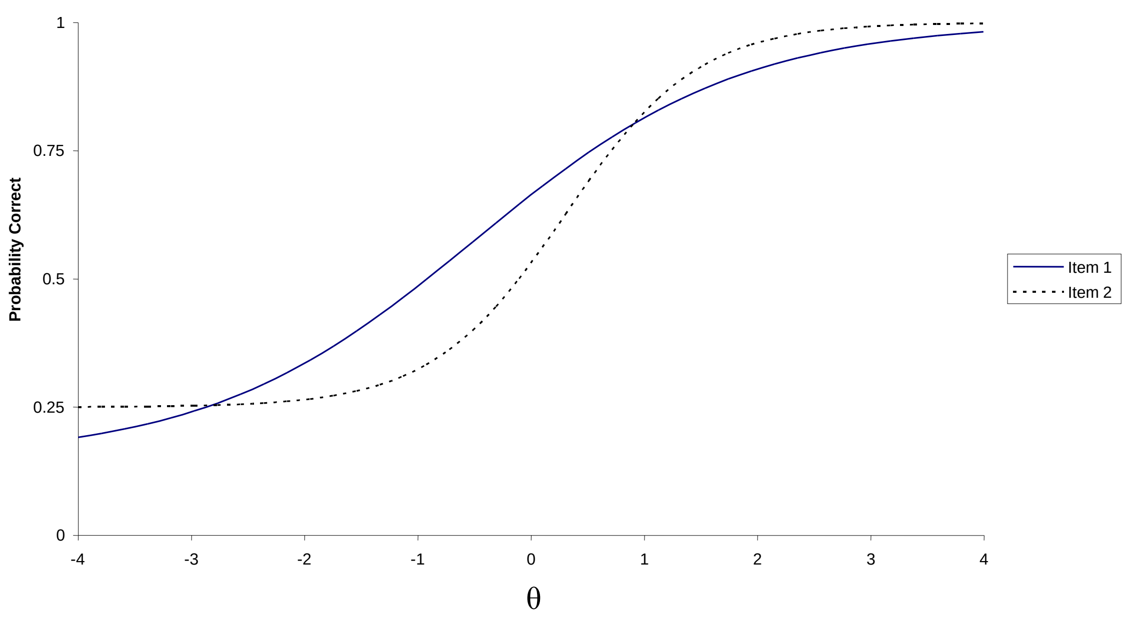
The 3PL model is mathematically defined as the probability of person *i* correctly answering item *j*:

$$P_{ij} = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7 a_j (\theta_i - b_j)\right]}, \tag{6.4}$$

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 θ > 1.0 or θ < -3.0 have a higher expected probability of getting item 2 correct.

Figure 6.4. 3PL Item Response Functions for Two Sample Dichotomous Items
Item 1: a=.5 b=-.5 c=.15
Item 2: a=1 b=.3 c=.25



The figure also shows item 2 clearly has a non-zero asymptote (*c* = .25). Item 1 also has a non-zero asymptote (*c* = .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 .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.

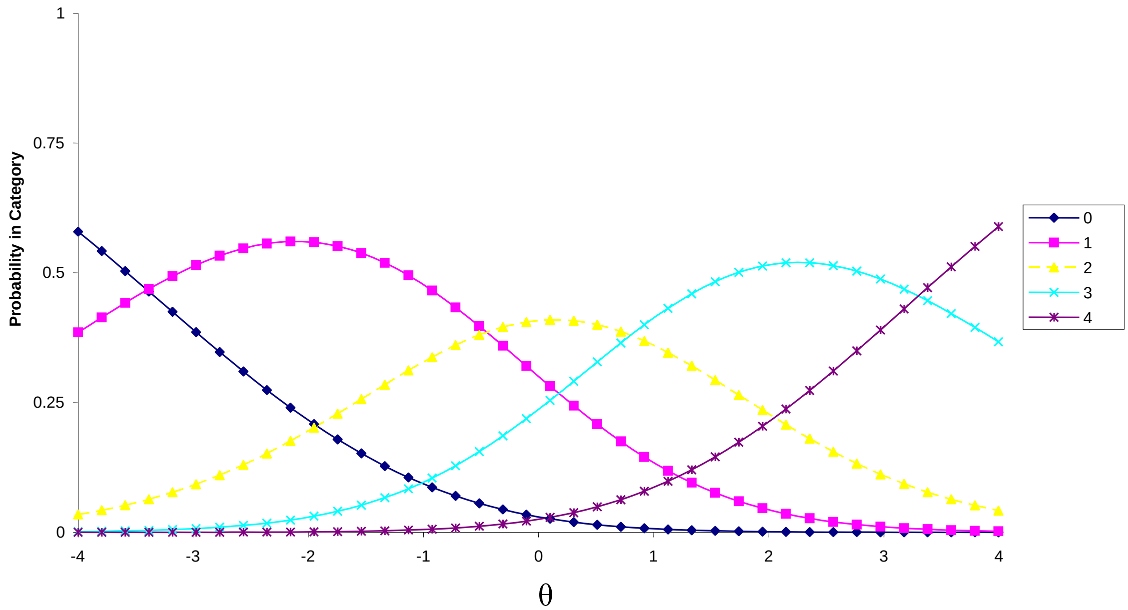
The polytomous IRT model described in this section is the GPC model. Instead of having a single probability correct, as in the 3PL model, the GPC model has a separate probability for each possible response category. The GPC model is mathematically defined as the probability of person *i* scoring in response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

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

where *m* is the number of response categories for the item and $d_{j1} = 0$ (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* = .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.

Figure 6.5. Generalized Partial Credit Model Category Response Functions for Sample Polytomous Item
a=.4; b=.3; d1=0; d2=3.7; d3=.75; d4=-.5; d5=-3

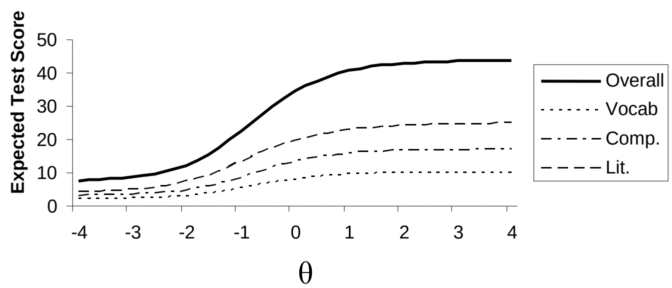


Calibration of items for the 3PL/GPC models is achieved using the computer program MULTILOG (Thissen, 1991), which estimates parameters simultaneously for dichotomous and polytomous items via a statistical procedure known as marginal maximum likelihood. Simultaneous calibration of these items automatically 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 II (MCA-II). 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 ($\theta = 0.0$) can be expected to get a score of around 35 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 three subscores are also graphed in Figure 6.6.

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



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. For the Test of Emerging Academic English (TEAE) Reading, the RPC model is used where each stimulus is considered as a polytomous item. The RPC model is also well suited to model the performance task-based Minnesota Test of Academic Skills (MTAS). However, the Minnesota Comprehensive Assessments-Modified (MCA-Modified) is composed of multiple choice items, so the dichotomous Rasch model is used for these tests. 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 is also used for the 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 the 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." (under the section "Latent Trait Estimation") The spring 2006 administration of the Minnesota Comprehensive Assessments-Series II (MCA-II) Mathematics and Reading is the baseline year for grade 11 mathematics and for all grades in reading. In administrations after 2006, the ability score metric is equated back to the spring 2006 base administration. In the case of assessments based on the Rasch measurement model (MTAS and MCA-Modified), the number right and model based scoring approaches are mathematically equivalent.

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 MCA-III makes 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 score. This is the case even for students getting 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 in chapter 7, "Equating and Linking." (under the section "Latent Trait Estimation").

Minnesota Comprehensive Assessments-Series II and 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 II (MCA-II) and 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. The MCA-II and the MCA-III have slightly different transformations to the reporting metric. These transformations are described in the next section.

Minnesota Comprehensive Assessments-Series II and Series III Transformation

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

$$Scale = (\theta_{EQ} - \theta_{Std2}) \bullet Spread + Center + Grade \bullet 100\,, \tag{6.6}$$

where θ_{EQ} 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 both MCA-II and MCA-III, the transformation formula uses cuts scores on the θ scale (see Chapter 5, "Performance Standards"). For MCA-III, the Commissioner of Education approved cut scores were already on the θ scale. For MCA-II, 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

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

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 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.

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 MCA-II uses raw to scale score conversion, some additional scoring rules are necessary. For MCA-II, 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.

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 3 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 399 (the HOSS). Because only students who score all correct are awarded a 399, students who get all but one correct would be assigned a score of 398.

One difference between the MCA-II and MCA-III scale transformations is how very low scores are assigned. For MCA-III, all students are assigned a θ score by the scoring algorithm and so no further manipulation of the score is necessary. However, MCA-II 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 MCA-II, 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:

$$Scale(y) = LOSS + y \bullet \frac{A - LOSS}{x}. \quad (6.8)$$

For both MCA-II and MCA-III, non-integer value scale values are rounded to the nearest integer value. Because 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 II 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 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 2008, a vertical scale is reported for Reading Minnesota Comprehensive Assessments-Series II (MCA-II) in grades 3 through 8 (note that the vertical scale for Mathematics MCA-II was discontinued with the move to MCA-III—a new vertical scale for MCA-III will be given for 2011-2012). 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-II 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.

Each student’s vertical IRT scale score is computed using the student’s post-equated ability estimate $\hat{\theta}_{EQ}$, the scaling constants given in the tables below, and the following formula.

$$\hat{\theta}_{VS} = a \bullet \hat{\theta}_{EQ} + b \tag{6.9}$$

The constants *a* and *b* for each grade are found below in table 6.1. Note that because $\hat{\theta}_{EQ}$ has already been placed on the 2006 operational IRT scale through the procedures described in chapter 7, “Equating and Linking,” the theta metric of the vertical scale is also on the 2006 IRT scale. The minimum and maximum values from table 6.1 represent the lower and upper limits for the vertical IRT scale score. Minimums and maximums were chosen to allow sufficient range for scores within a grade level while at the same time preventing the maximum for a particular grade from exceeding the maximum of the next higher grade.

TABLE 6.1. Reading MCA-II Vertical Scaling Constants on the Theta Metric

Grade	a	b	Minimum	Maximum
3	0.9863	−0.8216	−6.9132	1.2841
4	1.0273	−0.3361	−6.6705	2.8199
5	1.0000	0.0000	−6.4278	3.3903
6	1.0106	0.2550	−6.1851	4.2751
7	0.9627	0.4860	−5.9424	4.3791
8	0.9378	0.6081	−5.6997	4.5696

The transformation to the reported Progress Score is achieved through the equation,

$$PS = \hat{\theta}_{VS} \bullet A_{PS} + B_{PS}\ , \tag{6.10}$$

where PS is the student’s Progress Score. Non-integer values are rounded to the nearest integer value. The constants A_{PS} and B_{PS} are given in table 6.2 below. The scaling constants were chosen so that a score of 3500 on the Progress Score corresponded with the Grade 3 *Meets the Standards* cut score, and so that a score of 4000 corresponded with the Grade 8 Meets cut. Table 6.3 gives the minimum and maximum scores on the Progress Score. Only students with all correct raw scores are awarded the maximum Progress Score. If other raw scores map to Progress Scores at or exceeding the maximum, students with these scores are assigned Progress Scores one less than the maximum.

As was the case for horizontal scaling, linear interpolation is used to scale raw scores below chance level. Boundary points for the interpolation are *x*—the lowest raw score *above* chance level—and 0, for a raw score of zero correct. The raw score *x* is assigned Progress Score *A* using the above transformation equation. A raw score of zero is assigned the Lowest Observable Progress Score (LOPS). For a raw score *y* between zero correct and *x*, Progress Scores are assigned using the following interpolation equation,

$$PS(y) = LOPS + y \bullet \frac{A - LOPS}{x} \tag{6.11}$$

For non-integer values, the Progress Score value is rounded to the nearest integer value. Note that in cases where the Progress Score for *x* is already the LOPS value, interpolation will not be necessary.

TABLE 6.2. Progress Score Scaling Constants for Reading MCA-II

Subject	A_{PS}	B_{PS}
Reading All Grades	247.208	3909

TABLE 6.3. Reading MCA-II Progress Score Minimums and Maximums

Grade	Reading	
	Minimum	Maximum
3	2200	4226
4	2260	4606
5	2320	4747
6	2380	4966
7	2440	4992
8	2500	5039

On the student ISR, the Progress Score is given for grades 3 through 8 for each year that the student has taken the MCA-II, beginning with the inception of the MCA-II in 2006. For example, if a student took the MCA-II as a third grader in 2006, as a fourth grader in 2007, as a fifth grader in 2008, as a sixth grader in 2009, and a seventh grader in 2010, 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 Comprehensive Assessments-Modified and the Minnesota Test of Academic Skills Scaling

The general transformation formula used to obtain scale scores for the Minnesota Comprehensive Assessments-Modified (MCA-Modified) and the Minnesota Test of Academic Skills (MTAS) is as follows:

$$Scale = (\theta_{eq} - \theta_{Std2}) \bullet Spread + Center \text{ ,} \tag{6.12}$$

where θ_{eq} 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 MCA-Modified and the MTAS use the same transformation equation.

Chapter 5, “Performance Standards”, describes the process of setting the standards for the MCA-Modified and 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 MCA-Modified and the MTAS scales..

As with the MCA-II and 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 (6.12) allows two of the three scale cut scores to be fixed. For all grades and subjects of the MCA-Modified and 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 prior to 2008, 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 MCA-Modified and the MTAS was selected to force the first scale cut score to be equal to 190. The formula used to find the *Spread* is

$$Spread = 10 / (\theta_{Std2} - \theta_{Std1}) \text{ ,} \tag{6.13}$$

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.

Test of Emerging Academic English Reading Scaling

The Test of Emerging Academic English (TEAE) Reading number correct score is transformed to place it on a standardized vertical or developmental scale. The equating procedures used allow TEAE tests of different years and grades to be reported on the same scale, facilitating comparisons over time and recording students’ growth in language proficiency. The standardized vertical scale encompasses all grades, regardless of which level in the series the student completes. The Language Proficiency Test Series (LPTS) Reading scale, upon which the TEAE Reading scale is based, is a vertical scale. The TEAE scale was equated to the LPTS scale. This equating was accomplished by administering LPTS test forms along with the TEAE forms to all students participating in the 1999 TEAE pilot test administration.

The TEAE Reading scale score is derived from the student’s θ estimate, using the formula:

$$SS = 50 \bullet \theta_{est} + 150 \text{.} \tag{6.14}$$

Scale scores on the TEAE Reading test can range from 1 to 450.

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 II (MCA-II) and 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 (6.7), 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 to be implemented in 2011-2012, which are based on vertical scales, are intended to provide an appropriate basis for making comparisons across years within a subject area.

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.

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. 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 mathematics from those who took the MCA-II in past years. Thus, for example, it is not appropriate to compare the grade 5 mathematics score from the current year to the grade 5 average from previous years. However, limited and focused linking procedures or prediction analyses may still serve useful purposes.

Minnesota Test of Academic Skills and Minnesota Comprehensive Assessments-Modified

The same caveats, cautions and guidelines presented above for the MCA-II and the MCA-III also apply to interpreting the scales scores of the Minnesota Comprehensive Assessments-Modified (MCA-Modified) and the Minnesota Test of Academic Skills (MTAS). Although the MCA-Modified and the MTAS uses identical scale numbers to represent *Meets* and *Partially Meets* cuts across subjects and grade levels, the scale score metrics cannot be assumed to be equivalent across grades, subjects, or testing programs. As with the MCA-II and the MCA-III, the best way to compare across grades or subjects for the MTAS or MCA-Modified is through achievement levels. It must be remembered, however, that the MTAS achievement levels refer to the Alternate Achievement Standards. That is, the MTAS measures student progress on state grade-level content standards but at reduced breadth, depth and complexity and judged against a different definition of proficiency. Similarly, the MCA-Modified is based on modified achievement levels. The distinctions between the achievement levels and standards used on these tests much be kept in mind when making comparisons between the MCA, the MCA-Modified, or the MTAS tests.

Test of Emerging Academic English and Minnesota Student Oral Language Observation Matrix

The Test of Emerging Academic English (TEAE) and Minnesota Student Oral Language Observation Matrix (MN SOLOM) forms administered in 2011 may be assumed to be equivalent to those used in previous years. TEAE Reading forms have been equated statistically. TEAE Writing forms use a consistent scoring rubric, and prompts are selected to match difficulty across forms. The MN SOLOM checklist is unchanging across administrations. As a result, valid comparisons can be made within a grade level across years (between, for example, grade 3 scores in 2011 and grade 3 scores in 2006). Further, the existence of a developmental (vertical) scale for the TEAE Reading test allows users to make valid year-to-year comparisons of TEAE Reading scores for individuals or groups of students across grades. They can do so even when the students shift from one form of the test to another, as occurs when the student moves from grade 4 to grade 5 or grade 6 to grade 7.

Conversion Tables, Frequency Distributions and Descriptive Statistics

The Yearbooks provide tables for converting raw scores to derived scale scores 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 tests 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 is 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 having difficulty similar to 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.

The Test of Emerging Academic English (TEAE) rotates previously administered forms from year to year. Because the test forms administered this year have already been operationally employed and have been through the full equating process, no further equating is required.

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 II (MCA-II), each new assessment is constructed from a pool of items equated to the base test form (2006 for Reading and Grade 11 Mathematics; 2008 for Science).

Test Construction and Review

Test construction begins by selecting the operational (or base) items for an administration. These items are given on every test form for that administration, and they count toward the individual student's score. Using the items available in the item pool, psychometricians from Minnesota's testing contractor construct new forms by selecting items meeting 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 to ascertain whether the test has similar psychometric properties (for example, difficulty) as 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 to ensure each test will be of the highest possible quality.

Field-Test Items

Once a newly constructed item has survived committee review and is ready for field-testing, it is embedded in a test booklet among the operational test items. For example, in a particular grade's Reading MCA-II 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 reading passages 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-II the role of linking is usually reserved for items that have been administered operationally in a previous year.

Forms are spiraled within testing sites (usually classrooms) across the state so a large representative sample of test takers would respond to the field-test items. For example, at grade 10, with a statewide enrollment of approximately 65,000, approximately 4,300 students would respond to each form. 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 fielded 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 fielded 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, 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-II administration there are at least eight internal linking items, and eight to sixteen external linking items. Linking items are chosen so the set of linking items gives good coverage of the benchmarks as well as approximating the overall difficulty and information spread of the operational items.

Some administrations of the MCA-II and the MCA-III in grades 3–8 have included off-grade vertical linking items. They are multiple-choice items from an adjacent grade's test that are used to link grades on the Progress Score. Off-grade vertical linking items do not count toward a student's score. For administrations after 2006 for MCA-II or 2011 for MCA-III, not all students will take off-grade vertical linking items. The tests for grades 10 and 11 do not contain vertical linking items since no growth measure 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 districts are excluded from the equating because their data arrived late or they failed to clear the scoring and editing process in time to be used in the equating. This, however, only represents a small percentage of total students by grade and subject (usually less than one percent).

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 Minnesota Assessments, the general rule is that students must respond to at least four machine-scoreable questions in each segment of the test in order to be classified as "attempted." In addition, the responses of home school and private school students are excluded from the calibration data set. Home school and private school students are not required to take the MCA-II or MCA-III to be included in statewide summary statistics or included 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 Equating Procedures

Once the statewide data file has been edited for exclusions, a statistical review of all operational items is conducted before beginning item response theory (IRT) calibration. Items are evaluated for printing or processing problems. A multiple-choice 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. Gridded-response and figural response items are flagged for low mean scores or low item-total correlations. Constructed-response items are flagged for unusual score distributions. Any flagged items are reviewed in the published test books to ensure the item was correctly printed. Also, flagged items 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 the MCA-II and MCA-III, the commercial software MULTILOG version 7 is used for all item calibrations. The 3-parameter logistic model is fit to student responses to multiple-choice (MC) and most technology enhanced (TE) items, the 2-parameter logistic model is fit to responses to gridded-response (GR) items and open-ended TE items, and the generalized partial credit model is fit to responses to the constructed-response (CR) items. These models are described in chapter 6, "Scaling." All operational items for a test (MC, TE, GR and CR) and external linking items are calibrated simultaneously. After obtaining the linking item parameter estimates on the current administration's operational scale, another scaling is performed 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 appears 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 raw scores can be placed on the reporting metric as described in chapter 6, "Scaling."

MTAS and MCA-Modified Equating

The commercial software package WINSTEPS (Linacre, 2006) is used for the calibration of the Minnesota Comprehensive Assessments-Modified (MCA-Modfied) items and the Minnesota Test of Academic Skills (MTAS) performance tasks. As described in chapter 6, "Scaling," the IRT model used for calibration is the dichotomous Rasch model or the Rasch Partial Credit model (Masters, 1982). For some MCA-Modified and MTAS administrations, a combined operational and field-test design is employed. After item or task calibration, MDE staff select the nine tasks at each grade-level to be designated as operational. The base year for the Reading MTAS, the Science MTAS, and grade 11 Mathematics is 2008. For grades 3 to 8 of Mathematics MTAS and for all grades and subjects of MCA-Modified, the base year is 2011.

Equating to the base year is accomplished using conceptually similar procedures to those used with MCA-II. 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 to insure 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 to 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 .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

Placing Field-Test Items on Operational Scale

The next step in the equating process is to place the item parameter estimates for the field-test items onto the same scale as the equated operational test items. All items, operational and field-test, are calibrated simultaneously. The Stocking-Lord procedure is used to find the scaling constants to transform the operational item parameter estimates of the combined calibration to the equated item scale. These same constants are then applied to the field-test items.

Item Pool Maintenance

The next step is to update the item pool with the new statistical information. The new 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.

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 II (MCA-II), linking is used for two purposes: (1) scaling across grades with the Progress Score; and (2) linking the Reading MCA-II and GRAD to the Lexile® reading scale.

Linking Grades 3 through 8 with the Progress Score

Vertical scales, also called growth scales, are designed to help evaluate how much students improved from one year to the next. Use of vertical scales is becoming more common as the desire of educators and parents to obtain information about student growth increases. The Reading MCA-II uses a vertical scale called the Progress Score. Linking for the Progress Score was accomplished by using common items on adjacent grades on the 2006 MCA-II 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 for the 2006 administration, 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 2006 operational scale. After off-grade items were scaled for each grade, another scaling process was conducted to place the items of grades 3 through 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, 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-II may contain vertical linking items for purposes of maintaining the scale of the Progress Score.

Linking Reading MCA-II and GRAD to the Lexile® Scale

In the spring of 2010, MetaMetrics, Inc. conducted a study to link scores on the Reading MCA-II and GRAD to the Lexile scale. Lexiles are a widely-used metric of reading difficulty used to inform reading instruction and selection of appropriate reading materials for students. A detailed description of the Minnesota linking study is provided in the document *Linking the Reading MCA-II with the Lexile Framework*, which is available upon request from MDE. Minnesota students at schools that volunteered to participate in the study completed grade-specific Lexile linking tests subsequent to their participation in the census administration of the Reading MCA-II and GRAD assessments. In brief, MetaMetrics used linear regression models to develop predictions of Lexile scores from Reading MCA-II and GRAD scale scores at each grade level. Selection of this particular linking approach reflected MDE concerns about the psychometric equivalence of MCA-II and Lexile reading constructs as well as the intended purpose of the linkage, i.e., prediction of Lexile scores. This approach and its implementation were approved by Minnesota's Technical Advisory Committee. MetaMetrics constructed conversion tables that provide predicted Lexile scores and associated 68% prediction intervals for all obtainable Reading MCA-II and GRAD scale scores. The predicted Lexile scores and prediction intervals will be reported for individual students taking those Minnesota reading assessments. It should be noted that the reported prediction intervals were empirically determined, and differ from the fixed 150-point Lexile score ranges (Lexile score -100 to Lexile score +50) typically employed in Lexile reports. More detailed information about the Lexile Framework and interpretation of Lexile scores is available at the Lexile website (www.Lexile.com).

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) uses 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 in chapter 6, "Scaling." 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 score 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.

Minnesota's testing contractor used a proprietary software program called IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2007) to conduct pattern scoring for the spring 2011 administration of the mathematics MCA-III for grades 3-8. The program has been extensively tested and compared to commercially available software programs (e.g., MULTILOG, PARSCALE; Tong, Um, Turhan, Parker, Shin, Chien, & Hsu, 2007). The report concluded that with normal cases the ISE program was able to replicate MULTILOG and PARSCALE theta estimates. However, "in problem cases, such as monotonically decreasing likelihood functions, in which MULTILOG and PARSCALE both produced theta estimates, ISE was able to produce the estimates that yielded the largest likelihood function, in alignment with the definition of the maximum likelihood algorithm" (p. 9). In addition, "with problem cases in which MULTILOG and PARSCALE failed to produce theta estimates, ISE was able to produce an estimate that yielded the largest likelihood from the likelihood function of a given response pattern" (p. 9). With regard to the CSEM, ISE produced similar results to MULTILOG. More information about the ISE program can be found in the user manual, the technical manual, and the evaluation report, which are available upon request from Pearson.

Raw-to-theta Transformation

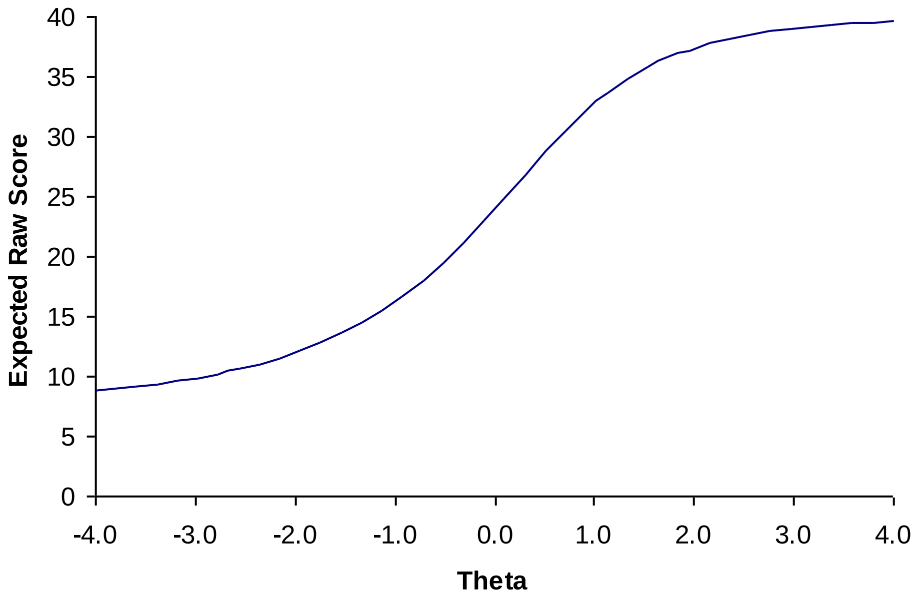
The raw-to-theta transformation can be described as a reverse table lookup on the test characteristic function. The test characteristic function can be defined as

$$TCF(\theta)=\sum_{j=1}^{N}\sum_{k=0}^{m-1} kP_{jk}(\theta)\,, \tag{7.1}$$

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 $P_{jk}(\theta)$ 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 7.1 presents the test characteristic function for a hypothetical 40-

item multiple-choice test. For example, based on Figure 7.1, persons with θ proficiency equal to 1.0 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.

Figure 7.1: Example Test Characteristic Function for 40-Item Test



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.

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 (that is, 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,

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

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)).

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

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 yielding 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 assumption that inter-item covariances constitute true-score variance and the fact that the average true score variance of items is greater than or equal to the average inter-item covariance. The formula for coefficient alpha 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, $s_{Y_i}^2$ is the sample variance of the i^{th} item (or component) and s_X^2 is the observed score sample variance for the test.

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.”

The Yearbook provides coefficient alpha for each grade and subject by gender and ethnicity. Within each table, coefficient alpha estimates are provided for the entire test, as well as each major subscale. Included with coefficient alpha in the tables is the number of students responding to the test, the mean score obtained by this group of students, the standard deviation of the scores obtained for this group, and the standard error of measurement (SEM).

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 (ten or fewer). Results involving subscores must be interpreted carefully, as in some cases these measures have low reliability due to the limited number of items attached to the score.

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 standard deviation of the total test (standard deviation of the raw scores) and $\rho_{XX'}$ is the reliability estimate for the set of test scores.

Use of the Standard Error of Measurement

The 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 true score would result in a range of values most likely to contain the student’s observed score. The observed score may be expected to fall within one SEM of the true score 68 percent of the time, assuming that measurement errors are normally distributed.

For example, if a student has a true score of 48 on a test with 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 student’s true score would result in a score range of 43.81 to 52.19 (that is, 48 ± 4.19). Furthermore, if it is assumed the errors are normally distributed and if this procedure were replicated across repeated testings, this student’s observed score would be expected to fall within the ±1 SEM band 68 percent of the time (assuming no learning or memory effects). Thus, the chances are better than 2 out of 3 that a student with a true score of 48 would have an observed score within the interval 43.81-52.19. . 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; an interval of ± 1.96 SEMs around the true score covers the observed score with 95 percent probability and is referred to as a 95 percent confidence interval. It is *not* the case that a ±1 SEM band around the *observed score* will include the true score 68% of the time (Dudek, 1979). Whereas true and error scores are uncorrelated, observed and error scores *are* correlated, as error is a component of observed score. Thus, observed score is a biased estimator of true score, and the correct approach to constructing a confidence band for true score requires centering the confidence band on the observed score adjusted for unreliability. Still, it is common practice to use a confidence band around the observed score as a rough approximation to the true score range.

The SEM is reported for Minnesota assessments in the Yearbooks in the summary statistics tables. The SEM is reported for total scores, subscores and scores of each breakout group.

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 the Minnesota Comprehensive Assessments-Series II (MCA-II), Minnesota Test of Academic Skills (MTAS), or Test of Emerging Academic English (TEAE), the mathematical statement of CSEM is

$$\text{CSEM}(O_X \mid \theta) = \sqrt{\left[\sum_{X=0}^{MaxX} O_X^2\, p(X \mid \theta)\right] - \left[\sum_{X=0}^{MaxX} O_X \cdot p(X \mid \theta)\right]^2}\,, \tag{8.6}$$

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. $p(X \mid \theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud and 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 MCA-III, which employs pattern scoring based on the 3PL measurement model, the CSEM of the scale score is calculated from the CSEM of the obtained θ estimate:

$$CSEM(Scale) = Spread \cdot CSEM(\theta). \tag{8.7}$$

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

$$CSEM(\theta) = 1/\sqrt{I(\theta)}\ . \tag{8.8}$$

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 (that is, 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 (that is, 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 $S_e = \frac{10}{\sqrt{20}} = 2.2$. (8.10)

A confidence bound around the class average would indicate that one could be 68 percent 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 percent 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 sixty responses for reading and every ninety 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 constructed-response and/or 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 II (MCA-II) and Series III (MCA-III), the Minnesota Comprehensive Assessments-Modified (MCA-Modified) 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*. For the Test of Emerging Academic English (TEAE) Reading, the levels of achievement are Level 1, Level 2, Level 3 and Level 4. For TEAE Writing, the levels of achievement are Level 1, Level 2, Level 3, Level 4 and Level 5. 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 item response theory (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) (that is, 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 $f(\theta)$ is the density function associated with the true score. Because $f(\theta)$ is unknown, the observed probability distribution of student theta estimates is used to estimate $f(\theta)$ 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 8.9. 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*. The numbers under the achievement level labels (for example < 21) are the range of raw scores that apply for the particular achievement level In this example, total decision consistency is 77.7 percent (sum of diagonal elements).

TABLE 8.1. Example Classification Table

Achievement Level		True Category				Expected %	% Accuracy
		D (<21)	P (21-30)	M (31-41)	E (>41)		
Expected Category	D (< 21)	11.7	3.4	0.0	0.1	15.2	77.7
	P (21–30)	1.9	16.1	6.1	0.2	24.3	
	M (31–41)	0.0	3.4	29.4	6.1	38.9	
	E (> 41)	0.0	0.0	3.6	17.9	21.6	
True %		13.7	22.8	39.2	24.3		

NOTE: Calculation based on Rudner (2005), equation 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 90.1 percent. The percentage of examinees incorrectly classified as *Partially Meets the Standards* or lower, when their true score indicates *Meets the Standards* or above, is 3.4.

The Yearbook contains a table with the overall classification accuracy for each grade and subject of MCA-III, MCA-II, MCA-Modified, 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 if the test measures what it purports to measure. During the process of evaluating if 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, the test content may not span the entire range of the construct to be measured, etc. 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 equally 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 section "Cautions for Score Use") and chapter 6, "Scaling" (in the section "Scale Score Interpretations and Limitations").

Demonstrating that a test measures what it is intended to measure and 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 (say college grade point average (GPA)) would not be available until the following year.

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 like the Minnesota Comprehensive Assessments-Series II (MCA-II) or 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-II or 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-II or 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 allowing students to demonstrate their mathematical ability.

Evaluating content validity is a subjective process based on rational arguments. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity only speaks 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 could only be reached if it could be shown or argued that the student put forth their best effort, the student was not distracted during the test and the test did not contain a bias preventing 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 (that is, 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 and the National Council on Measurement in Education, 1985) provides the following list of possible sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain 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 which 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 proven, 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 fifth edition *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and the National Council on Measurement in Education, 1999) recommends establishing the validity of a test through the use of a *validity argument*. This term is defined in the *Standards* as "An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores."

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 of 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 like 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 of 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 II (MCA-II) or 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-II and the MCA-III are 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-II and 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. These sections are the evidence for the scoring of performance items and 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 to describe 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 unidimensional. To test this assumption, a principal components analysis is performed. The scree plots for the principal component analyses for each subject and grade are given in the Yearbooks. The topography of the scree plot indicates the dimensionality of the data. At the point where the scree plot becomes flat, adding further dimensions to the analysis become irrelevant. The Yearbook scree plots show the first dimension is markedly different from the rest, with the slope flattening beginning with the second dimension. This type of result in a scree plot is evidence Minnesota assessments measure a single dimension.

Another check for unidimensionality 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 (that is, 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 Minnesota Comprehensive Assessments-Series III (MCA-III) 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 the spring of 2011, a mode comparability study was conducted using a matched group study design to compare students taking one of the online operational test forms versus a similar form given on paper. 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 item not impacted by mode. The complete MCA-III comparability report can be found at the MDE website.

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. Static tests, such as the Test of Emerging Academic English (TEAE), developed prior to No Child Left Behind (NCLB), did not have strong alignment to standards and did not undergo a comprehensive review process.

A sequential review process for committees is used by MDE and was outlined in chapter 2. 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 II (MCA-II) 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 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.

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 the IRT model fits well. Item-fit statistics and tests of unidimensionality 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 comprising 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 demographic subgroups making 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 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 Minnesota Comprehensive Assessments-Modified (MCA-Modified) uses the Minnesota Academic Standards, but has modified achievement standards. The MCA-Modified is an assessment designed to provide increased access to grade-level assessment tasks for students with disabilities and is intended to address the needs of students for whom neither the MCA nor the MTAS is an appropriate assessment choice. Chapter 2, "Test Development," describes in detail the process of aligning the modifed achievement standards to the general standards and serves as validity evidence documentation for the MCA-Modified.

The use of different item types also increases the validity of Minnesota assessments. The combination of multiple-choice, gridded-response 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-II or MCA-III. Because no other assessment is likely to be found to measure the standards as well as the MCA-II or 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-II, or 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-II or the MCA-III, the converging evidence from them would provide justification for extrapolation.

In 2003, a study investigating criterion validity was conducted by the National Center on Educational Outcomes at the University of Minnesota for the TEAE Reading and Writing. A comparison was made of two language proficiency tests and two state achievement tests: Language Assessment Scales (LAS) used to place students in language proficiency classes, TEAE Reading and Writing, MCA Reading and BST Reading. Results of the study showed underlying reading skills measured by the LAS and TEAE were closely related, while the writing skills were not related (Albus, Klien, Liu, & Thurlow, 2004). Strong relationships were observed between the TEAE Reading and MCA and BST Reading scores. The study involved a small number of students with little representation from greater Minnesota, thus limiting their generalizability.

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-II and 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-II and the MCA-III are a valid measures of student achievement on the standards, individual and school-level scores are not valid if students do not take the test seriously. The incorporation of the Graduation-Required Assessment for Diploma (GRAD) into the MCA-II increases the consequences of the test for high school students; this may mitigate concerns about student motivation affecting test validity. Also, as students are made fully aware of the potential No Child Left Behind (NCLB) ramifications of the test results for their school, this threat to validity should diminish.

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 Yearbooks include frequency of response tables for each scorer of each constructed-response item. The 2009 Yearbook data show non-response rates for Minnesota Assessments to be approximately 6 percent or less.

One of the most important inferences to be made concerns the student’s proficiency level, especially for accountability tests like the MCA-II, MCA-III, the MCA-Modifed, 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 discusses the details of setting performance standards, and chapter 6 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-II and 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 impacts 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 generalizability validity arguments for Minnesota assessments are also quite strong. Reliability indices, model fit and dimensionality studies provide consistent results, indicating the Minnesota assessments are properly scored and 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-II or 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: Constructed-Response Items and Written Compositions

Some Minnesota assessments, including the Minnesota Comprehensive Assessments-Series II (MCA-II) in science, require students to construct their own response to some of the test questions. For example, examinees may be required to provide a short written response to demonstrate the application of a scientific concept. For writing tests such as the Graduation-Required Assessment for Diploma (GRAD) Written Composition, students are required to write essays based on a given prompt. The procedure for scoring constructed-responses and written essays is described in this chapter. Also described at the end of the chapter is the phasing out of human scores constructed response items from the MCA-II assessments, to conform with Minnesota law enacted by 2009 Minnesota Legislative Session, House File 2. This section from the chapter not only details the timing for when constructed response items are eliminated from the different subject areas, but also describes the steps taken to ensure that the loss of constructed response items from the tests does not change the content and constructs being measured.

Scoring Process

Outlined below is the scoring process that the Minnesota's testing contractor follows. This procedure is used to score responses to all constructed response and written composition items for the Minnesota assessments.

Rangefinding and Rubric Review

For pilot and field-test items, rangefinding is done as part of the scoring process. Small scoring teams at PSC, led by a scoring director, review the rubric for a particular item and review a sampling of the student papers for that item before assigning scores. Problematic issues are discussed with the Minnesota Department of Education (MDE). After a consensus has been reached, the team scores all papers for that item. Group discussion takes place for problematic papers. After rangefinding, the scoring director constructs an exemplar set with papers for each score point for each item. Those sets are sent to MDE for review and approval before scorer training begins.

Prior to scoring the operational assessment, the subject teams of the PSC conduct rangefinding and rubric review activities with panels of Minnesota educators. In conjunction with MDE, PSC conducts a review of the rubrics used immediately prior to rangefinding. This establishes a baseline among all the participants. Minnesota's testing contractor reviews the rubrics with MDE and the participants on an as-needed basis throughout the course of rangefinding.

Rangefinding materials are chosen from field-test responses. The PSC staff assembles those materials with enough copies so all members of the rangefinding committees have working copies. The thoughtful selection of papers during rangefinding and the subsequent compilation of anchor papers and other training materials are essential to ensuring that scoring is conducted consistently, reliably and equitably.

Teams review a sufficient number of papers from the field tests in order to select a representative sample of the papers for inclusion in the training sets. Often, this number is in excess of 200 papers.

The PSC's scoring team selects exemplar papers for constructed-response items for reading and mathematics as well as writing test prompts. Exemplar papers are selected from field-test materials to provide a representative sample of a wide range of Minnesota's school districts.

The primary task in the selection of training papers is the identification of anchor papers—examples that clearly and unambiguously represent the solid center of a score point as described in the rubric. Those anchor papers form the basis not only of scorer training but of subsequent discussions as well. The rangefinding team compiles careful notes during its preparation of training sets, and those notes are used to support decisions when replacement responses must be identified.

The goal of the rangefinding meetings is to identify a pool of student responses that illustrate the full range of student performance in response to the prompt or item and generate consensus scores. This pool of responses will include borderline responses—ones that do not fit neatly into one of the score levels and represent some of the decision-making problems scorers may face. As the final step in selecting the exemplar and anchor papers, the team members review all the papers that have been assigned the same score point as a check for intra-year consistency of decision-making.

All reasonable steps are taken throughout preparation of the rangefinding materials and during the meetings to ensure security, including storing the materials in locked facilities and locking unattended meeting rooms. All rangefinding materials are accounted for at the conclusion of each session.

Following rangefinding and the approval of selected training papers, anchor sets are assembled. Drawing from the pool of additional resolved student responses, scoring leaders construct the practice sets to be used in scorer training. As those sets are assembled, they are forwarded to MDE for review and approval, as further assurance that panel decisions have been accurately enacted.

Recruiting and Training Scorers

Highly qualified scorers are essential to achieving and maintaining a high degree of reliability in scoring students' responses. Thus, the careful selection of professional scorers to evaluate the constructed-response items and writing tasks is critical in scoring the Minnesota assessments. Minnesota's testing contractor has compiled a personnel database containing the academic training and professional experience of more than 4,500 college graduates who have completed the stringent selection process for scorers. This process requires that each candidate successfully complete a personal interview, a written essay assignment and a grammar and editing or a mathematics and science test when appropriate. Such pre-screening of candidates ensures only scorers of the highest caliber are selected. Throughout the selection process, Minnesota's testing contractor actively emphasizes the need for ethnic and racial diversity among professional scorers. Included in this diverse pool is a core group of veteran scorers whose insight, flexibility and dedication have been demonstrated while working on a range of performance assessments.

Scoring supervisors are chosen from the pool of scorers based on demonstrated expertise in all facets of the scoring process, including strong organizational abilities and training skills. Individuals chosen to perform these assignments possess practical skills, leadership abilities and sensitivity to interpersonal communication requirements. Supervisors also possess the essential capability of assimilating and helping scorers understand the particular scoring requirements of MDE.

Upon being hired, scorers sign a confidentiality agreement in which they pledge to keep all information and student responses confidential. Scorers and scoring supervisors are trained to thoroughly learn the rubric and score responses according to the scoring guides developed for the specific assessment.

At the beginning of each scoring project, all scoring supervisors and scorers assigned to the project complete project-specific training.

Training

Thorough training is vital to the successful completion of any scoring. Subject leaders follow a series of prescribed steps to ensure training is consistent and of the highest quality. The PSC staff develops its training materials to facilitate learning through visual, auditory and kinesthetic channels.

Prior to scorer training, the PSC subject leaders conduct scoring supervisor training. A primary goal of this session is to ensure scoring supervisors clearly understand the scoring protocols and the training materials. This ensures all responses are scored in a manner consistent with the scores assigned to the anchor papers and according to the intentions of MDE. Scoring supervisors read and discuss the assessment items along with the rubrics used to score them. They are asked to carefully read and annotate all training materials so they can readily assist in scorer training and respond to scorers' questions during training and scoring.

The training agenda includes an introduction to the assessment program whose tests are being scored. It is important for scorers to have an understanding of the history and goals of the assessments and the context within which students' responses are evaluated. This gives them a better understanding of what types of responses can be expected. The scorers receive a description of the scoring criteria applied to the responses. Next, the trainers present the first item to be scored and the scoring rubric itself.

The primary goal of training is to convey to the scorers the decisions made during training paper selection about what type(s) of responses correspond to each score point and to help scorers internalize the scoring protocol so they may effectively apply those decisions. Scorers are better able to comprehend the scoring guidelines in context, so the rubric is presented in conjunction with the anchor papers. Anchor papers are the primary points of reference for scorers as they internalize the rubric. There are three to four anchor papers per item for each score point value. Trainers direct scorers' attention to the score point description from the scoring guide, as well as the illustrative anchor papers, thereby enabling scorers to immediately connect the language of the rubric with actual student performance.

After presentation and discussion of the anchor papers, each scorer is shown a practice set. Practice papers represent each score point and are used during training to help scorers become familiar with applying the rubric. Some papers clearly represent the score point. Others are selected because they represent borderline responses. Use of these practice 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.

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. Those reports include:

- *Daily and Cumulative Inter-Rater Reliability Reports by Item and Scorer*. These reports provide information about how many times scorers were in exact agreement, assigned adjacent scores or required resolutions. The reliability is computed and is monitored daily and cumulatively for the project.
- *Daily and Cumulative Frequency Distributions*. These reports show how many times each score point has been assigned to the item being scored by readers. The frequency distributions are produced both on a daily basis and cumulatively for the entire scoring project. This report allows scoring supervisors and subject leaders to see whether scorers have a tendency to score consistently high or low.

The most immediate method of monitoring a scorer's performance is through backreading by scoring supervisors. If a scoring supervisor discovers that a scorer is consistently assigning scores other than those the scoring supervisor would assign, he or she retrains that scorer, using the original anchor papers and training materials. This immediate check and remedial correction also provide an effective guard against scorer drift.

With the help of the individual scorer reliability and validity reports, the scoring lead staff can closely monitor each scorer's performance. In order to document retraining efforts for scorers with low reliabilities, the PSC maintains a Scorer Intervention Log. Entries on this form describe the feedback given a scorer regarding his or her problematic scoring and enumerate the interventions taken.

Readers are dismissed when, in the opinion of the subject leaders, those readers have been counseled, retrained, given every reasonable opportunity to improve and are still performing below the acceptable standard.

Appeals

The Performance Scoring Center (PSC) responds to appeals within five working days of notification. Once an appeal has been identified, the appropriate scoring director 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 non-disclosure 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

The Elimination of Constructed-response Items from the MCA-IIs

Following the 2009 Minnesota Legislative Session, House File 2 was signed into law. Among the numerous education policy and funding provisions in this bill was a qualification in how state funds can be used to support the assessment program. House File 2 prohibited the use of state funds in hand-scoring constructed-response items (CRs) on the Minnesota Comprehensive Assessments-Series II (MCA-II) in reading, science, and mathematics, with the exception of mathematics grades 3 to 8 of the 2009–2010 school year. It also required that any savings from this prohibition be redirected into the development of computerized statewide testing.

To progress in the direction required by House File 2, the Minnesota Department of Education (MDE) developed a plan to phase out human scoring of constructed-response items within the MCA-II. Human scoring of constructed-response items in this transition phase are scored using federal assessment funding. Appropriate state and federal budget allocations allow MDE to phase out human scoring of constructed-response items with a considered approach and still move forward with the development of computerized tests. The plan was prepared with comments by the Technical Advisory Committee (TAC), the Assessment Advisory Committee (AAC), and the Local Assessment and Accountability Advisory Committee (LAAAC), as well as consultation with the United States Department of Education (USDE) regarding the federal approval status of Minnesota's assessment system.

Reading

- Constructed-response items were eliminated beginning in the 2009–2010 school year. Technology-enhanced items will be considered for the Minnesota Comprehensive Assessments-Series III (MCA-III), scheduled to begin in 2012–2013.

Mathematics

- Constructed-response items continued through the completion of the MCA-II (2009–2010). Mathematics was computer delivered in grades 3–8 in 2010–2011 with the MCA-IIIs. In 2010-2011, the grade 11 MCA-II no longer had constructed-response items. Administered via paper, it will still be aligned to the previous set of content standards (2003) as students complete their high school careers, as required by Minnesota statute.

Science

- Science will continue with constructed-response items through 2010–2011, until the end of the current test series (MCA-II).

Transition Schedule

	2009–2010	2010–2011	2011–2012	2012–2013
Reading	No CRs, MCA-II	No CRs, MCA-II	No CRs, MCA-II	MCA-III
Math	Continue CRs in 3–8 & 11, MCA-II	MCA-III (3–8) & MCA-II (11)	MCA-III (3–8) & MCA-II (11)	MCA-III (3–8) & MCA-II (11)
Science	Continue CRs, MCA-II	Continue CRs, MCA-II	MCA-III	MCA-III

Rationale for Transition Plan

- Reading – From initial analysis, this subject is projected to be least susceptible to standards distribution/alignment re-work due to loss of constructed-response items. Based on this, it may be possible to forego resetting achievement standards on the test next year. This revision likely will require resubmission to the USDE peer review process.
- Mathematics – This subject was allowed to continue constructed-response items in grades 3-8 in 2009-2010. Continuing the use of constructed-response items in grade 11 in 2009-2010 would allow for additional planning in grade 11 for one year: grade 11 is the grade identified in analysis as the most susceptible to standards distribution/alignment. This would require only a single resubmission to USDE peer review in 2010-2011 for all of math rather than only grade 11 in the 2009-2010 year.
- Science – This subject would retain constructed-response items on the test for the remaining two years of Series II since analysis revealed the loss of these items is most susceptible to standards distribution/alignment issues and the item pool is not sufficiently deep to immediately support the replacement of constructed-response items as the test is only two years old. This would not require resubmission of peer review any earlier than anticipated in the 2011-2012 year.
- This plan provides a schedule that does not use any state dollars to score constructed-response items. It allows the state to move toward computer testing without compromising the quality of either endeavor.

- This plan establishes a logical transition that is easy for educators in the field to follow. Human-scored, constructed-response items are eliminated from the Minnesota Comprehensive Assessments in the following timetable by subject:

2009-2010	Reading
2010-2011	Mathematics
2011-2012	Science

Procedure in Revising the Test Specifications

Reading

Maximum item counts were modified for the Comprehension and Literature sub-strands to account for the loss of constructed response items. Item counts for the Vocabulary sub-strand were not adjusted because constructed response items were not permitted for this sub-strand. The following rules were followed in adjusting the item counts:

- For each constructed response point lost for a particular sub-strand, a point was added to the maximum multiple choice item count. This change allowed the total number of points for each grade to remain unchanged.
- The total number of items for each grade was modified to equal the total number of points on the test. This was necessary as multiple choice questions are only worth one point each.
- The maximum item count for each benchmark in the Comprehension and Literature sub-strands was increased by one to account for the extra items needed to replace the constructed response items. The exception was grade 6, where maximum item counts for each benchmark in the Literature sub-strand were increased by two. The greater increase in grade 6 was due to the need to replace eight points of constructed response items and only having three Literature sub-strand benchmarks in the grade.
- No other test specification was changed.

Mathematics (Grade 11)

The removal of 5 CR items coupled with the addition of 8 Gridded Response (GR) items and 2 Multiple Choice (MC) items resulted in a total of 70 items on the 2011 test (30 MCA items, 25 common items and 15 GRAD only items).

These assumptions were made in deciding upon a distribution of the additional items for the 2011 test:

- The count ranges in the MCS only (Strand; Sub-strand; Benchmark) SSBs could be changed since the CRs were removed from only these SSBs
- A GR should be placed in each allowable SSB since there are 10 GRs on the new test and there are equally only 10 benchmarks that allow for GRs.
- The count ranges in individual SSBs would not be changed in order to maintain a distribution of benchmarks close to the distribution of benchmarks in previous tests.

The table below from the current Test Specifications shows the recommended increases in the ranges of MCA only SSBs. The minimums and maximums of the ranges in strands III and IV were increased by 2 since the elimination of the CR items accompanied by the addition of the GR items to these strands would result in a net gain of 2. The range in strand V was also increased by 2 to ensure that strand V would not decrease in relative importance due to the loss of CRs.

Although the ranges for the individual SSBs were not changed because of the additional items, a change to the ranges for the following sub-strands is suggested.

Original			**2011**	
II.B	5 – 8 items		II.B	5 – 8 items
III.A	7 – 9 items		III.A	8 – 10 items
III.B	12 – 14 items		III.B	14 – 16 items
IV.A	8 – 10 items		IV.A	10 – 12 items
IV.B	6 – 8 items		IV.B	8 – 10 items
V.A	1 – 2 items		V.A	1 – 2 items
V.B	11 – 13 items		V.B	13 – 15 items
V.C	1 – 2 items		V.C	2 – 3 items

In most circumstances the increased sub-strand ranges reflect the number of benchmarks that could contain GR items, since a GR item was included in each available benchmark. In some cases, these sub-strand increases were also made to allow both MC and GR items in the same SSB.

Test Construction Procedures to Build Aligned Assessments Based on New Specifications

Reading

In June of 2009, Minnesota's Technical Advisory Committee discussed the issue of eliminated constructed-response items at its regularly-scheduled meeting. The TAC recommended an independent alignment review of the scheduled operational forms for the 2009-2010 year. MDE adjusted test construction activities to June of 2009 rather than fall of the year in order to facilitate this alignment review and take additional steps should the review warrant. MDE did not use items that were field tested in the spring of 2009 due to time constraints. This created some limitations on the quantity of items and passages available to build the 2010 operational forms.

Mathematics (Grade 11)

In June of 2009, Minnesota's Technical Advisory Committee discussed the issue of eliminated constructed-response items at its regularly-scheduled meeting. The TAC recommended an independent alignment review of the scheduled operational forms for the 2010-2011 year. MDE adjusted test construction activities to June of 2009 rather than fall of the 2010 year in order to facilitate this alignment review and take additional steps should the review warrant. MDE did not use items that were field tested in the spring of 2009 due to time constraints. This created some limitations on the quantity of items and passages available to build the 2011 operational form for grade 11.

Independent Alignment Review

Reading and Mathematics (Grade 11)

In the summer of 2009, the MCA-IIs for reading were constructed for all grades, 3-8 and 10 – using only multiple-choice items. Also, the MCA-II for mathematics 2011 was constructed using only multiple-choice and gridded items. Upon the completion of those operational forms, MDE hired the services of Dr. Norman Webb to conduct an independent alignment review of all grade-level assessments to determine if the assessments attend to the following:

- Cover the full range of content specified in the State's academic content standards, meaning that all of the standards are represented legitimately in the assessments;
- Measure both the content (what students know) and the process (what students can do) aspects of the academic content standards;
- Reflect the same degree and pattern of emphasis apparent in the academic content standards (e.g., if the academic standards place a lot of emphasis on operations then so should the assessments);
- Reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the State's academic content standards, meaning that the assessments are as demanding as the standards.

Dr. Webb and his associated conducted their alignment review in Roseville, MN, from July 28-31, 2009. On July 31, Dr. Webb provided the MDE with a preliminary report of their findings. An overview of the findings provided on July 31, 2009, is provided below.

Grade 3

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
3.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	WEAK
3.I.C - Comprehension. The student will understand the mea ...	YES	NO	YES	YES
3.I.D - Literature. The student will actively engage in th ...	YES	YES	YES	YES

Grade 4

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
4.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	NO	YES
4.I.C - Comprehension. The student will understand the mea ...	YES	YES	YES	YES
4.I.D - Literature. The student will actively engage in th ...	YES	YES	YES	WEAK

Grade 5

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
5.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	YES
5.I.C - Comprehension. The student will understand the mea ...	YES	WEAK	YES	YES
5.I.D - Literature. The student will actively engage in th ...	YES	NO	YES	YES

Grade 6

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
6.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	YES
6.I.C - Comprehension. The student will understand the mea ...	YES	WEAK	YES	WEAK
6.I.D - Literature. The student will actively engage in th ...	YES	NO	YES	YES

Grade 7

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
7.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	YES
7.I.C - Comprehension. The student will understand the mea ...	YES	WEAK	YES	YES
7.I.D - Literature. The student will actively engage in th ...	YES	NO	YES	YES

Grade 8

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
8.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	WEAK
8.I.C - Comprehension. The student will understand the mea ...	YES	YES	YES	YES
8.I.D - Literature. The student will actively engage in th ...	YES	NO	YES	YES

Grade 10

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
10.I.B - Vocabulary Expansion. The student will use a varie ...	YES	YES	YES	YES
10.I.C - Comprehension. The student will understand the mea ...	YES	YES	YES	YES
10.I.D - Literature. The student will actively engage in th ...	YES	WEAK	YES	YES

Mathematics (Grade 11 in 2011)

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
11.I - MATHEMATICAL REASONING	YES	YES	NO	YES
11.II - NUMBER SENSE, COMPUTATION AND OPERATIONS	NO	YES	NO	YES
11.III. - PATTERNS, FUNCTIONS AND ALGEBRA	YES	YES	YES	YES
11.IV - DATA ANALYSIS, STATISTICS AND PROBABILITY	YES	YES	YES	YES
11.V. - SPATIAL SENSE, GEOMETRY AND MEASUREMENT	YES	YES	YES	YES

Revision of Operational Test Forms to Strengthen Alignment

Reading

Upon review of the preliminary alignment results and after consultation with Dr. Webb, MDE reviewed the operational forms in an attempt to strengthen their alignment in critical areas. Specifically, MDE conducted the following procedure to investigate its ability to strengthen the operational forms.

1. Identified areas of critical focus from the preliminary results. The areas of focus are those Depth of Knowledge (DOK) cells were the analysis indicated NO in alignment. This resulted in the following grades and sub-strands:

Grade 3	C: Comprehension
Grade 5	D: Literature
Grade 6	D: Literature
Grade 7	D: Literature
Grade 8:	D: Literature

2. Reviewed the item pool of available, appropriate cognitive complexity that existed within the passages currently assigned to the operational forms.
3. Examined those supplementary items in the pool for their ability to replace items that currently existed in the operational form without compromising the other test specification requirements (e.g., item counts, benchmark distributions) or best practices of test construction (e.g., cuing items or psychometric targets). A critical factor in identifying eligible replacement items was discrepancies that may have existed between an item's benchmark coding derived from the Minnesota Assessment Advisory Panels and the benchmark coding assigned during the alignment study. In order for an item to be eligible for replacement, MDE required the benchmark coding assigned by the state be identical to benchmark coding assigned by majority of reviewers during the alignment study.
4. Replaced operational items currently in the form with higher level, cognitively complex items assigned to the passage in order to strengthen the DOK alignment. This effort was hampered by not being able to use constructed-response items (frequently assigned to a Level C cognitive complexity) or the items produced through the 2009 spring field testing efforts. The results of this review were as follows within each sub-strand identified in Step 1 above:

Grade 3	1 Level A item replaced with 1 Level B item
Grade 5	No eligible replacement items in pool
Grade 6	No eligible replacement items in pool
Grade 7	No eligible replacement items in pool
Grade 8:	No eligible replacement items in pool

Mathematics (Grade 11)

Upon review of the preliminary alignment results and after consultation with Dr. Webb, MDE reviewed the operational forms in an attempt to strengthen their alignment in critical areas. Since the grade 11 mathematics form for 2011 did not show any weaknesses in Depth of Knowledge, MDE made no adjustments for alignment. It should be noted in the operational form – according to the design of the test specifications and the intent of the Minnesota Academic Standards – that Mathematical Reasoning (Strand I) and Number Sense (Strand II) are skills expected to be demonstrated within the remaining four strands and as such are embedded within them.

Statistical Comparisons Prior to Administration

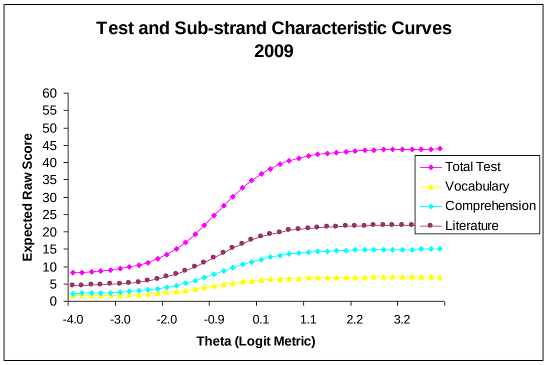
After the operational forms were finalized, based on the procedures described above, statistical methods were used to compare the predicted psychometric characteristics of of 2010 and 2011 Mathematics test forms as well as 2010 Reading forms against the obtained results from 2009. Three comparisons, based on IRT item parameter estimates, were used to evaluate the similarity of the psychometric characteristics of the revised and previous versions of the tests. For the first comparison, estimates of the conditional standard error of measurement at the three cut scores were calculated for both years, based on the test information function. For the other two comparisons, the estimated test and strand-level characteristic curves and information functions were plotted. The tables and graphs below present the results for each grade. In every case, the results show that the revised tests are predicted to be psychometrically comparable to their previous versions.

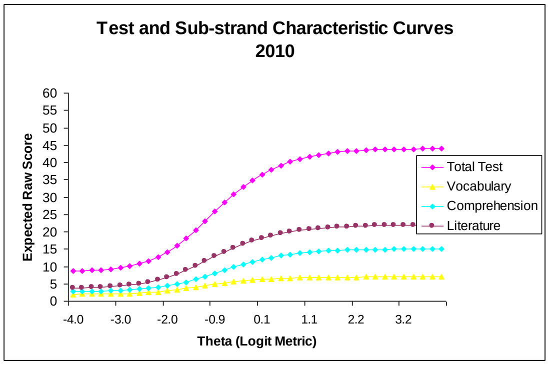
Reading Grade 03

(a) CSEM at cut score

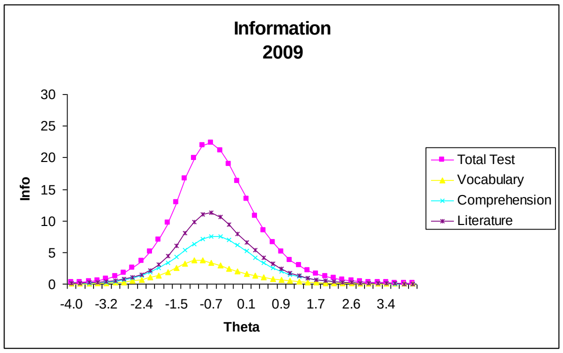
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.4013	0.25	0.23
Meets	-0.8449	0.21	0.22
Exceeds	-0.0065	0.22	0.28

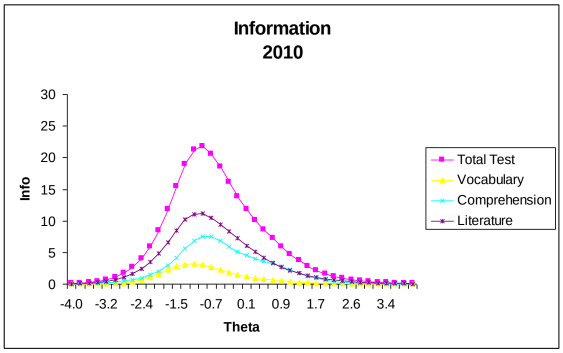
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



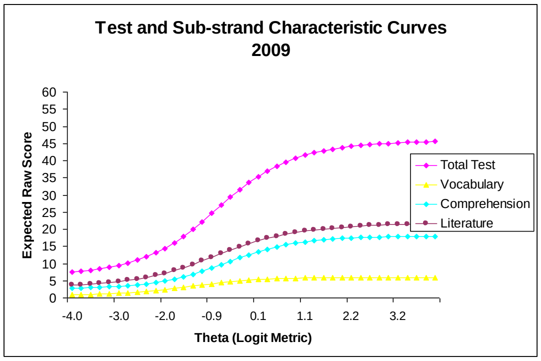


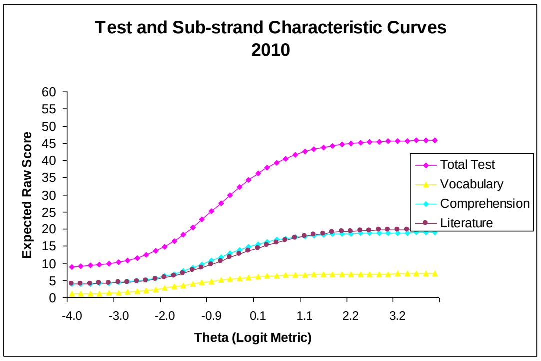
Reading Grade 04

(a) CSEM at cut score

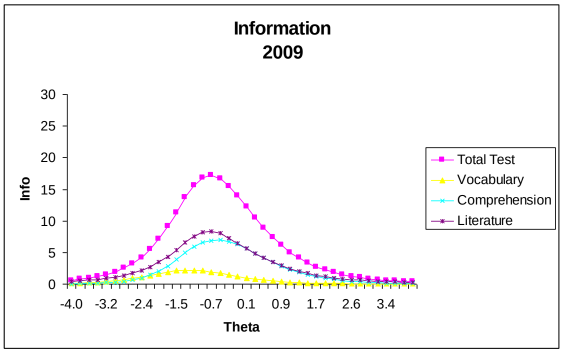
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.3632	0.27	0.28
Meets	-0.6527	0.24	0.25
Exceeds	0.3188	0.31	0.31

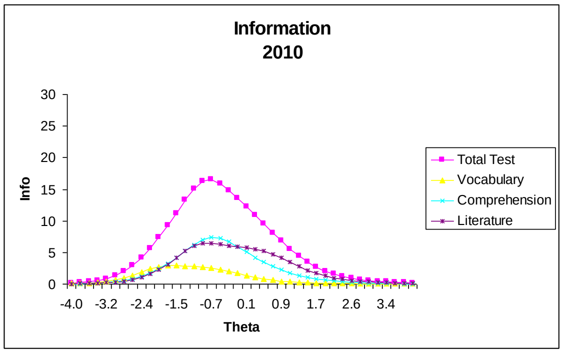
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



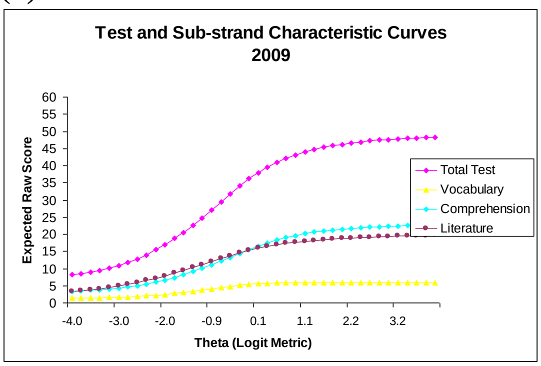


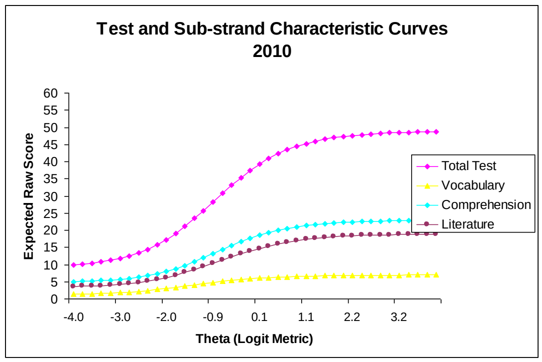
Reading Grade 05

(a) CSEM at cut score

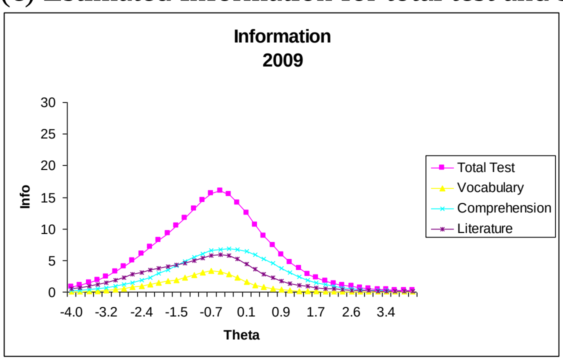
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.4522	0.31	0.28
Meets	-0.6738	0.25	0.25
Exceeds	0.5049	0.34	0.33

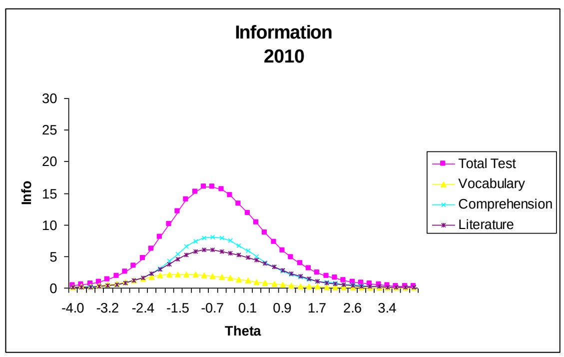
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



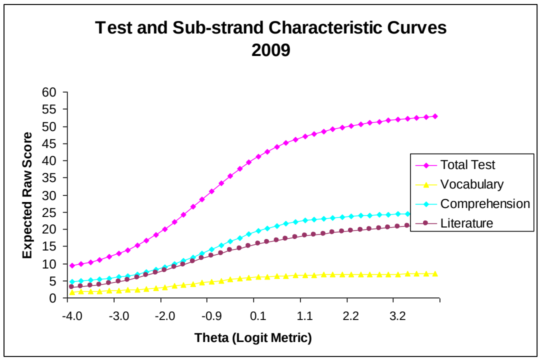


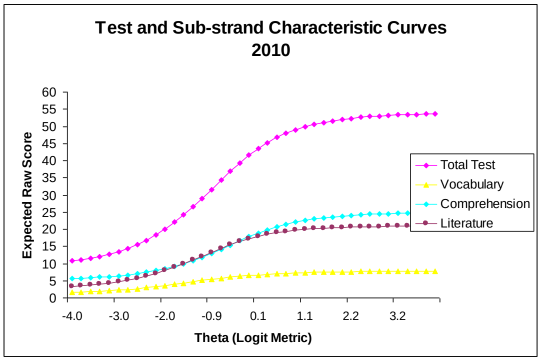
Reading Grade 06

(a) CSEM at cut score

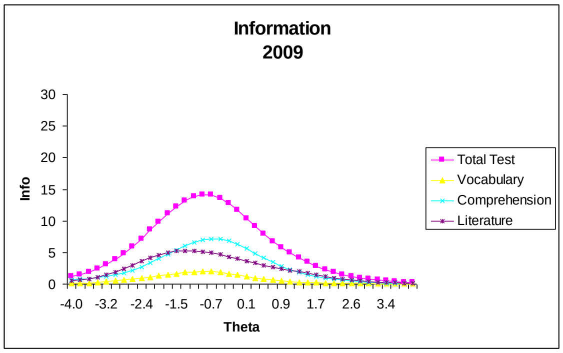
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.37	0.28	0.28
Meets	-0.5055	0.27	0.25
Exceeds	0.4848	0.36	0.32

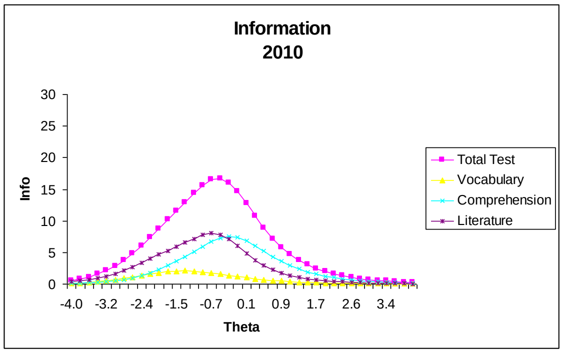
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



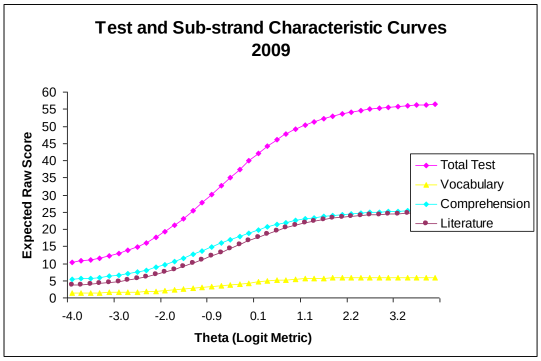


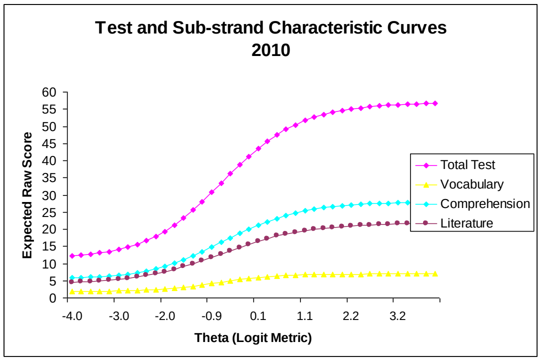
Reading Grade 07

(a) CSEM at cut score

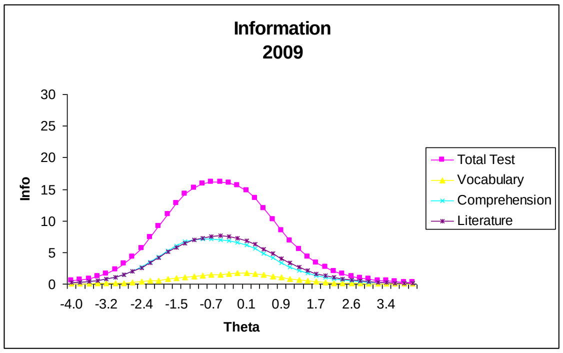
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0964	0.26	0.26
Meets	-0.3308	0.25	0.24
Exceeds	0.4868	0.29	0.28

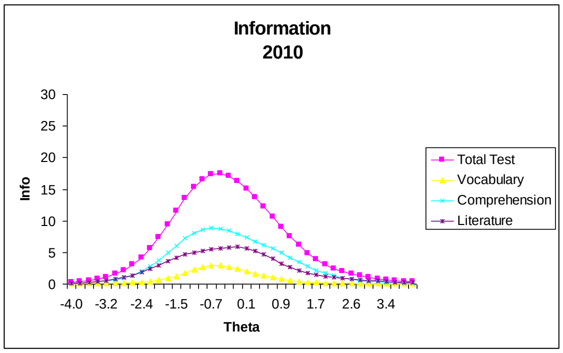
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



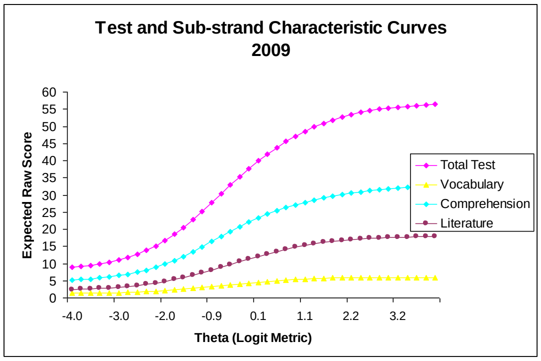


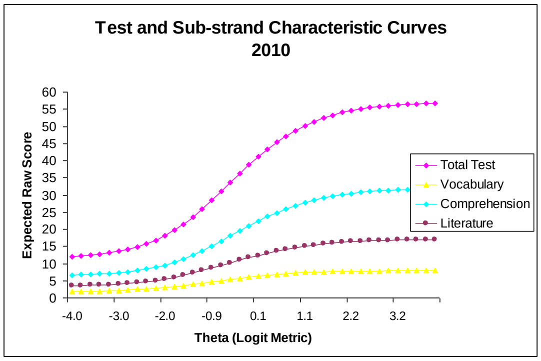
Reading Grade 08

(a) CSEM at cut score

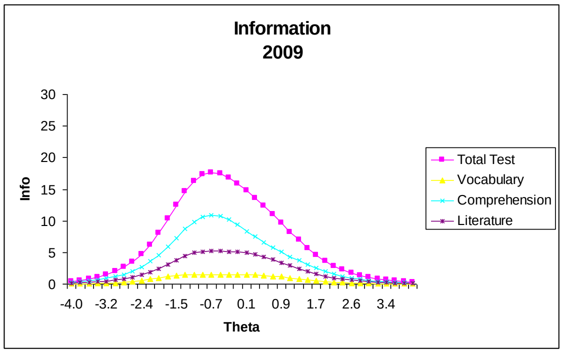
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0619	0.25	0.27
Meets	-0.2563	0.25	0.26
Exceeds	0.5755	0.29	0.30

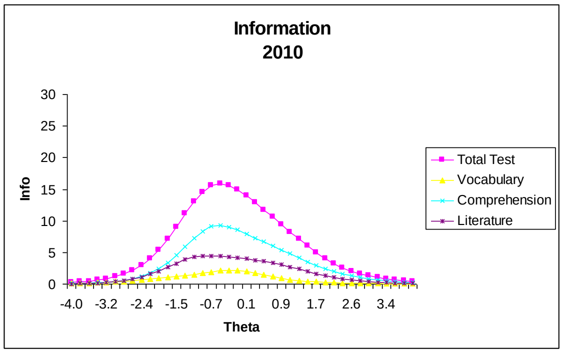
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



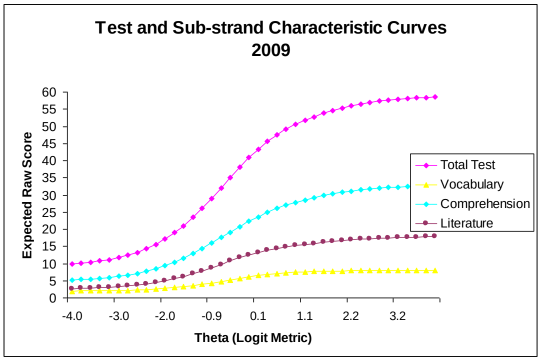


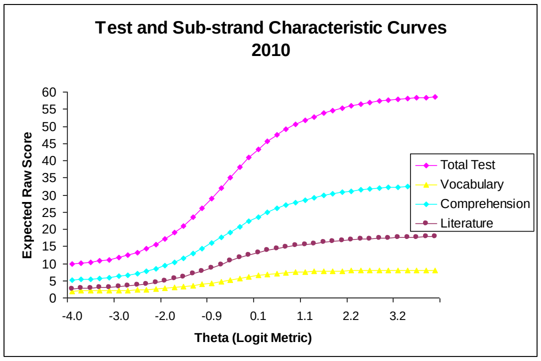
Reading Grade 10

(a) CSEM at cut score

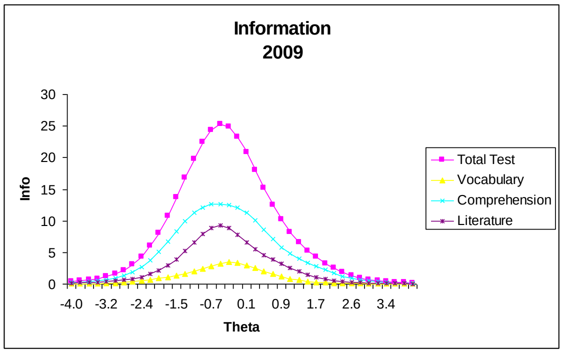
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0438	0.23	0.25
Meets	-0.2914	0.23	0.21
Exceeds	0.585	0.30	0.27

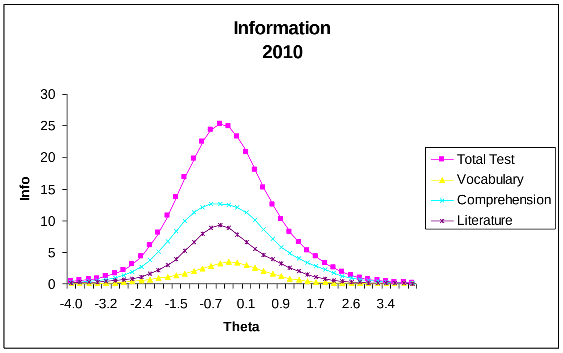
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



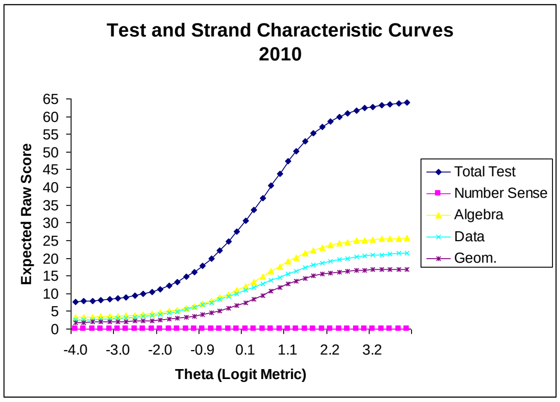


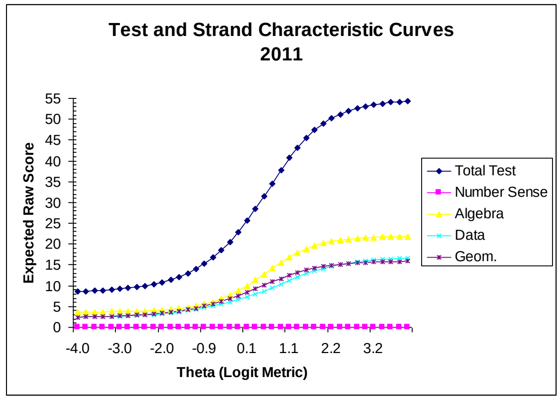
Math Grade 11

(a) CSEM at cut score

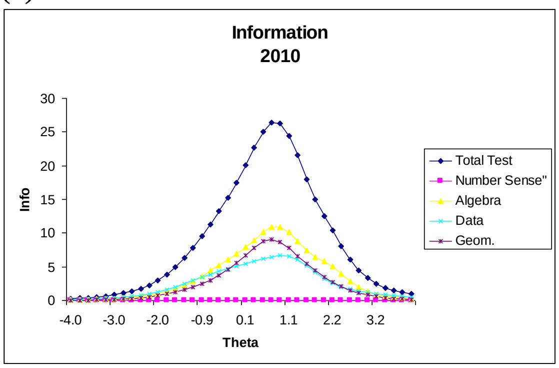
	Theta Cut	2010 CSEM	2011 CSEM
Partially Meets	0.0012	0.27	0.27
Meets	0.5467	0.26	0.22
Exceeds	1.3011	0.27	0.24

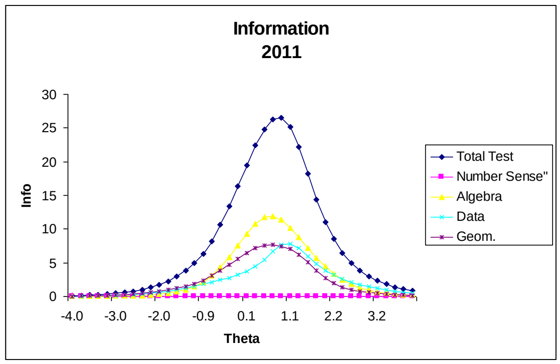
(b) TCC for total test and strands





(c) Information functions for total test and strands





TAC Review

MDE reviewed the results of these activities with its TAC in November of 2009 for the reading and mathematics (grade 11) assessments. The Minnesota TAC found the procedures to date acceptable and technically defensible. Although the TAC made some recommendations for the display of the test and strand characteristic curves, the members made no substantive recommendations or revisions requests to the work done leading up to the spring 2010 administration of Reading MCA-II and the spring 2011 administration of grade 11 Mathematics MCA-II..

Statistical Comparisons After Administration

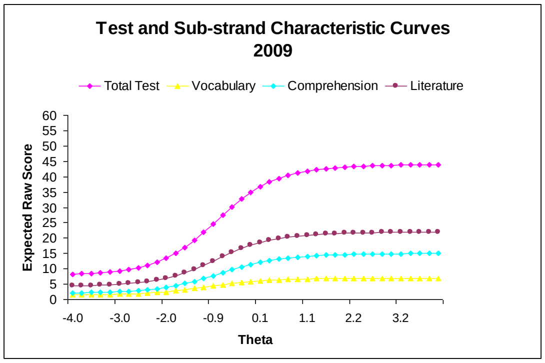
Statistical methods were used to compare the obtained psychometric characteristics of 2010 and 2009 Reading forms and the 2010 and 2011 Mathematics forms. Three comparisons, based on IRT item parameter estimates, were used to evaluate the similarity of the psychometric characteristics of the revised and previous forms. For the first comparison, estimates of the conditional standard error of measurement at the three cut scores were calculated for both years. The MCA-II tests use test characteristic curve (TCC) scoring, with the raw score cut set at the value whose associated theta value on the TCC is closest to the theta cut-score. Using the test characteristic method to estimate CSEM resulted in different estimates for the 2009 forms than values reported for the prior to administration results, but it was deemed to be the more appropriate method to use for the after administration results as it reflects the final reporting score method. For the other two comparisons, the estimated test and strand-level characteristic curves and information functions were plotted. The tables and graphs below present the results for each subject and grade. In each case, the results indicate that the revised tests are psychometrically comparable to their previous versions.

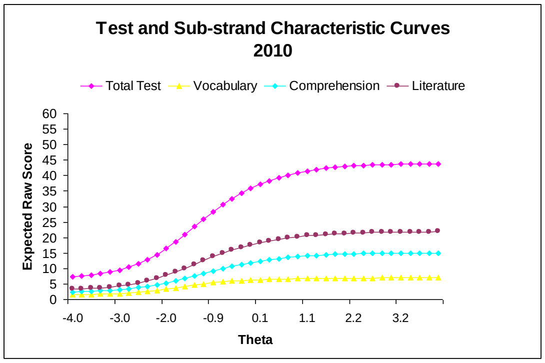
Reading Grade 03

(a) CSEM at cut score

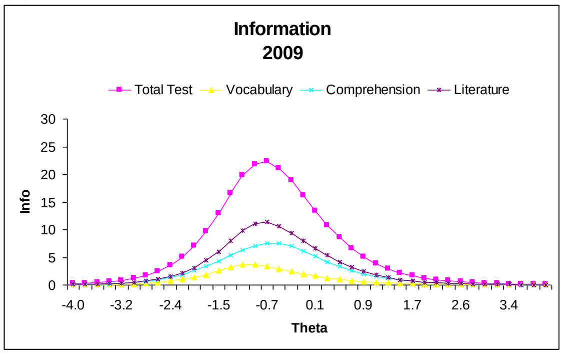
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.4013	0.29	0.26
Meets	-0.8449	0.27	0.27
Exceeds	-0.0065	0.39	0.36

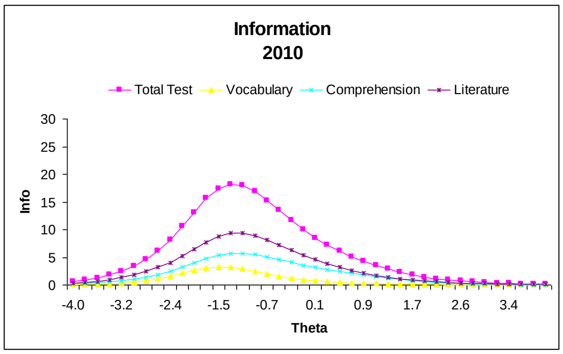
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



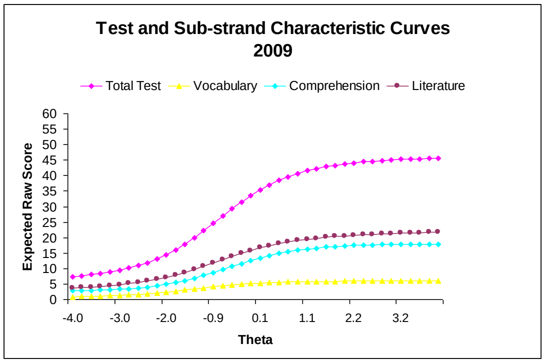


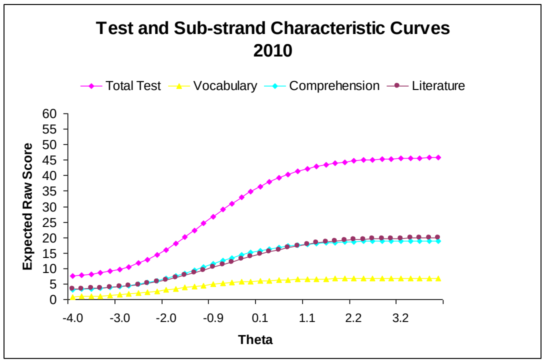
Reading Grade 04

(a) CSEM at cut score

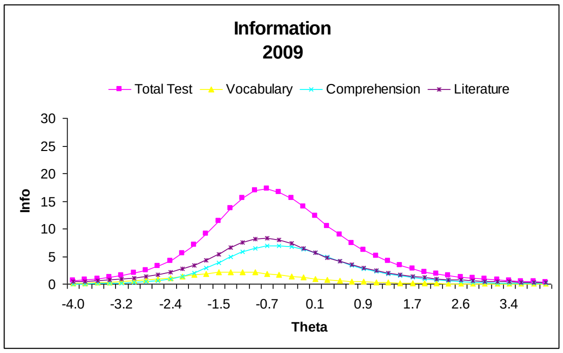
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.3632	0.31	0.31
Meets	-0.6527	0.29	0.30
Exceeds	0.3188	0.41	0.38

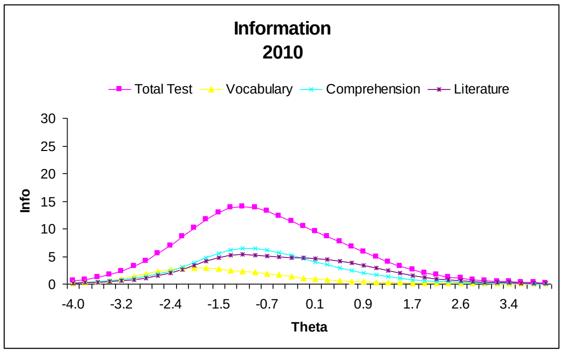
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



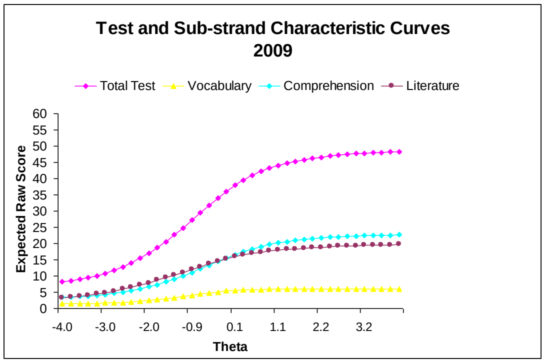


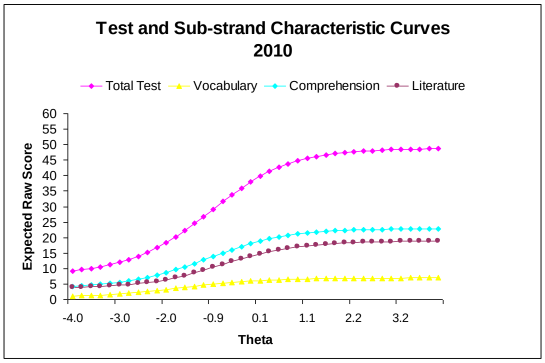
Reading Grade 05

(a) CSEM at cut score

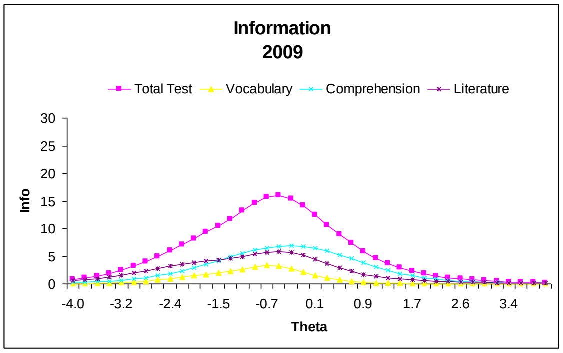
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.4522	0.33	0.32
Meets	-0.6738	0.28	0.28
Exceeds	0.5049	0.50	0.51

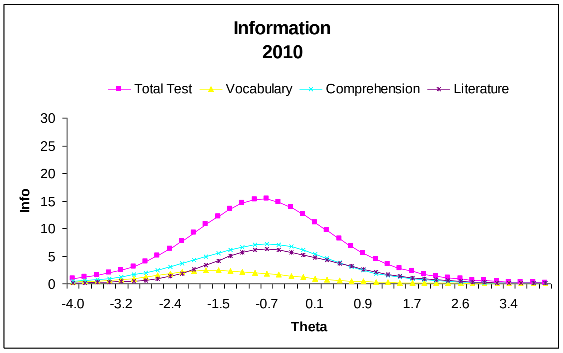
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



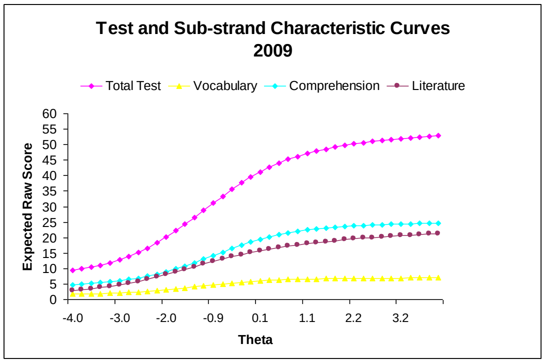


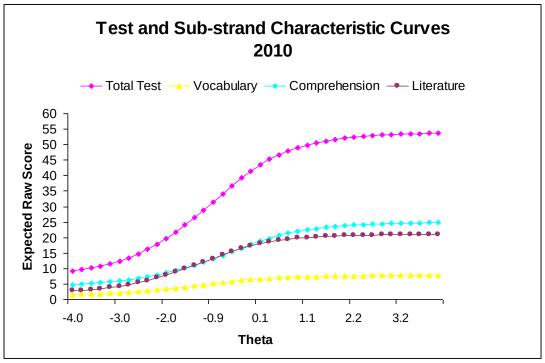
Reading Grade 06

(a) CSEM at cut score

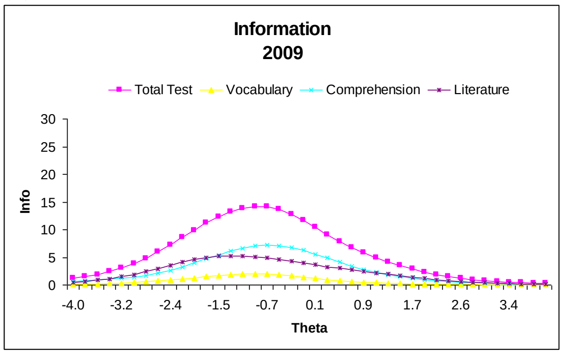
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.37	0.33	0.30
Meets	-0.5055	0.34	0.27
Exceeds	0.4848	0.47	0.38

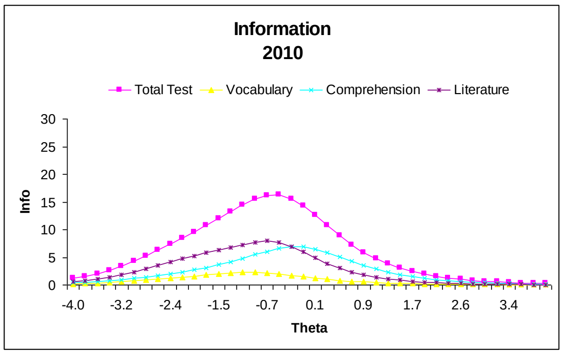
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



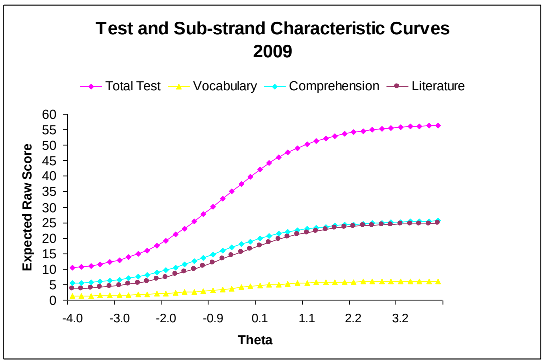


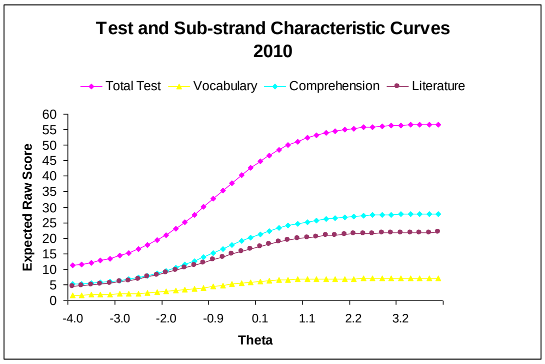
Reading Grade 07

(a) CSEM at cut score

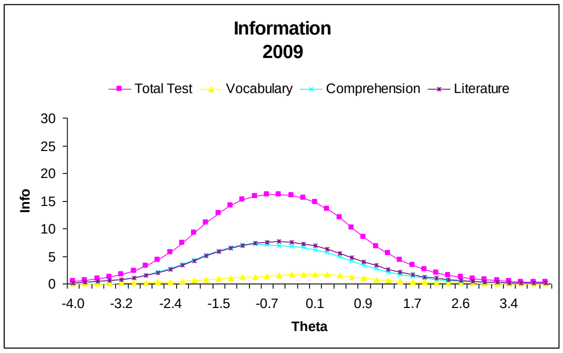
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0964	0.33	0.29
Meets	-0.3308	0.32	0.27
Exceeds	0.4868	0.35	0.34

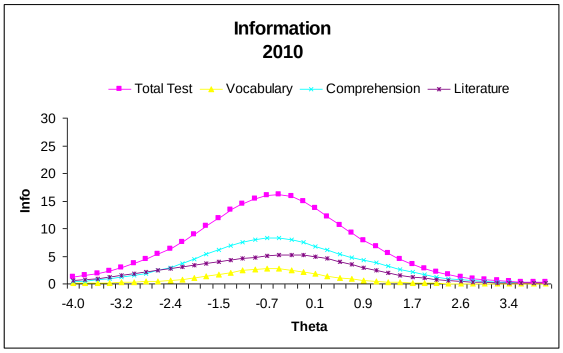
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



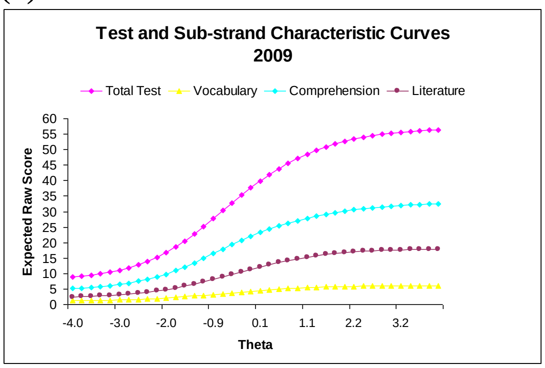


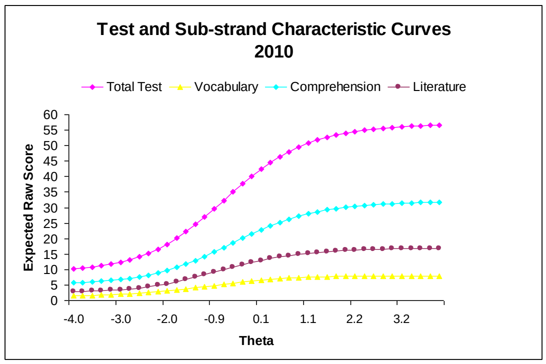
Reading Grade 08

(a) CSEM at cut score

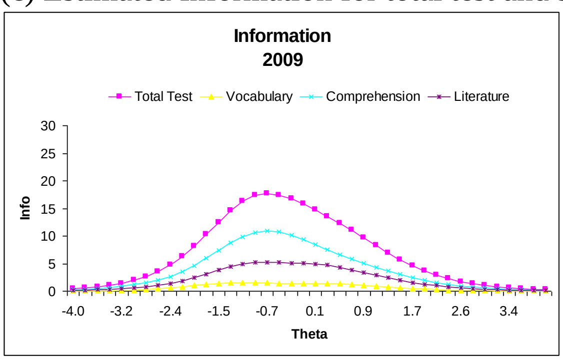
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0619	0.30	0.29
Meets	-0.2563	0.31	0.28
Exceeds	0.5755	0.37	0.35

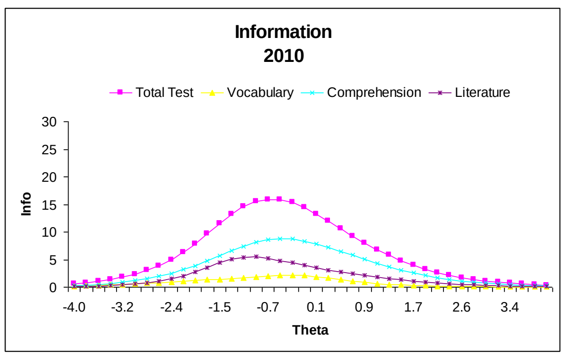
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



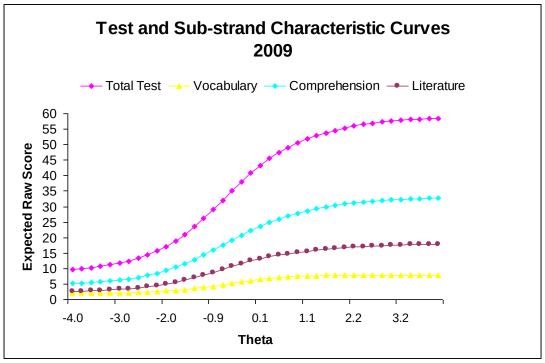


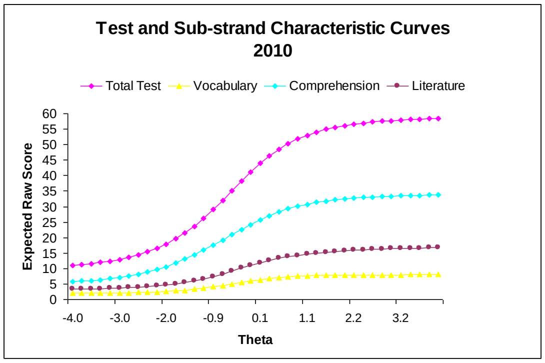
Reading Grade 10

(a) CSEM at cut score

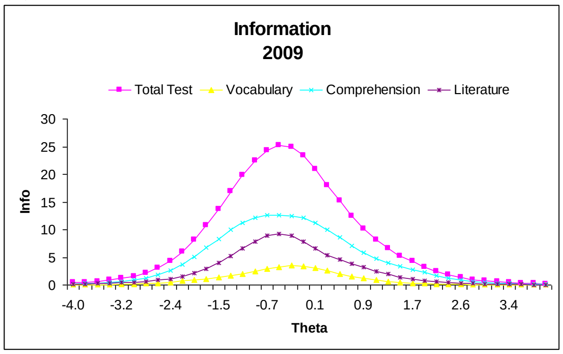
	Theta Cut	2009 CSEM	2010 CSEM
Partially Meets	-1.0438	0.28	0.29
Meets	-0.2914	0.27	0.23
Exceeds	0.585	0.36	0.32

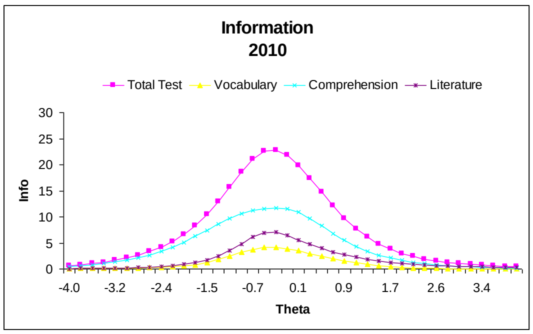
(b) Estimated TCC for total test and strands





(c) Estimated Information for total test and strands



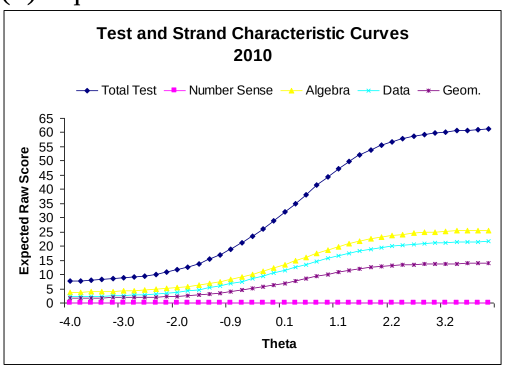


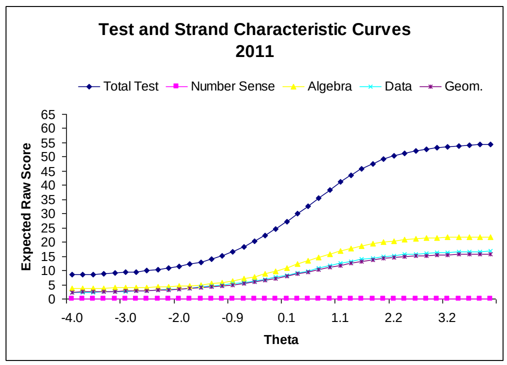
Math Grade 11

(a) CSEM at cut score

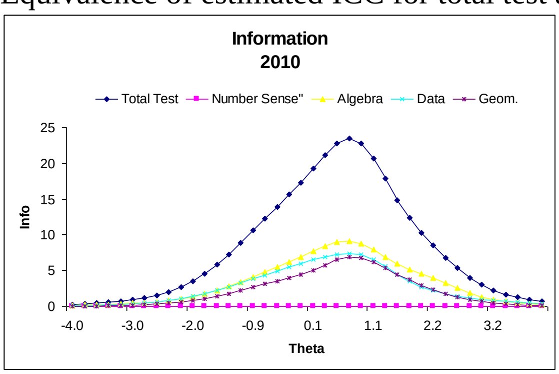
	Theta	2010 CSEM	2011 CSEM
Partially Meets	0.0012	0.24	0.25
Meets	0.5467	0.22	0.22
Exceeds	1.3011	0.22	0.21

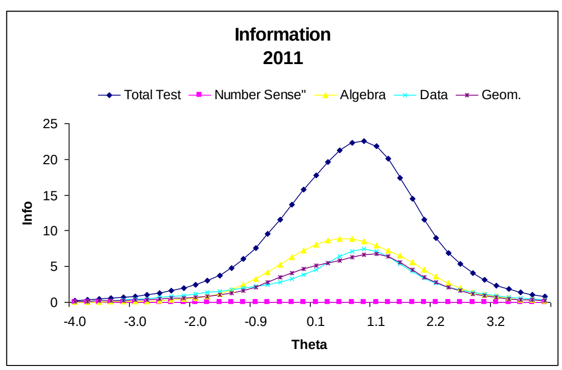
(b) Equivalence of estimated TCC for total test and strands





Equivalence of estimated ICC for total test and strands





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 of 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, the Minnesota's testing contractor's commitment to quality is evidenced by initiatives in two major areas:

- Task-specific quality standards integrated into individual processing functions and services
- 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 et. al., 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 Non-Scannable Documents

Minnesota's testing contractor contracts with outside vendors for the printing of non-scannable documents because of the large volume of printed materials necessary for the Minnesota Assessment system. To ensure the accuracy of these documents, Minnesota's testing contractor holds periodic meetings with all of their printers to reiterate the high expectations for printing quality and to remind them of the penalties associated with the failure to perform to standards. The following quality controls are implemented to facilitate the successful performance of outside printing companies.

- Minnesota's testing contractor provides design and schedule requirements to printers well in advance of the delivery of copy so that the schedule for printing can be arranged.
- If any changes are made by the Minnesota Department of Education (MDE) and Minnesota's testing contractor with regard to a print schedule, then the printer is notified immediately.

- Corrections submitted by MDE are added to any of the corrections Minnesota's testing contractor sends to the printer.
- All page proofs, final proofs and specimens of printed materials are proofread in their entirety by the forms support department and are submitted to MDE for review.
- Sample printed materials are examined for the required paper type, ink color, collation and copy. If discrepancies are noted, the printer is notified immediately to make allowances for corrections and reprint where required.
- Whenever possible, electronic transfer of copy is used to minimize human error and to expedite the printing process.

An additional quality check of all outside printing materials is made during the packaging operation at Minnesota's testing contractor. Each box of materials is spot checked to verify printing and collating accuracy.

Quality Control in Data Preparation

To ensure an accurate accounting of the hundreds of thousands of Minnesota assessment documents that Minnesota's testing contractor receives, year-round Data Preparation staff perform a series of receipt and check-in procedures. All incoming materials are carefully examined for a number of conditions, including damage, errors, omissions, accountability and secured documents. When needed, corrective action is promptly taken according to specifications developed jointly by Minnesota's testing contractor and MDE.

Quality Control in Production Control

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 Workflow Management System (WFM). 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. WFM efficiently carries the planning and control function to first-line supervisory personnel so that key decisions can be made properly and rapidly. Since WFM 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.

Quality Control in Editing and Data Input

As Minnesota assessments' answer documents are scanned, the data are electronically transcribed directly to data files, creating the project's database. After scanning, a three-step data editing process is performed to verify that all data on the project database are complete and accurate. During this process, the data are examined for omissions, inconsistencies, gridding errors and other specified error-suspect conditions.

The first quality control step consists of a complete computer editing of the data to verify all documents are accounted for and all possible "suspects" or omissions have been checked.

In the second editing step, Minnesota's testing contractor's editing personnel review the errors detected during the first step and indicate the necessary corrections to be made. The editing staff inspects both the computer-generated edit log and the actual source document that is listed on the edit log as being "suspect" or containing possible errors. The edit log indicates the actual field or information that may be in error. The editing staff visually checks this particular piece of information against the source document. At this point, double grids, erasures and smudge marks are flagged. Subsequently, one of the following actions is taken.

- Correctable error: If an error is correctable by the editing staff according to editing specifications, then the corrections are handwritten on the edit log, checked by a lead staff member and the required changes are made by the Data Input department. These editing specifications are custom-designed to meet Minnesota Department of Education (MDE) requirements.
- Error Not Correctable According to Specifications: If an error is not correctable according to the specifications, the Project Director will be notified and contact with MDE will be made. The correction information will be obtained from MDE for the item in question. The specifications for the types of error corrections requiring contact with MDE are developed jointly.

- Non-correctable error: If a “suspect” is found, but no alterations are possible according to the specifications, the proper procedure to allow this type of data to remain on the records is initiated, and no corrective action is necessary. An example of this would be an answer document containing double-gridded student demographic information.

Once the necessary corrections have been entered in the edit log and checked by a lead staff member, the batch is forwarded to the Data Input department, where corrections are key-entered and key-verified on data entry terminals. At this point, the updated batch files will contain only valid information. The data entry screens are designed to enhance operator speed and accuracy: fields to be entered are titled to reflect the actual source document. When all corrections for a batch have been entered and verified, then the correction file is submitted to the mainframe computer for updating of the batch data file.

A third edit step, called post-edit, takes place as the data file is being updated. During this step, the entire data file is again re-edited according to an editing procedure approved by MDE.

Quality Control in Performance Assessment Services and the Performance Scoring Center—Constructed-Response Tasks

Quality control permeates all steps of the performance assessment scoring process. It starts with a scorer recruiting and screening process designed to locate and employ the most highly qualified individuals available. Scorers receive careful, exacting and thorough training in the specific items and rubrics at the beginning of each scoring project, regardless of their previous scoring experience. Training is provided by those individuals on our staff who, after fulfilling rigorous internal guidelines for presentation skills and knowledge, have become qualified trainers. During scoring, scorers are constantly monitored to ensure they are scoring accurately and consistently. More complete details regarding Performance Scoring Center (PSC) quality control procedures for constructed-response scoring are presented in chapter 10.

Quality Control for Computer-Based Assessments

Minnesota’s testing contractor uses a proven data verification plan to fully test all functions, outputs, processes and procedures of the Assessment Network computer-based testing system before tests are administered to students. As detailed below, Minnesota’s testing contractor meets the critical need for data verification by means of time-tested, comprehensive quality control techniques and processes. Minnesota’s testing contractor Pearson’s Software Quality & Testing (PSQT)—an internal, dedicated testing group—monitors and performs all aspects of the data verification process. All PSQT data verification processes ultimately have one goal: to ensure that the system functions as expected so that all scoring and reporting data are accurate.

The “master document” for Minnesota’s testing contractor’s data verification process is the Validation and Verification (V&V) test plan. V&V activities include, but are not limited to, documentation reviews, code inspections, structural walkthroughs and testing, unit testing, integration testing, systems testing, and acceptance testing. The sample V&V plan itself details a front-to-back testing process, with each phase defined by activity, responsibility and methods and tools.

TABLE 11.1. Validation and Verification Test Plan

Phase	V&V Activity	Groups Responsible	Methods and Tools
Requirements	Testability analysis	Development, Product Line, PSQT	Inspections, Review
	Completeness review	Development, Product Line, PSQT	Inspections, Review
Design	Testability analysis	PSQT, Development, Product Line	Inspections, Review
	Completeness review	PSQT, Development, Product Line	Inspections, Review
Code/Unit Test	Verification reviews (code)	PSQT, Development	Inspections
	Unit test	Development	Various
System Test	Subsystem (integration)	PSQT, Requirements Analysts	V&V Plan, Manual testing, Automated Testing
	System Test	PSQT, Development	Review
	User Acceptance Test	Product Line	Review

Minnesota's testing contractor implements the V&V test plan and receives MDE approval at least one month prior to each test administration. The V&V test plan includes test cases for all aspects of web-based testing, including student registration, login procedures, computer-based training modules (online tutorials, Practice Tests, etc.), image capturing (TestNav delivery system) and data transfer.

For example, test cases reflecting all combinations of demographic fields (for example, gender, disability, ethnic codes and special population classifications) are written and tested. Cases written for mock data reflect both valid and invalid demographic data combinations and correct and incorrect responses to test items.

The verification of mock data process also includes taking the test online to gather electronic expected results data. Any unexpected results will be routed back to Minnesota's testing contractor's software development staff for investigation and examination of test case viability. Expected results testing is a highly systematic, controlled approach to testing, designed to systematically identify anomalies in the data. One major benefit to this kind of testing is that the PSQT test group creates cases based on its interpretation of the project requirements and does so independently of the development group responsible for producing the score data. This methodology affords the checks and balances necessary to first produce a common understanding of the project requirements and then yield accurate and valid test results based on those requirements.

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.

- Based on the equating sample (approximately 80 percent of all data), Minnesota's testing contractor performs a "key-check" analysis to ensure the appropriate scoring key is being used.
- 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.
- The post-equated and pre-equated values of anchor items are compared and differences beyond expectation are investigated and resolved.
- New post-equated raw score to scaled score tables are generated and compared to the pre-equated tables. Any unexpected differences are resolved.
- Expected passing rates or rates of classification are generated and compared to previous years.
- An equating summary is provided to MDE and the National Technical Advisory Committee (TAC) for review.

Quality Control in Scoring and Reporting

All Minnesota assessment reports are quality-controlled by Minnesota's testing contractor staff. Before reporting, conversion programs with mock data are run to ensure that accurate reports are being produced. Calculations are also verified to ensure they are being performed according to the specifications of MDE. In addition, a random sample of reports are selected during processing and checked against raw data to verify the accuracy of the actual reports. Test files are used to produce reports for the software quality-assurance team to review. The reports generated from the test files are checked against SAS checker programs as well as file compares. These mockups are sent to MDE for their approval. This approval is specifically related to the format and look of the report. Once these mockups are approved, the data is checked again using production data. Data files are provided to MDE prior to the districts receiving their reports. This data is used by MDE to confirm the reported data is correct as well as prepare reports for a state press conference regarding the release of results.

Score reports and analyses are assembled by Minnesota's testing contractor's pre-mailing staff. Strict quality control is observed during pre-mailing to ensure all score reports and analyses shipments are complete. Once all score reports are assembled and quality-checked, they are distributed using quality shipping procedures.

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

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.

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 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)

An individual's 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 2013–14.

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 to demonstrate proficiency. On the 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-Retest Reliability Estimate

A statistic that represents the correlation between scores obtained from one measure when compared to scores obtained from the same measure on another occasion.

Test Specifications

A detailed description of a test that helps to 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.

Test-Centered Standard Setting Methods

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.

True Score

That 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, 1999) 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 S*tandards*.

References

Albus, D., Klein, J. A., Liu, K., & Thurlow, M. (2004). *Connecting English language proficiency, statewide assessments, and classroom proficiency* (LEP Projects Report 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved September 19, 2008, from the website: http://education.umn.edu/NCEO/OnlinePubs/LEP5.html

American Educational Research Association, American Psychological Association, & the 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, & the National Council on Measurement in Education. Joint Technical Committee. (1999). *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, p. 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 non handicapped groups. *The Journal of Special Education*, 21(3), 9–21.

Browder, D.M., Gibbs, S., Ahlgrim-Delzell, L., Courtade, G., Mraz, M., & Flowers, C. (in press). Literacy for students with significant cognitive disabilities: What should we teach and what should we hope to achieve? *Remedial and Special Education*.

Chien, M., Hsu, Y, & Shin, D. (2007). ISE [computer software]. Iowa City, IA: Pearson.

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.

Dudek, F.J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*, 335-337.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., & Swaminithan, 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, Huynh, Meyer, J. Patrick, & Gallant, Dorinda 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.

Jaeger, R. M., (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement*, (3rd Edition, 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.

Kolen, M. J. (2004) POLYEQUATE [Computer Software]. Iowa City, Iowa: 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 edition. 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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.

MacArthur, Charles A., & Cavalier, Albert 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, Charles A., & Cavalier, Albert R. (2004). Dictation and Speech Recognition Technology as Test Accommodations. *Exceptional Children*, 71(1), 43–58.

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). Available online 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), pp. 333–342.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210.

Thissen, D. (1991). *MULTILOG user's guide*. Chicago: Scientific Software.

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.

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.

Tong, Y., Um, K., Turhan, A., Parker, R., Shin, D., Chien, Y., & Hsu, Y. (2007). *IRT Score Estimation Evaluation Document* (Pearson Educational Measurement Internal Report). Iowa City, IA.

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.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Verhelst, N.D., & Verstralen, H.H.F.M. (1997). *Modeling sums of binary responses by the partial credit model* (Measurement and Research Department Report No. 97-7). Cito, Arnhem, The Netherlands.

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 cut scores has evolved since the 1980s. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum.