DEPARTMENT OF EDUCATION

2023–24 Technical Manual for Minnesota's Statewide Assessments

Prepared by Pearson | December 2024

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Below is a list of abbreviations that appear in this technical manual.

2PL ..... two-parameter logistic model
3PL ..... three-parameter logistic model
AIR ..... American Institutes for Research
ALD ..... achievement level descriptor
AMPI ..... Alternate Model Performance Indicator
ARC ..... Assurance, Rationale, and Context
AYP..... Adequate Yearly Progress
BST ..... Basic Skills Test
CAL..... Center for Applied Linguistics
CAT..... computer adaptive test
CCR ..... career and college readiness
CR..... constructed-response
CRM ..... conditional randomesque method
CSEM..... conditional standard error of measurement
DIF..... differential item functioning
DOK..... depth of knowledge
DRC ..... Data Recognition Corporation
DSR ..... District Student Results
EAP..... expected a priori
EL ..... English learner
ELD..... English Language Development
ELP ..... English language proficiency
ESEA..... Elementary and Secondary Education Act
ESSA..... Every Student Succeeds Act
FIB..... fill-in-the-blank
GPA..... grade point average
GPC ..... generalized partial-credit [model]
GRAD ..... Graduation-Required Assessment for Diploma
HOSS ..... highest observable scale score
HumRRO ..... Human Resources Research Organization
IEP..... Individualized Education Program
ILSSA ..... Inclusive Large Scale Standards and Assessment
IRT..... item response theory
ISR..... Individual Student Report
KSA..... knowledge, skills, and abilities
LAAC ..... Local Assessment Advisory Committee
LCI ..... Learner Classification Inventory
LIEP ..... Language Instruction Educational Program
LOSS..... lowest observable scale loss
MARSS ..... Minnesota Automated Reporting Student System

MC ..... multiple-choice
MCA ..... Minnesota Comprehensive Assessment
MDE ..... Minnesota Department of Education
MH ..... Mantel-Haenszel
MLE ..... maximum likelihood estimation
MN SOLOM ... Minnesota Student Oral Language Observation Matrix
MTAS ..... Minnesota Test of Academic Skills
MTELL ..... Mathematics Test for English Language Learners
NCLB ..... No Child Left Behind
OIB ..... ordered item booklet
OLPA ..... Optional Local Purpose Assessment
PCA ..... principal component analysis
PSD ..... posterior standard deviation
SATWG..... State Assessments Technology Work Group
SBAC ..... Smarter Balanced Assessment Consortium
SEM..... standard error of measurement
SSR ..... School Student Results
STEM..... science, technology, engineering, and mathematics
SWDs..... students with disabilities
TAC..... technical advisory committee
TCF ..... test characteristic function
TE ..... technology-enhanced
TEAE..... Test of Emerging Academic English
Test WES..... Test Web Edit System
UAT ..... user acceptance testing
WIDA AMS .... WIDA Assessment Management System
WPM..... weighted penalty model

Purpose

This technical manual provides information about the development and measurement characteristics of Minnesota's statewide assessment system. It is organized into two parts: (a) chapters within this document that provide general information about the construction of the statewide assessments, statistical analysis of the results, and the meaning of the test scores and (b) a separate document with appendices organized as the *Yearbook Tables for Minnesota's Statewide Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Yearbook Tables expandable header)) that provides detailed statistics on the various assessments for a given academic year.

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

A large-scale standardized assessment requires evidence to support the meaningfulness of the inferences made from the scores (validity) and the consistency with which the scores are derived (reliability, equating accuracy, and freedom from processing errors). That evidence is reported in this manual.

This manual does not include all the information available regarding the assessment program in Minnesota. Additional information can be found on the Minnesota Department of Education (MDE) website. Questions may also be directed to the Division of Academic Standards, Instruction and Assessment at MDE by email (mde.testing@state.mn.us).

MDE is committed to following generally accepted professional standards when creating, administering, scoring, and reporting test scores. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) is one source of professional standards. As evidence of our dedication to responsible and fair testing practices, an annotated table of contents linking the sections of this manual to the *Standards* is provided after the glossary.

Chapter 1: Background

With the enactment of the Elementary and Secondary Education Act of 1965 (ESEA), as amended by the No Child Left Behind Act of 2002 (NCLB), Minnesota accountability and statewide assessment requirements were dramatically increased. The state was required to 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 to be considered academically proficient—in the core subjects.

According to ESEA, by 2005–06, all students are required to take annual mathematics and reading/English language arts (ELA) tests in grades 3–8 and once during high school. Minnesota provides a reading assessment but believed adding the three additional tests of writing, speaking, and listening was not in the state's best interest, as this would increase testing time for students. In federal peer review, Minnesota received "substantially meets" for both the general and alternate reading assessments in February 2017 and "meets" for the general reading assessments in November 2020. By 2007–08, students were required to be tested in science at least once in each of the following grade spans: grades 3–5, 6–9, and 10–12. Under the ESEA English Language Proficiency Assessments, the state was required to develop and assess English language proficiency (ELP) standards for all students identified as English learners (ELs). This requirement establishes additional tests for EL students.

A more recent update has occurred with the passage of the Every Student Succeeds Act of 2015 (ESSA). Similar to Title I assessments under ESEA, students are required to complete standards-based accountability assessments aligned to the Minnesota Academic Standards. The standards were approved for reading, mathematics, and science in 2010, 2007, and 2009, respectively. Under ESEA and Minnesota Statute 120B.30, all public school students are required to be assessed in both reading and mathematics yearly in grades 3–8 and once in high school as part of Minnesota's accountability system. ESSA and Minnesota Statute 120B.30 also require students to be assessed in science at grades 5 and 8 and once in high school.

The reading, mathematics, and science standards-based accountability tests are given online. Paper accommodated versions are available for students who are unable to take the test online due to a disability. The grades 3–8 Mathematics Minnesota Comprehensive Assessment (MCA) has been adaptive (i.e., each item is selected based on student performance on items taken earlier in the test) since 2011–12, starting with the 2015–16 test administration. The grade 11 Mathematics Minnesota Comprehensive Assessments-Series III (MCA-III) and the grades 3–8 and 10 Reading MCA-III are also adaptive.

Similar to the Title III assessments under NCLB, ESSA requires that Minnesota develop a series of ELP accountability assessments for students identified as ELs, Minnesota uses assessments created by the WIDA consortium that include ACCESS for ELLs (ACCESS) and WIDA Alternate ACCESS for ELLs. The first online operational administration of ACCESS was 2015–16.

1.1. Minnesota Statewide Assessment History

Prior to ESEA in 1965, 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. Below is 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, subd.7 [2024]). The Minnesota Department of Education (MDE) developed a set of test specifications to measure the minimum 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 establish higher academic standards, educators, parents/guardians, and community members from across Minnesota collaborated to develop the Profile of Learning, Minnesota's first version of academic standards, and classroom-based performance assessments to measure these standards. Minnesota developed its assessment program to evaluate student progress toward achieving academic excellence, as measured by the BST and performance assessments of the Profile of Learning.

1997

The Minnesota legislature mandated a system of statewide testing and accountability for students enrolled in grades 3, 5, and 7 (Minn. Stat. 120B.30 [2024]). This legislation required all Minnesota students in those grades to be tested annually using a single statewide test by grade and subject for the purpose of system accountability.

1998

MDE developed the MCAs to fulfill the mandates of the statewide testing statute enacted in 1997. Since 1998, Minnesota students in grades 3 and 5 started testing annually using a single statewide test for the purpose of statewide system accountability. 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 Statutes 120B.30 [2024] (revisor.mn.gov/statutes/cite/120b.30). Special education students were required to participate in all testing according to the recommendations of their Individualized Education Program (IEP) or 504 plan. EL students who were in the United States for less than three years were exempted from the BST.

2001

The Division of Special Education Policy developed Alternate Assessments—checklists for mathematics, reading, writing, and functional skills—to be used in place of the MCA or BST for students whose IEP and 504 plan teams determined it was appropriate. This year also saw the first operational administration of Test of Emerging Academic English (TEAE) and the Minnesota Student Oral Language Observation Matrix (MN SOLOM).

2004

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

2006

In response to NCLB legislation, the Minnesota statewide assessment system was expanded in 2005–06. Students in grades 3–8, 10, and 11 were assessed with the first MCA-II in mathematics and reading. Information from these tests was used to determine proficiency levels in each school and district for the purpose of determining Adequate Yearly Progress (AYP) and to evaluate student, school, and district success in Minnesota's standards-based education system for NCLB. This assessment system would be expanded in future years to meet additional requirements under NCLB.

2007

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

2008

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

2009

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

2010

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

2011

This year saw the first operational administrations of the Mathematics MCA-III and MCA-Modified for mathematics and reading. Districts chose to administer the Mathematics MCA-III either on computer or on paper accommodated forms. The computer version included TE items. The grades 5–8 Mathematics MCA-Modified was computer-delivered. The Mathematics MCA-III, grades 5–8 Mathematics MCA-Modified, and grades 3–8 Mathematics MTAS assessed the *2007 Minnesota K–12 Academic Standards in Mathematics*.

2012

For districts opting to participate online, the grades 3–8 Mathematics MCA-III was administered as a computer adaptive test (CAT) that offered students up to three testing opportunities, with the highest score used for score reporting and accountability purposes. This year also saw the first operational administration of the Science MCA-III in grades 5, 8, and high school that continued to be administered online and assessed the *2008 Minnesota K–12 Academic Standards in Science*. A new English language proficiency accountability assessment was introduced in 2012 known as ACCESS for ELLs (ACCESS), an online ELP accountability assessment given to students identified as ELs in grades K–12 and administered annually in states that are members of the WIDA consortium. Test development for ACCESS is performed by the Center for Applied Linguistics (CAL), and Data Recognition Corporation (DRC) manages the printing, scoring, reporting, online test delivery, and distribution of all ACCESS test materials. Administration of the Test of Emerging Academic English (TEAE) and the Minnesota Student Oral Language Observation Matrix (MN SOLOM) was discontinued.

2013

This year saw the first operational administration of the Reading MCA-III, MCA-Modified, and MTAS aligned to the *2010 Minnesota K–12 Academic Standards in English Language Arts (ELA)*. Districts chose to administer the Reading MCA-III either on computer or paper accommodated forms. The computer version included TE items, whereas the paper accommodated version included only multiple-choice (MC) items. A study linked Reading MCA-III scores to the Lexile scale to permit inferences about Lexile Reading scores based on scores from the Minnesota reading assessments. Grades 5–8 and 10 Reading MCA-Modified were delivered on computer. This year also saw the first operational administration of the Optional Local Purpose Assessment (OLPA) for Mathematics administered as an adaptive test that offered students up to two testing opportunities. The administration of the Mathematics MCA-III was changed in spring 2013 to be a single-opportunity test. Alternate ACCESS was introduced this year as an alternate ELP accountability assessment, administered individually to students identified as ELs with significant cognitive disabilities in grades 1–12.

2014

This year saw the first operational administration of the grade 11 Mathematics MCA-III, MCA-Modified, and MTAS aligned to the *2007 Minnesota K–12 Academic Standards in Mathematics*. Districts chose to administer the grade 11 Mathematics MCA-III either on computer or paper accommodated forms. The computer version included TE items. This year also marked the last operational administration of the Mathematics and Reading MCA-Modified. This was also the last year that districts were allowed to choose between paper accommodated and online assessments for the Mathematics, Reading, and Science MCA-III.

2015

The Mathematics, Reading, and Science MCA-III were administered online only (except for the paper accommodated forms for eligible students). Two career and college readiness (CCR) assessments in grades 8 and 10, Explore and Plan, were census-administrated in fall 2014. The college entrance exam, ACT Plus Writing, was administered to all grade 11 students in spring 2015. A college placement diagnostic exam, COMPASS, was given to some students after the grade 10 Plan and prior to the grade 11 ACT Plus Writing. A study was conducted to determine if students participating in COMPASS were academically ready for career and college based on their performance on the grades 8 and 10 assessments. This was the last academic year in which the GRAD retests were still available as an option to meet graduation assessment requirements for students who first enrolled in grade 8 through 2010–11. The first administration of the Reading OLPA as a single-opportunity, fixed-form online test took place.

2016

This year marked the first operational administration of the adaptive grades 3–8 and 10 Reading MCA-III. During the 2015–16 operational year, the Reading MCA-III was only administered as a multistage CAT, except for eligible students who took the test on paper accommodated forms. This was the first year the grade 11 Mathematics MCA-III was administered as an adaptive assessment to all students, except for students who were eligible to take the assessment on paper accommodated forms. The first operational administration of the online ACCESS for ELLs 2.0 (ACCESS 2.0) was 2015–16, although paper accommodations were still available for eligible students. WIDA also conducted a standard setting study to re-examine proficiency level scores of ACCESS 2.0. This was the first year to include the off-grade items (i.e., items above or below a student's grade) for grades 3–8 mathematics and reading. Progress scores, which were calculated from on- and off-grade items, were reported for grades 3–8 in mathematics and reading, while CCR scores, the same scale scores as the MCA-III accountability scale scores, were reported for grade 11 mathematics and grade 10 reading. The adaptive grades 3–8 Mathematics OLPA item bank was increased, and the grade 11 Mathematics OLPA was administered as a linear, fixed-form assessment only.

2017

This year marked the final operational administration of the Mathematics and Reading OLPA for all grades. This was also the first year that high school grade 10 reading and grade 11 mathematics scores could be used for course placement into Minnesota state colleges and universities.

2018

A study was conducted to link scores on the Mathematics MCA-III to the Quantile® scale to permit inferences about Quantile Mathematics scores based on scores from the Mathematics MCA-III.

2019

The MCA Benchmark Report was redesigned for 2019 to use a different calculation method to measure school and district performance on benchmarks. Because of the change in calculation methodology, 2019 and later benchmark reports cannot be compared to benchmark reports before 2019.

2020

By executive order from the governor on March 27, 2020, at 5 p.m., Minnesota canceled its statewide assessments for the remainder of the 2019–20 school year. The U.S. Department of Education approved a waiver to Minnesota for the federally mandated standardized statewide assessments, accountability, and reporting requirement due to the COVID-19 pandemic. Additionally, starting with the 2020 administration, progress scores on the MCA for grades 3–8 mathematics and reading were no longer reported. Based on the federal guidance, MDE provided an extension of the spring 2020 ACCESS testing window, due to the COVID-19 pandemic. The extended testing window was open from August 3, 2020, to September 25, 2020. The extended window was provided so that students could demonstrate proficiency in and exit their Language Instruction Educational Program (LIEP).

2021

In accordance with the requirements from the U.S. Department of Education that all states administer statewide assessments, as explained in the February 22, 2021, letter to all states, the statewide assessments (MCA, MTAS, ACCESS, and Alternate ACCESS) were administered and reported. Minnesota submitted a waiver for the accountability sections of its ESSA State Plan. On April 21, 2021, MDE was notified that the waiver was approved. Under this waiver, Minnesota was still required to collect and report data on student achievement. However, due to the effects of the COVID-19 pandemic on data collection and usability, data collected during the 2020–21 school year were not used for statewide accountability purposes. This meant the next round of identification of schools for support and improvement under ESSA was delayed until fall 2022. Due to the unknown impact the COVID-19 pandemic might have on test participation or performance, MDE provided the *2021 Statewide Assessment Reporting Guidance* document (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > District Resources > Archive) that provided guidance for districts with information on using 2021 results appropriately and in context.

All testing was required to be done in person; there was no remote (i.e., administration at home) option. To ease the scheduling of testing sessions during the pandemic, greater flexibility was offered by extending the testing window by one week. The MCA and MTAS testing windows were extended through May 21, 2021, and the testing window for ACCESS and Alternate ACCESS was extended to April 16, 2021. Also, due to safety concerns related to the pandemic, no test administration audits were conducted. Furthermore, 2021 saw the first administration of the Science MCA-IV field test items.

2022

Regular testing and reporting resumed in spring 2022. The Science MCA-IV field testing continued to include constructed response (CR) items. With the transition to new academic standards, MDE is developing the Alternate MCA, a redesigned alternate assessment that will replace the MTAS. The Science Alternate MCA field test items were administered for the first time in spring 2022. The timeline for the first administration of the Science MCA-IV, based on the *2019 Minnesota Academic Standards*, has been updated to school year 2024–25. Similarly, the timeline for the first administration of the Reading MCA-IV, based on the *2020 Minnesota Academic Standards*, has been updated to school year 2025–26. Additionally, the timelines for the first administration of the Alternate MCA will follow the same timelines as the Science and Reading MCA-IV.

2023

In spring 2023, Science Alternate MCA field testing continued, and the Reading Alternate MCA field test items were administered for the first time. With this transition to including new field test items, symbolated materials for the Reading MTAS were discontinued this year. The Science MCA-IV field testing continued to include CR items in spring 2023, and the 12-point regular print paper test book was offered for the Science MCA for the first time. The WIDA Alternate ACCESS field test was administered as a stand-alone assessment for all students who participated in the Alternate ACCESS for ELLs, along with kindergarten students who were considered likely to participate in an Alternate English language proficiency assessment. Personalized video individual student reports for students with MCA and MTAS scores were also provided for the first time.

2024

This year marked the first operational administration of the WIDA Alternate ACCESS, an ELP assessment developed for ELs with significant cognitive disabilities. The test is individually administered and designed to measure language skills for accountability purposes. In addition, revisions to Minnesota Statutes, section 120B.30, removed the requirement to include off-grade items in the grades 3–8 Reading and Mathematics MCAs. These off-grade items were not included in the 2024 administration. Furthermore, supports and accommodations for statewide assessments were reorganized. This new framework replaced the former categorization (general supports, linguistic supports, and accommodations) with two main categories (universal supports and accommodations) and include additional subcategories that better organize and describe available supports. Linguistic supports were re-categorized as universal supports to reduce the complexity in determining supports for all learners. It also recognized a more inclusive population of multilingual learners (MLs), as some universal supports may be appropriate for ELs, former ELs, and/or students who have participated in dual-language education programs. Finally, the Student Readiness Tools (SRTs) replaced the items samplers and student tutorials for the MCAs as the resources used to familiarize students with the assessments.

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

Table 1.1. Minnesota Statewide Assessments Chronology

Year	Event
1995–96	• First administration of the grade 8 Mathematics and Reading BST • First administration of the grade 10 BST Written Composition
1997–98	• First administration of the grades 3 and 5 MCAs
1998–99	• Development of high school test specifications for the grades 10–11 MCAs • Field test of the TEAE
2000–01	• First administration the MCA/BST Written Composition • Field test of the grade 11 Mathematics MCA and grade 10 Reading MCA • First administration of the TEAE and First administration of the MN SOLOM
2001–02	• Second field test of the grade 10 Reading MCA and grade 11 Mathematics MCA

Year	Event
2002–03	• First administration of the grade 10 Reading MCA and grade 11 Mathematics MCA • Field test of the grade 7 Mathematics and Reading MCA • Revision of the grade 11 mathematics test specifications
2003–04	• First field test of the grades 4, 6, and 8 Mathematics and Reading MCA • First operational administration (reported) of the grade 7 Mathematics and Reading MCA, grade 10 Reading MCA, and grade 11 Mathematics MCA
2004–05	• Second field test of the grades 4, 6, and 8 Mathematics and Reading MCA
2005–06	• First operational administration of the grades 3–8, 10, and 11 Mathematics and Reading MCA-II
2006–07	• First administration of the grade 9 Written Composition GRAD • Last year of the grade 10 BST Written Composition as a census test • Field test of the MTELL and MTAS • First operational administration of the Mathematics and Reading MTAS • First operational administration of the MTELL
2007–08	• Field test of the MTAS • First administration of the grades 5, 8, and high school Science MCA-II • First administration of the Reading GRAD • First operational administration of the Science MTAS
2008–09	• First operational administration of the Mathematics GRAD
2009–10	• Field test of TE Mathematics MCA-III items • Field test of the Mathematics and Reading MCA-Modified • Lexile® linking study
2010–11	• First operational administration of the grades 3–8 Mathematics MCA-III • Districts given the choice of computer or paper accommodated delivery of Mathematics MCA-III • First operational administration of the Mathematics and Reading MCA-Modified • Last operational administration of the TEAE and First administration of the MN SOLOM
2011–12	• First operational administration of the Science MCA-III and MTAS-III • First year of the Mathematics MCA-III online assessments being delivered as a multi-opportunity computer adaptive assessment • First operational administration of ACCESS as an English language proficiency accountability assessment

Year	Event
2012–13	• First operational administration of the Reading MCA-III, MCA-Modified, and MTAS-III aligned to the *2010 Minnesota K–12 Academic Standards in ELA* • Districts given the choice of computer-based or paper delivery of the Reading MCA-III • Lexile linking study for the Reading MCA-III • First operational administration of the Mathematics OLPA being delivered as a multi-opportunity, computer-based adaptive assessment • The online Mathematics MCA-III reverts to being a single-opportunity assessment • First operational administration of the Alternate ACCESS as an English language proficiency accountability assessment • Census administration of the grade 10 Reading GRAD discontinued • Last year of the grade 9 Written Composition GRAD as a census test
2013–14	• First operational administration of the grade 11 Mathematics MCA-III, MCA-Modified, and MTAS aligned to the *2007 Minnesota K–12 Academic Standards in Mathematics* • Districts given the choice of computer or paper delivery of the grade 11 Mathematics MCA-III • Final operational administration of the Mathematics and Reading MCA-Modified • Discontinuation of the census administration of the grade 11 Mathematics GRAD
2014–15	• Census administrations of the Explore, Plan, and ACT Plus Writing • Final administrations of the Mathematics, Reading, and Written Composition GRAD retests • First operational administration of the Reading OLPA as a single-opportunity, fixed-form online test • First year developing the Reading MCA-III as a computerized adaptive assessment
2015–16	• First operational administration of the adaptive version of the grades 3–8 and 10 Reading MCA-III • First operational administration of the adaptive version of the grade 11 Mathematics MCA-III • First operational administration of the online ACCESS 2.0 • Calculation of first-year progress scores from on- and off-grade items for Reading and Mathematics MCA in grades 3–8 (while the CCR scores, the same scale scores as the MCA-III, are reported for grade 10 reading and grade 11 mathematics) • Increased item pool for the adaptive grades 3–8 Mathematics OLPA • First operational administration of the grade 11 Mathematics OLPA
2016–17	• Last operational administration of the grades 3–8 and 11 Mathematics OLPA • Last operational administration of the grades 3–8 and 10 Reading OLPA • First year the Minnesota state colleges and universities used MCA-III scores for course placement and acceptance
2017–18	• Quantile linking study for Mathematics MCA-III
2018–19	• Redesign of benchmark reports

Year	Event
2019–20	• Closure of the MCA and MTAS administrations on March 27, 2020, for the remainder of the school year due to the COVID-19 pandemic • Approval of a waiver to Minnesota by the U.S. Department of Education for the federally mandated standardized statewide assessments, accountability, and reporting requirements • Progress scores on the grades 3–8 Mathematics and Reading MCA no longer reported • Extension of the spring 2020 ACCESS testing window through the 2020 ACCESS extended testing window from August 3, 2020, to September 25, 2020, from MDE due to the limited testing window resulting from the COVID-19 pandemic so that students could demonstrate proficiency and exit the Language Instruction Educational Program (LIEP)
2020–21	• Waiver granted to Minnesota from the U.S. Department of Education stating that, due to the effects of the COVID-19 pandemic on data collection and usability, data collected during the 2020–21 school year were not to be used for statewide accountability purposes, although Minnesota was still required to collect and report data on student achievement • Testing window extended by one week to ease scheduling during the pandemic; no test administration audits were conducted • Field test items for the Science MCA-IV administered for the first time
2021–22	• Regular testing and reporting resumed in spring 2022 • Science Alternate MCA-IV field test items administered for the first time in spring 2022
2022–23	• Reading Alternate MCA-IV field test items administered for the first time in spring 2023 • The 12-point regular print paper test book offered for the Science MCA for the first time • Personalized video individual student reports for students with MCA and MTAS scores provided for the first time • Last operational administration of off-grade items for Reading and Mathematics MCA in grades 3–8
2023–24	• First operational administration of the WIDA Alternate ACCESS • First year of offering the Student Readiness Tools (SRTs) as a replacement for MCA item samplers and student tutorials • Reorganization of supports and accommodations for statewide assessments

1.2. Organizations and Groups Involved

The following major groups and organizations are involved with the Minnesota assessment program. Each contributor serves a specific role, and their collaborative efforts contribute significantly to the program's success. One testing service provider constructs and administers all tests, while other service providers provide other independent services.

1. Human Resources Research Organization (HumRRO)
2. Local Assessment Advisory Committee (LAAC)
3. Minnesota Department of Education (MDE)
4. Minnesota educators
5. Minnesota's testing service providers
6. National Technical Advisory Committee (TAC)

1.2.1. Human Resources Research Organization

HumRRO is a separate service provider working with MDE to complete quality assurance checks associated with elements of the Minnesota statewide assessment system and accountability program. In collaboration with MDE and Minnesota's testing service provider, HumRRO conducts quality checks during calibration, equating, and scoring of Minnesota's standards-based accountability assessments, including the MCA-III and MTAS-III. HumRRO has also conducted (a) alignment studies to evaluate the congruence between the items on the Minnesota statewide assessments and the skills specified in the Minnesota Academic Standards, (b) form review, and (c) psychometric research as requested by MDE. HumRRO has been in this role since 2006.

1.2.2. Local Assessment Advisory Committee

The LAAC advises MDE on assessment technical issues. Table 1.2 presents the members of this committee.

Table 1.2. Local Assessment Advisory Committee

Name	Position	Organization
Sherri Dahl	Principal	Kelliher Public Schools
Johnna Rohmer-Hirt	District Research, Evaluation, and Testing Achievement Analyst	Anoka-Hennepin Public Schools
Kara Arzamendia	Director of Research, Evaluation and Assessment	St. Paul Public Schools
Donna Roper	Executive Director of Research, Evaluation & Assessment	St. Cloud Area School District
Stacey Lackner	Director of Research and Evaluation	Wayzata Public Schools
Katie Rotvold	Curriculum and Instruction Coordinator	Faribault Public Schools

1.2.3. Minnesota Department of Education

MDE's Division of Academic Standards, Instruction and Assessment has the responsibility of carrying out the requirements in the Minnesota statutes and rules for statewide assessments. The division oversees the planning, scheduling, and implementation of all major assessment activities and supervises the agency's contracts with the service providers (Pearson, HumRRO, and WIDA). Division staff, in collaboration with an outside service provider, also conducts quality control activities for every aspect of the development and administration of the assessment program. Division staff, in conjunction with MDE's Office of the Inspector General, are also active in monitoring the security provisions of the assessment program.

1.2.4. Minnesota Educators

Minnesota educators—including teachers, curriculum specialists, administrators, and members of the best practice networks, who are working groups of expert educators in specific content areas—play a vital role in all phases of the test development process. Committees of Minnesota educators review the test specifications and provide advice on the model or structure for assessing each subject. They also work to ensure that the test content and item types align closely with best practices in classroom instruction.

Draft benchmarks were widely distributed for review by educators, 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 reviewed the items to judge appropriateness of content and difficulty and to eliminate potential bias. Items were revised based on input gathered from these committee meetings. After items were field tested, Minnesota educator committees convened to review each item and its associated data for appropriateness for inclusion in the item bank from which the test forms are built.

Many Minnesota educators have served on one or more of the educator committees involved in item development for statewide assessments. Those who wish to participate may sign up by registering on the Join an MCA/Alternate MCA Review Committee page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Join an MCA/Alternate MCA Review Committee).

1.2.5. Minnesota's Testing Service Providers

Pearson served as a testing service provider for MDE beginning in 1997 and as the primary service provider for all of Minnesota's statewide assessments from 2005 through the close of the 2010–11 test administration cycle. After that, the American Institutes for Research (AIR) served as MDE's primary testing service provider through the close of the 2013–14 test administration cycle. AIR worked with Data Recognition Corporation (DRC)—a service provider primarily responsible for printing, distribution, and processing of testing materials—to manage all standards-based accountability assessments in Minnesota. Beginning with the 2014–15 test administration cycle, Pearson again became the primary service provider, providing Minnesota's standards-based accountability assessments and resources for district and school assessment coordinators.

MDE's testing service providers are responsible for the development, distribution, and collection of all test materials and for maintaining the security of tests. The service providers work with MDE to develop test items and forms, maintain item pools, produce ancillary testing materials that include test administration manuals and interpretive guides, administer tests to students online and on paper accommodated forms, collect and analyze student responses, and report results. Service providers are responsible for scoring all student tests, including paper accommodated tests that are entered online by the administrator and online tests that employ both MC items and other machine-scorable item types. The testing service provider may also conduct standard setting activities, in collaboration with panels of Minnesota educators, to determine the translation of scores on statewide assessments into achievement levels on the Minnesota Academic Standards. Refer to Chapter 5: Performance Standards for information on standard setting.

1.2.6. National Technical Advisory Committee

The National TAC serves as an advisory body to MDE by providing recommendations on technical aspects of large-scale assessment, including item development, test construction, administration procedures, scoring and equating methodologies, and standard setting workshops. The National TAC also provides guidance on other technical matters such as practices not already described in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) and continues to provide advice and consultation on the implementation of new state assessments and meeting the federal requirements set forth by ESSA. Table 1.3 presents the members of the National TAC.

Table 1.3. National Technical Advisory Committee

Name	Position	Organization
Wesley Bruce	Consultant	Indiana
Dr. Gregory J. Cizek	Professor of Educational Measurement and Evaluation, School of Education	University of North Carolina at Chapel Hill
Dr. Claudia Flowers	Professor Emeritus, Educational Research and Statistics	University of North Carolina at Charlotte
Dr. Mark Reckase	Professor Emeritus, Measurement and Quantitative Methods, College of Education	Michigan State University
Dr. Michael Rodriguez	Professor and Dean, College of Education & Human Development	University of Minnesota

1.3. Minnesota Statewide Assessments

MDE provides general information about statewide assessments on the Statewide Testing page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing). Minnesota's testing service provider also maintains a website that provides information about Minnesota's statewide assessments. Materials available on these websites include testing schedules, achievement level descriptors (ALDs), test specifications, technical manuals, other technical reports, and information for parents/guardians.

The standards-based accountability assessments are used to evaluate school and district success in Minnesota's accountability system related to the *Minnesota K–12 Academic Standards in Mathematics, Reading, and Science*. Additional alternate assessments exist for special populations of students, such as students with the most significant cognitive disabilities. All students in grades 3–8, 10, and 11 are required to take standards-based accountability assessments according to their eligibility status. Minnesota's standards-based accountability assessments are listed in Table 1.4 and described below. In addition, ACCESS and WIDA Alternate ACCESS are administered to ELs to measure progress toward the ELP requirement of ESSA. They also serve as evidence of proficiency for determining an individual student's continuing EL eligibility or exit. All ELs are required to take an ELP accountability assessment.

Table 1.4. Standards-Based and ELP Accountability Assessments

Test	Subject	Grades
MCA-III	Mathematics	3–8, 11
	Reading	3–8, 10
	Science	5, 8, 9–12[1]
MTAS-III	Mathematics	3–8, 11
	Reading	3–8, 10
	Science	5, 8, 9–12[1]
ACCESS	English Language Development	K–12
WIDA Alternate ACCESS	English Language Development	K–12

1.3.1. Minnesota Comprehensive Assessments-Series III (MCA-III)

1.3.1.1. Mathematics

The Mathematics MCA-III is aligned to the *2007 Minnesota K–12 Academic Standards in Mathematics* and has been given in grades 3–8 since spring 2011 and in grade 11 since spring 2014. Students are asked to respond to items involving mathematical problem-solving. They answer items about concepts and skills in four different strands:

- Numbers and Operations
- Algebra
- Geometry and Measurement
- Data Analysis and Probability

The Numbers and Operations strand is not assessed for grade 11. Originally, the Mathematics MCA-III could be administered in either online or paper modes based on district choice. Currently, only online administration is available (except for paper accommodated forms). The online Mathematics MCA-III includes MC items and TE item types that allow measurement of higher-level thinking and concepts.

The 2011 online and paper accommodated administrations were fixed forms that included 50 operational items. Beginning in 2012, the online test was administered adaptively and included 42 scored items for grades 3–8. The paper accommodated version included 50 operational MC and fill-in-the-blank (FIB) items in grades 5 and above. A unique feature of the 2012 online Mathematics MCA-III administration was that students were permitted to take the CAT up to three times and use their highest score for accountability purposes. Since 2013, only a single testing opportunity has been allowed. Beginning in 2014, the grade 11 Mathematics MCA-III assessment was administered. The grade 11 Mathematics MCA-III was first administered operationally as an adaptive assessment in 2016 with 47 items. The paper accommodated version of the assessment contains 56 items.

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

Beginning in 2015–16, the Mathematics MCA-III online adaptive test is administered as a fully adaptive test, which is legislatively defined as including off-grade content as appropriate to students' demonstrated levels of ability, and the grade 11 test is also mandated to be adaptive, though off-grade content is not required. Starting in 2023–24, the off-grade part in the grades 3–8 tests is removed as a result of a legislative decision. The Mathematics MCA-III CAT algorithm uses the weighted penalty model (WPM; Shin et al., 2009) to select each item one-by-one during the assessment and the conditional randomesque method (CRM; Shin and Chien, 2017) that controls item exposure. The Mathematics MCA-III also has controls in place for calculator and noncalculator sections of the assessment for grades 3–8 to ensure that students are not allowed to use a calculator in the noncalculator section. Students are administered four noncalculator items in the adaptive online assessments for grades 3–8 mathematics. Grade 11 mathematics does not contain noncalculator benchmarks, so a calculator is available for all items.

1.3.1.2. Reading

The Reading MCA-III is aligned to the *2010 Minnesota K–12 Academic Standards in ELA* and has been given in grades 3–8 and 10 since spring 2013. Students are asked to read both literature and informational text. For literature, students use strategies to analyze, interpret, and evaluate fictional texts such as short stories, fables, poetry, and drama. For informational text, students use strategies to analyze, interpret, and evaluate nonfiction such as expository and persuasive text and literary nonfiction. Originally, the Reading MCA-III could be administered in either online or paper modes based on district choice. Currently, only online administration is available, except for students eligible to take the accommodated paper form.

Beginning in 2015–16, the Reading MCA-III online adaptive test is administered as a fully adaptive test, which is legislatively defined as including off-grade content as appropriate to students' demonstrated levels of ability, and the grade 11 test is also mandated to be adaptive, though off-grade content is not required. Starting in 2023–24, the off-grade part in the grades 3–8 tests is removed as a result of a legislative decision. The Reading MCA-III CAT uses the same algorithm as the Mathematics MCA-III CAT except it adapts section by section rather than item by item.

The online Reading MCA-III was administered in an adaptive mode starting in 2016 with 40 operational items for grades 3–5, 45 operational items for grades 6–8, and 51 operational items for grade 10. The online Reading MCA-III includes MC items and TE item types that allow measurement of higher-level thinking and concepts. Within the assessment, the total word count, passage length and Lexile level, and passage counts are held within defined limits so that all students have similar test forms. The numbers of operational passages for the Reading MCA-III are as follows: four to seven passages for grades 3–8 and four to eight passages for grade 10. The total number of scored items for the paper accommodated administrations for the Reading MCA-III were 48 items for grades 3–5, 54 items for grades 6–8, and 60 items for grade 10. The paper accommodated assessment is administered in four or five separate segments that may be given on different days.

The Reading MCA-III is given as an adaptive assessment in which the items are administered as three testlets, each of which contains one or more passages and their associated items. Similar to the Mathematics MCA-III, the Reading MCA-III has controls to regulate the number of MC and TE items.

1.3.1.3. Science

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

The grades 5, 8, and high school MCA-III was initially administered operationally in spring 2012 as online fixed forms. The assessments had 41, 51, and 68 operational items in grades 5, 8, and high school, respectively. The scored operational item types for science include MC and TE items. Minnesota revised its academic standards in science in 2009 and implemented them in May 2010. Most notably, the revised standards explicitly include engineering knowledge and skills so that they align with the emphasis on science, technology, engineering, and mathematics (STEM) necessary for success in the twenty-first century. In grades 5 and 8, students answer items about concepts and skills in four different strands:

- Nature of Science and Engineering
- Physical Science
- Earth and Space Science
- Life Science

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

- Nature of Science and Engineering
- Life Science

1.3.2. Minnesota Test of Academic Skills (MTAS-III)

The Mathematics and Reading MTAS-III have been developed for grades 3–8 and high school, and the Science MTAS-III has been developed for grades 5 and 8 and high school. The mathematics and science tests consist of a series of discrete items. In reading, the tasks are designed to assess comprehension of the MTAS-III passages.

1.3.2.1. Mathematics

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

1.3.2.2. Reading

The Reading MTAS-III is aligned to the *Minnesota K–12 Academic Standards in ELA* and administered in grades 3–8 and 10. Each test contains a set of nine scored performance tasks designed to measure student understanding of literary or informational text. The content strands are the same as those tested by the Reading MCA-III and mirror their pattern of emphasis but with a reduction in the depth and complexity of concepts measured.

Reading passages for the MTAS-III differ from those appearing on the MCA-III. The MTAS-III 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-III 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-III includes both fiction and nonfiction passages. Passage topics are age-appropriate and generally familiar to the population assessed. Concepts presented in the passages are literal.

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 to 8 and to high school by using grade- and age-appropriate vocabulary and subject matter. The word count and length of the passages are also increased, further adding to the complexity. The performance tasks can be administered on different days according to the needs of the student.

1.3.2.3. Science

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

1.3.3. ACCESS for ELLs

ACCESS is a test of English language proficiency in reading, writing, listening, and speaking administered to ELs in grades K–12. The four language domains are aligned to the WIDA 2012 Amplification[2] of the WIDA English Language Development (ELD) Standards (WIDA, 2012) that describe performance in five areas:

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

[2] Because of the many steps involved in the process of aligning ACCESS to the updated framework, ACCESS is not yet fully aligned with the 2020 edition, even though state education agencies were to begin implementing the 2020 edition at the classroom level. ACCESS will continue using the 2012 version of the standards until 2025–26.

ACCESS has six English language proficiency levels: 1–*Entering*, 2–*Emerging*, 3–*Developing*, 4–*Expanding*, 5–*Bridging*, and 6–*Reaching*. Students performing at Level 1 can communicate using single words, phrases, and simple statements or questions. The performance of students at Level 6 will demonstrate a range of grade-appropriate language use for a variety of academic purposes and audiences. The degree of strategic competence in academic language use among students performing at a Level 6 is such that it facilitates both their access to content area concepts and ideas and their ability to successfully relate information and ideas for each content area. The whole number indicates the student's language proficiency level based on the WIDA ELD Standards. The decimal indicates the proportion within the proficiency level range that the student's scale score represents, rounded to the nearest tenth.

For all four domains in K–12, the stimuli and items are aligned to the content areas listed above and represent the range of proficiency levels included in a form's tier. Graphic stimuli and supports play a large role in the tests for all domains, tiers, and grades.

The kindergarten test is a one-on-one administration using paper materials. The grades 1–12 online tests are administered in the following grade clusters: grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. Paper accommodations are available for students who have an IEP or 504 plan or who have been enrolled in U.S. schools for less than one year. Paper accommodations are administered for the Listening, Reading, and Writing domains in the following grade clusters: grade 1, grade 2, grade 3, grades 4–5, grades 6–8, and grades 9–12. Paper accommodations for the Speaking domain are administered in the following grade-level clusters: grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12.

For the online grades 1–12 ACCESS, the Listening and Reading domains must be administered first (in either order), with the Speaking and Writing domains following (in either order). The student's tier placement for Speaking and Writing are determined by their performance on the Listening and Reading tests. Students are placed into one of three tiers for Speaking (Pre-A, A, or B/C) and one of two tiers for Writing (A or B/C). For the paper grades 1–12 ACCESS, the Listening, Reading, Writing, and Speaking domains can be administered in any order. The Speaking domain must be administered in an individual setting. Educators determine which tier is most appropriate for each student before test materials are ordered based on a student's current proficiency level. The kindergarten test is not tiered but is adapted to each student's performance during the administration.

In grades 1–12, the Listening, Reading, Writing, and Speaking online tests are administered to groups of students. In the Listening test, students listen to audio stimuli and answer MC and TE items. The Reading test contains MC and TE items related to passages, and the Writing domain contains a set of tasks to which students respond. Students in grades 1–3 handwrite their responses to the writing test in a paper booklet.

For grades 4–5, states may select a default of handwritten or keyboarded responses for students. Minnesota has selected handwritten responses in paper booklets for students in grades 4–5, with the option for schools to override this default for students who are comfortable with keyboarding. Students in grades 6–12 keyboard their responses to the Writing test online, but students in grades 6–12 who are inexperienced, unfamiliar, or uncomfortable with keyboarding may handwrite their responses in a Writing response booklet. This decision must be made on an individual student basis, in conjunction with the student, and prior to the student attempting to test. The Speaking test consists of speaking prompts delivered online; students respond by speaking into the microphone of their headsets, and their responses are recorded and sent to the service provider for scoring.

All four domains on the Kindergarten ACCESS are administered within the context of two stories. The test is individually administered, and materials include manipulatives and an activity board. The test is scored by the test administrator.

1.3.4. WIDA Alternate ACCESS

WIDA Alternate ACCESS is an individually administered English language proficiency accountability assessment developed specifically for ELs who have significant cognitive disabilities in grades K–12. It is based on the WIDA English Language Development (ELD) Standards Framework, 2020 Edition. WIDA Alternate ACCESS assesses the domains of Reading, Writing, Listening, and Speaking. The tests for all four domains are aligned to the WIDA ELD Standards describing performance in Language for Social and Instructional Purposes, Language for Language Arts, Language for Mathematics, Language for Science, and Language for Social Studies. The tests are administered in four grade clusters:

- Grades K–2
- Grades 3–5
- Grades 6–8
- Grades 9–12

Unlike ACCESS, WIDA Alternate ACCESS has only one form per grade cluster. Each individually administered domain test is semi-adaptive, meaning that administration ends when the student scores No Response, Incorrect, or Approaches on three consecutive tasks. The domain tests can be administered in any order and on different days, with no minimum or maximum break between the administrations. The Listening and Reading domains contain selected-response items, and the Speaking and Writing domains contain CR items. The domains are individually administered, and all sections are scored by the test administrator.

For WIDA Alternate ACCESS, the following English language proficiency levels apply: 1–*Entering*, 2–*Emerging*, 3–*Developing*, 4–*Expanding*, and 5–*Bridging*. A student who scores at the *Entering* proficiency level on WIDA Alternate ACCESS is not necessarily performing at the same level as a student who scores at the *Entering* proficiency level on ACCESS. A student performing at Level 1 may communicate using emerging expressions of letters or a word. Level 5 performance is characterized by connected ideas with recognizable organization, including simple and compound, connected sentences.

1.4. Graduation Assessment Requirements

To be eligible for a diploma from a Minnesota public high school in prior years, all students were required to fulfill graduation assessment requirements. There were different routes to meeting these requirements depending on what year students were first enrolled in grade 8 in 2012–13. Based on the revisions to Minnesota Statute 120B.30 enacted in 2013, the graduation assessment requirements transitioned from the GRAD requirements to the Career and College Assessments. Legislation in 2015 modified the requirement for the Career and College Assessments so they are no longer administered statewide. Consistent with legislation, student progress scores in mathematics and reading for grades 3–8 and MCA-III student scores in grades 10 and 11 for mathematics and reading provided career and college readiness indicators from 2016 to 2019. Starting in 2016–17, MCA-III scores in grade 10 reading and grade 11 mathematics could be used for course placement into Minnesota state colleges and universities. In 2019, legislation removed the requirement for student progress scores for grades 3–8, beginning with results from the 2019–20 administrations. The grade 10 reading and grade 11 mathematics option for MCA-III scores continue to be used for course placement into Minnesota state colleges and universities.

1.5. Modes of Assessment

The standards-based accountability assessments and ELP accountability assessments are administered in one of the following modes depending on the student characteristics, subject, grade, and assessment being given: online adaptive, online fixed form, and data-entry fixed form.

1.5.1. Online Adaptive Assessments

The items in the operational pools for the statewide assessments have been calibrated so that their difficulty level, or ability to discriminate between students of similar proficiencies, and susceptibility to guessing are known. Based on these items' statistics, the point on the ability continuum where their measurement accuracy is highest and their degree of accuracy relative to other items can be calculated. As students work through the test, their ability is measured after each item has been answered, and the next item (for mathematics) or testlet (for reading) is selected adaptively from the pool by selecting the item or testlet with the highest degree of precision at each student's current level of ability (subject to content coverage requirements and item exposure controls).

CATs are a specific mode of test delivery where each item is adaptively selected for administration based on the responses to the prior items in the test. Because a series of items are taken by a given student in a restricted range of the distribution, adaptive online assessments must have item pools with several items at each point across a wide range of student proficiency. Unlike fixed-form assessments, adaptive assessments rarely result in multiple students being administered identical tests because the item selection algorithms select each item on the test based on prior performance. Item exposure rules also limit the exposure of individual items in the item pool.

1.5.1.1. Advantages of Online Adaptive Assessments

The grades 3–8 and 11 Mathematics MCA-III and grades 3–8 and 10 Reading MCA-III are administered in an online adaptive form that has several advantages. Adaptive assessments are a flexible mode of assessment that accurately measures students' proficiency while also meeting the blueprint constraints listed in the test specifications. The primary advantage of an adaptive assessment is that each student is provided an assessment tailored to their proficiency level. Adaptive tests accomplish this by adaptively selecting each item (for mathematics) or testlet (for reading), depending on the student's performance on previous items on the assessment. This process results in an individualized assessment that best measures each student's proficiency with the items on the test while selecting items that satisfy the test specifications.

The CATs administered to Minnesota students use an algorithm that adaptively selects items to match both student proficiency and the content blueprint according to the test specifications. Because items can both vary in the proficiency they best measure and their ability to discriminate between students of similar proficiency levels, these factors can influence which items are selected for the adaptive assessment. As the ability of an item to discriminate between students of similar proficiency levels increases, so does the information of that given item. More specifically, as the information of an item increases, the error associated with the item decreases, conditional on proficiency. Likewise, as the item information decreases, the error will increase, conditional on proficiency level. In other words, CAT algorithms select items that discriminate well between students of close proficiency and are appropriately difficult for a given student, thus resulting in items with high levels of information and more accurate measurement than a non-adaptive test, which has a fixed set of items administered to the student. Refer to Chapter 9: Reliability for more information regarding the information of items selected for a given assessment.

In addition to estimating student proficiency more accurately than linear, fixed-form assessments, CATs can also estimate student proficiency levels with fewer items. Because CATs only administer items that are appropriate to each student's proficiency level, they can accurately estimate that student's proficiency level with fewer items than a non-adaptive assessment. Similarly, adaptive tests can differentiate between students at or near the cut points on the test because each student's assessment is tailored to their proficiency level. A related advantage is that CATs allow the test developer to more efficiently balance test length and test precision given that adaptive tests can more accurately measure proficiency with fewer items. However, ensuring coverage of the content blueprints often requires several test items close to that of linear fixed-form tests because coverage of all relevant content areas must be achieved.

1.5.1.2. Adaptive Item Selection

The weighted penalty model (WPM; Shin et al., 2009) for content balancing is the algorithm used for the Minnesota CAT assessments. The WPM selects each ordinal item in the test for mathematics and testlet for reading by balancing test specification and item information, and the conditional randomesque method (CRM; Shin and Chien, 2017) is used to control item exposure. The CAT is configured to control for such things as starting person estimates (student theta), the range of theta, item exposure control, conditional ranges of theta, group sizes, and the balance of weights assigned to item information and the blueprint constraints. The testing service provider chooses and refines these parameters based on the results of simulated administrations of the CAT to a sample of students (known as simulees) drawn from a normal distribution with the mean and standard deviation drawn from the previous year's population results. The metadata and statistical parameters for the items in the CAT item pool are summarized and coded into the configuration file, as are the target item count ranges for each blueprint category (content strands, standards, and item types). The purpose of these simulations is to determine parameters that best balance the precision of the test against item exposure and usage rates.

1.5.1.3. Weighted Penalty Model

According to Shin et al. (2009), "The WPM approach attempts to balance content properties across all content categories as well as other non-statistical constraints, while simultaneously considering item information at each item-selection level and the scarcity of items relative to some constraints." Ideally, each item administered in each ordinal position during a test (or each testlet for reading) would contain the item that best measures a student's level of proficiency. However, this becomes more challenging when trying to control item exposure rates and adhere to content and test specification requirements that must be met during item selection. To form a list of candidate items to be selected at each point in the test, the WPM assigns penalties to each item for mathematics and testlet for reading in relation to the current estimated student theta and the remaining blueprint requirements that include target item count ranges for such things as the number of items in the strand/substrand and standard, the number of MC and TE items, and the number of items at the various depth of knowledge (DOK) levels. For each item for each individual, the WPM selects among the items with the lowest penalty values.

The WPM considers both statistical and non-statistical information about the item to select items for the test that address the test specifications and accurately measure the performance of a given student. The WPM adjusts the penalties for each item following each item administration, so an item with a high penalty value at the beginning of the test may have a low penalty value at the end of the test because some constraints were reached while others were not. The penalty value is a combination of both an item's information penalty value, resulting from the weight the WPM should place on item information, and content penalty value, indicating the weight the WPM should place on the test specifications in the computer control files. Moreover, ability estimates are recalculated after each student response to select items. Expected a priori (EAP) estimation is used to estimate student ability at the beginning of the test until a nonperfect or nonzero response string is observed. In other words, as soon as a student has received at least one item correct and one item incorrect, maximum likelihood estimation (MLE) is used to estimate student abilities.

The WPM has three stages: (a) calculate the weighted penalty value for each eligible item in the pool, (b) assign each eligible item into different groups (referred to as "color groups"), and (c) form a list of candidate items (it is here the CRM is used for item control). The remainder of this section describes the process the WPM takes prior to the selection of each item on the test for each student. The number of times the WPM calculates the penalty values depends on the number of items on the test.

For any given constraint *j*, the $Prevalence_j$ is the proportion of items in the pool that have the property associated with constraint *j*. $Upper_j$ and $Lower_j$ are also defined that are the upper bound and lower bound of the constraint, respectively. The Mid_j is the midpoint between $Upper_j$ and $Lower_j$. For example, consider a strand requirement with an item range of 7–14 items with a 42-item test when there are 700 items in the pool, and 175 of those items correspond to a specific strand. If no items had been administered prior, the calculations would be the following: $Upper_j$ = 0.02 (14/700), $Lower_j$ = 0.01 (7/700), Mid_j = 0.015, and $Prevalence_j$ = 0.25 (175/700).

Prior to the administration of any given item on the test, $Prop_j$ is first calculated, which is the expected proportion of items with constraint *j* that will have been administered if all remaining items in the test are selected in proportion to their prevalence:

$$Prop_j = (nadm_j + Prevalence_j \times nremaining)/testlength \tag{1.1}$$

where $nadm_j$ is the number of items to this point in the test administration having this property, *nremaining* is the number of items remaining to be administered in the test, and $testlength$ is the length of the test. To be more specific, if the constraint in question is the same strand as discussed above, with a range of 7–14 being addressed, $nadm_j$ represents the number of items already administered to the student from that strand. Therefore, if six items were already administered from that strand, $nadm_j$ would be six. Next, X_j, or the expected difference between $Prop_j$ and Mid_j across the full length of the test, is calculated:

$$X_j = (Prop_j - Mid_j). \tag{1.2}$$

For each item, the actual penalty value is then calculated for each constraint *j* using one of the following three equations depending on values for $Prop_j$, $Lower_j$, and $Upper_j$:

$$P_{ij} = (\frac{1}{kD_j} X_j^2 + \frac{D_j}{k}) \times Z_{ij}, if\ Prop_j < Lower_j \tag{1.3}$$

or

$$P_{ij} = (\frac{1}{kA_j} X_j^2 + \frac{A_j}{k}) \times Z_{ij}, if\ Prop_j \geq Upper_j \tag{1.4}$$

or

$$P_{ij} = X_j \times Z_{ij}, if\ Upper_j > Prop_j \geq Lower_j. \tag{1.5}$$

where D_j is $(Lower_j - Mid_j)$, A_j is $(Upper_j - Mid_j)$, *k* is an arbitrary constant that the service provider constrains to a value of 2, and Z_{ij} is a dummy-coded variable equal to 1 if item *i* has property *j*; otherwise, Z_{ij} equals 0. For example, the *j* in Z_{ij} could refer to Strand 1, so if a given item measures that strand, then Z_{ij} = 1. If the item instead measured Strand 2, it would equal 0. Finally, because the prior calculations were repeated for each content constraint, the total content penalty value is calculated as follows, which considers all content constraints:

$$F_i''' = \sum_{j=1}^{J} P_{ij} \times w_j, \tag{1.6}$$

where w_j is the weight for constraint *j*. These weights are determined by running a set of simulations prior to test administration to best balance estimating student ability as accurately as possible (by reducing the error of measurement) with meeting content constraints to achieve blueprint satisfaction. Lastly, the total content constraint penalty value is standardized as follows:

$$F_i' = \frac{F_i''' - \min(F_i''')}{max(F_i''') - \min(F_i''')}, \tag{1.7}$$

where $min(F_i''')$ and $max(F_i''')$ are the minimum and maximum F_i''' across all eligible items remaining in the pool.

The information penalty value is calculated separately from the content penalty value. For any item *i* for a given estimate of ability $(\hat{\theta})$, the standardized item information value is calculated as follows:

$$SI_i(\hat{\theta}) = \frac{I_i(\hat{\theta})}{I_{\max(\hat{\theta})}}, \tag{1.8}$$

where $I_i(\hat{\theta})$ is the information of item *i* given a specific $(\hat{\theta})$, and $I_{max}(\hat{\theta})$ is the maximum information value across all eligible items given a specific $(\hat{\theta})$is used to compute the information penalty value:

$$F_i'' = -SI_i(\hat{\theta})^2, \tag{1.9}$$

Lastly, the weighted penalty value (total penalty value) for a given item is calculated by combining the content and information values:

$$F_i = w' \times F_i' + w'' F_i'', \tag{1.10}$$

where w' and w'' are the weights for $F'$and F'', respectively, and refer to the content constraint and information weights, respectively.

These penalty values are then used to categorize items by color group, each of which represents the eligibility of an item depending on its status regarding the test specification constraints. Because the assessments have multiple constraints that are controlled for, a single item may have already reached the upper bounds for one set of constraints but may be under the minimum range for a different set of constraints. To reduce the frequency that a single constraint will be out of range, penalty values are categorized into groups that are identified by different colors.

The WPM selects items based on the penalty values for the following color groups: first green, then orange, then yellow, and finally, red. The items are ordered by the weighted penalty values within the color groups from smallest to largest. To assign items to color groups, items are first assigned the letter A, B, or C based on the following rules:

- If the lower bound of the constraint has not been reached, the item receives an A for this specific constraint.
- If the lower bound of the constraint has been reached but not the upper bound, B is assigned for this specific constraint.
- If the upper bound has been either reached or exceeded, C is assigned to this specific constraint.

After all constraints have been assigned flags, the items are placed into the color groups:

- Items with *all* assigned flags of A and B for all constraints are placed in the green group.
- Items with a combination of *either* assigned flags of A, B, and C *or* of assigned flags of A and C are placed in the orange group.
- Items with *all* assigned flags of B are placed in the yellow group.
- Items with *either* assigned flags of B and C *or* assigned flags of C only are placed in the red group.

The penalties for the MCA-III use a constant blueprint constraint weight and a constant information weight over the course of the test.

1.5.1.4. Conditional Randomesque Method

After the list of candidate items has been compiled, the CRM is used for item exposure control. Conditional on the current estimated theta (proficiency level), the CRM selects a group of items from the group of the most informative items, of which the group size is determined through the simulations prior to the operational administration of the assessment. The CRM then randomly selects the next item to be administered from that group.

For example, if a group size of three was determined through the simulation to perform optimally for a conditional range, the CRM will select three items from the most optimal candidate item group (from the green color group unless the green color group contains no more items) and then randomly select one of those three items. The remaining items in the group will be blocked from appearing on the remaining portion of the assessment. If the green color group has fewer than three items, the items will be selected from the next available grouping (e.g., orange). The group size is chosen with consideration of the number of items available at each point along the proficiency distribution that will satisfy the content and test specifications. If at any point in the administration of the test there are no available items because all eligible items have been previously blocked (but not previously selected), previously blocked items will be released.

Conditional on theta, the CRM controls the item exposure by a factor of 1/N where *N* is the group size. Therefore, a group size of two items results in an item exposure rate of 0.5 (or 1/2), and a group size of three items results in an item exposure rate of 0.33 (or 1/3). However, because the exposure controls are conditional, even a group size of one does not necessarily lead to an item exposure of 100% across all students, although it would across the students who happened to fall within or pass though the respective corresponding region of the ability continuum. Given that only a fraction of students is in any particular region of the ability continuum (or pass through at some point during the test), the exposure of items in such regions remains limited.

This algorithm controls item exposure because it allows for the random selection of items from among a set of informative items instead of only selecting the most informative item. If an item was not randomly selected from a group of multiple items, the most highly discriminating item (most informative) would have very high exposure, even though there might be other suitable items available for administration. Because the starting theta is controlled, the first item on the assessment will be selected based on the MDE-approved starting theta control value selected from the simulation. Although Minnesota uses adaptive assessments, the test length is fixed for all MCA-III assessments.

The group size parameter is controlled for conditionally on the current estimated level of theta. This allows for a more lenient level of exposure control to be employed in regions of the ability range addressed by relatively few items. Stringent exposure controls applied to regions of the ability continuum addressed by relatively few items would result in elevated levels of error in the ability estimates from the CAT compared to less stringent controls. This error occurs because items inappropriate to (e.g., relatively far away from) a student's true ability level would be included in the candidate group of items. Relaxation of the level of exposure control allows for the selection of more appropriate items and more precise measurement of student ability, albeit at the price of a higher level of exposure of the items in the affected region.

1.5.1.5. Online Adaptive Scale Score Estimates

The Mathematics and Reading MCA-III adaptive tests estimate theta for multiple reasons. First, all items in the adaptive CAT item pool for a given grade and subject are on the same scale. Item response theory (IRT) is used for field test equating, which places all items within a grade/subject on the same scale. (Refer to Chapter 7: Equating and Linking for equating details.) Because all items are on the same scale, a direct theta-to-scale-score transformation can occur. Generally, raw-score-to-scale-score tables are not created for three-parameter logistic (3PL) IRT adaptive-based tests like the mathematics and reading assessments.

Second, the CAT algorithm will attempt to meet the blueprint specifications set forth in the test specifications. One result of the algorithm meeting the specifications is that all students will be exposed to the same content on the assessments. Thus, it can be assumed that even though students are taking different tests, the tests are measuring the same content. Also, the data-entry assessments, described below, have been placed on the same metric as that of the operational adaptive items. Therefore, scores from the data-entry and adaptive forms of the assessment are on the same metric and can be directly compared.

1.5.2. Online Fixed-Form Assessments

The grades 5, 8, and high school Science MCA-III is administered as an online fixed-form assessment based on the online test specifications. The primary difference between the online adaptive and the online fixed-form assessments is that the items administered to the students in the fixed form are pre-selected and fixed to the form. The specific items found on a given assessment appear in the order corresponding to the form the student is administered. Each MCA-III form contains a different set of field test items but the same operational items. Great care is taken during the test construction process to ensure that items on the fixed-form assessments meet the test specifications and measure students from across the distribution of proficiency, with an emphasis of accurate measurement near the cut points. The presentation of items and online navigation system is identical to that of the online adaptive assessments.

1.5.3. Data-Entry Fixed-Form Assessments

The data-entry fixed-form assessments are administered in a one-on-one setting using a paper accommodated test form. The administrator enters the student's responses into the testing system. The grades 3–8 and 11 Mathematics MCA-III, the grades 3–8 and 10 Reading MCA-III, and the grades 5, 8, and high school Science MCA-III have data-entry forms that can be administered to eligible students (i.e., students who are unable to take the test on a computer). All MTAS-III assessments are assessed in this way. Like the online fixed-form assessments, the items on the data-entry fixed forms are fixed prior to administration. One primary difference between the online fixed-form and data-entry assessments is that the data-entry assessments are first given in a one-on-one setting and then manually entered into the computer by a test administrator in the district (often a teacher), whereas responses to the online assessments are provided directly by the student. A second difference is that the data-entry fixed-form assessments do not contain field test items, whereas both the online adaptive and online fixed-form assessments do.

Chapter 2: Test Development

Test development for each Minnesota assessment includes several activities designed to ensure the production of high-quality assessments that accurately measure the achievement of students regarding the knowledge and skills contained in the Minnesota Academic Standards. The standards are intended to guide instruction for students throughout the state, and the 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 reviewed 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. (achieve.org)
- Recommended *Standards for Information and Technology Literacy* from the Minnesota Educational Media Organization
- Content standards from other states

The following steps summarize the process followed to develop large-scale criterion-referenced assessments such as the MCA and MTAS:

1. **Development of test specifications and item development guides.** Committees of content specialists develop test specifications and item development guides that outline the requirements of the test, such as eligible test content, item types and formats, content limits, and cognitive levels for items. These documents are created 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 test structure is based on judgments made by Minnesota educators, students, and the public. Minnesota educators guide all phases of test development.
2. **Development of items, stimuli (passages and phenomena), and tasks.** Using the standards, test specifications, and item development guides, MDE Academic Standards, Instruction and Assessment staff and Minnesota's testing service provider work with the item development service provider to develop culturally affirming items, stimuli (including reading passages and science phenomena), and tasks.
3. **Item (and stimulus) content review.** All members of the assessment team review the developed items (and passages/phenomena for reading and science), discuss possible revisions, and make changes when necessary.
4. **Item (and stimulus) content review committee.** Committees of expert educators review the items and passage/phenomena (some of which are revised during content review) for appropriate difficulty, grade-level specificity, and potential bias and sensitivity issues. Committees of community members review passages and phenomena for Reading and Science to ensure culturally affirming content and review items to ensure inclusive language and content.

5. **Field testing.** Items are taken from the item content review committees, with or without modifications, and are field tested as part of the assessment program. Data are compiled regarding student performance, item difficulty, discrimination, reliability, and possible bias.
6. **Data review.** Committees review the items based on the field test data and make recommendations regarding the inclusion of the items in the item bank.
7. **New forms construction.** Items are selected for each test according to test specifications. Selection is based on content requirements and statistical (equivalent passing rates and equivalent test form difficulty) and psychometric (reliability, validity, and fairness) considerations. The Mathematics MCA-III item pool is finalized for operational administration, and new testlets for the Reading MCA-III are built from newly field tested and existing operational items. These are combined with selected testlets from previous administrations and together comprise the testlet pools used for the final operational administration.

2.1. Test Specifications

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

The test specifications developed by MDE 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 educators. Thus, the content standards and their strands, substrands, and benchmarks serve as the basis for the test specifications. Item types, cognitive levels of understanding to be tested, range in the number of items, and content limits are assigned to each benchmark within the standards.

The item formats are constrained by the test delivery system (online or paper). 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 to determine the assigned cognitive levels for each benchmark. Cognitive levels for benchmarks are determined independently of the item formats and difficulty of the content; this runs counter to many people's perceptions that cognitive level and content difficulty are equivalent concepts. For example, a benchmark measured at a high cognitive level could be assessed with different item formats, such as an MC or TE item. Similarly, the educator committees base the ranges in number of items and content limits on two things: (a) the emphasis that a benchmark is given in the classroom and (b) the type of curriculum content regularly taught to students in a grade level. This discussion guides the final information included in the test specifications.

Test specifications facilitate building a technically sound test that is consistent from year to year. They demonstrate MDE's respect for teacher concerns about the amount of time students spend taking tests, and they account for the grade and age of students involved and other pedagogical concerns. Test specifications define, clarify, and/or limit how test items will be written. The test specifications also provide a basis for interpreting test results.

2.1.1. Minnesota Comprehensive Assessments-Series III (MCA-III)

MDE held meetings with Minnesota educators to define general test specifications for each grade for the MCA-III. 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/substrands, standards, or benchmarks were not suitable for the large-scale assessments. These were clearly identified as content to be assessed in the classroom. After the measurable components of the standards were identified, teacher committees set item formats, cognitive levels, and content limits for each benchmark. Item prototypes were developed as part of the development of the test specifications. Committees of Minnesota educators reviewed drafts of these specifications, and their suggestions were incorporated into the final versions of the test specifications. The complete MCA-III test specifications documents are available on the MDE website (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Test Specifications).

2.1.2. Minnesota Test of Academic Skills (MTAS-III)

A systematic and iterative process was used to create the MTAS-III test specifications. Prior to the onsite benchmark extensions meetings, MDE met with stakeholder groups and their service providers (Minnesota's testing service provider and the Inclusive Large Scale Standards and Assessment [ILSSA] group) 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:

- The grade-level benchmark was assessed on the 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 (i.e., 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-III, including multiple substrands, cognitive levels, and benchmarks. The recommended benchmarks were taken to teacher groups that developed 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 educators analyzed the recommended benchmarks using their professional expertise and familiarity with the target student population and made changes to a subset of the recommended benchmarks in mathematics, reading, and science.

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

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

MDE recognizes that the students who take the MTAS-III are a heterogeneous group. To help ensure that every student in this group has access to the test items, student communication modalities were considered and accommodations were made. Six teacher groups, composed of curriculum experts and both special and general educators, were convened to write these entry points for three grade bands in mathematics and reading and each grade-level assessment in science. After approximately one half-day of training, the teacher groups wrote entry points for each selected benchmark included on the MTAS-III. 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 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 original specifications were published on the MDE website in December 2006. Since then, the test specifications have been updated in coordination with revisions to academic standards and specifications for the general assessments in mathematics, reading, and science. The complete MTAS-III test specifications documents are available on the MDE website (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Test Specifications).

2.1.3. ACCESS for ELLs and WIDA Alternate ACCESS

ACCESS is based on the WIDA 2012 Amplification[3] of the WIDA ELD Standards (social and instructional language, language of language arts, mathematics, science, and social studies) and WIDA performance definitions that describe the linguistic complexity, language forms and conventions, and vocabulary used by students at six proficiency levels. WIDA Alternate ACCESS forms are based on the WIDA ELD Standards Framework, 2020 Edition and WIDA alternate performance definitions at five proficiency levels. Documents describing these standards and performance definitions are available on the WIDA website (wida.wisc.edu/teach/standards/eld).

WIDA does not publish test blueprints or specifications on its website, but district staff who have logins to the secure portal of the WIDA website can access the Test Administration Manuals that contain limited information about the organization of the tests. The Test Administration Manuals are available in the Download Library under the "Resources" heading in the secure WIDA portal (portal.wida.us/IDP/Account/Login).

2.2. Item Development

This section describes the item writing process used during the development of test items (including stimuli/phenomena) and, in the case of the MTAS-III, performance tasks. Minnesota's testing service provider has the primary role for item and task development, but MDE and state review committees also participate in the item development process. Item and task development is a complex, multistage process.

For the Mathematics and Reading MCAs, items and tasks are written and internally reviewed by the testing service provider before submission to MDE. The Science MCA items and stimuli/phenomena are written by Minnesota educators before review by the testing service provider and submission to MDE. For each subject and grade, MDE receives an item tally sheet displaying the number of test items by benchmark and target. Item tallies are examined throughout the review process. Additional items are written by the testing service provider, if necessary, to complete the requisite number of items per benchmark.

2.2.1. Content Limits and Item Specifications

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

[3] Because of the many steps involved in the process of aligning ACCESS to the updated framework, ACCESS is not yet fully aligned with the 2020 edition even though state education agencies were to begin implementing the 2020 edition at the classroom level. ACCESS will continue using the 2012 version of the standards until 2025–26.

2.2.1.1. Minnesota Comprehensive Assessments

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

2.2.1.2. Minnesota Test of Academic Skills

Criteria outlined by the National Alternate Assessment Center served as a guide in the development of the MTAS-III to help ensure that items were based on the Minnesota Academic Standards. All the content of the MTAS-III is academic and derived directly from the standards in mathematics, reading, and science. The content limits of the MTAS-III provide clarification of the way the depth, breadth, and complexity of the academic standards have been reduced. In mathematics, this might concern the number of steps required of a student to solve a problem. In reading, this could involve a restriction in the number of literary terms assessed within a benchmark. In science, this might be addressed by requiring knowledge of only major aspects of the water cycle.

2.2.1.3. ACCESS for ELLs and WIDA Alternate ACCESS

The complexity of tasks called for by ACCESS is delineated by the WIDA performance definitions that describe the linguistic complexity, language forms and conventions, and vocabulary used by students at five proficiency levels. The sixth proficiency level represents the end of the proficiency scale continuum and is characterized by performance that meets all criteria through level five. The tasks for WIDA Alternate ACCESS are based on the WIDA alternate model performance indicators that describe the linguistic complexity, language forms and conventions, and vocabulary usage in the performance of ELs with significant cognitive disabilities. Documents describing the performance definitions for both assessments are available in the WIDA Resource Library (wida.wisc.edu/resources).

2.2.2. Item Writers

Minnesota's testing service provider uses item writers with extensive experience developing items for standardized achievement tests. The service provider selects item writers for their knowledge of the specific content area and for their experience in teaching or developing curricula for the relevant grades. For the Science MCA, the testing service provider hires Minnesota science educators who are trained in writer workshops facilitated by the service provider with input from MDE science staff. All item writers are approved by MDE and follow the same training process.

2.2.2.1. Minnesota Comprehensive Assessments

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

2.2.2.2. Minnesota Test of Academic Skills

MTAS item writers include both general and special education educators and must have experience with and a clear understanding of the unique needs of students with significant cognitive disabilities, in addition to meeting the standards for the MCA. Item writing assignments are divided between both general and special education educators to ensure coverage of the content breadth and maximum accessibility for students with significant cognitive disabilities. 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-III test specifications, and the MTAS-III essence statements. Evaluative feedback is provided to item writers from service provider content and alternate assessment specialists throughout the process to ensure submission of performance tasks that meet the grade level, content, and cognitive requirements.

2.2.2.3. ACCESS for ELLs and WIDA Alternate ACCESS

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

2.2.3. Item Writer Training

Minnesota's testing service provider 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. Minnesota's testing service provider also 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.

2.2.3.1. Minnesota Comprehensive Assessments

Minnesota's testing service provider conducts comprehensive item writer training for all persons selected to submit items for the MCA-III. Training includes an overview of the test development cycle and specific training in the creation of high-quality, culturally affirming MC and TE items. Experienced service provider staff members lead the trainings and provide specific and evaluative feedback to participants.

2.2.3.2. Minnesota Test of Academic Skills

Minnesota's testing service provider conducts item writer training for the MTAS-III that focuses on including students with significant cognitive disabilities in large-scale assessments. Item writers are specifically trained in task elements, vocabulary appropriateness, bias and sensitivity considerations, and significant cognitive disability considerations. The testing service provider recruits item writers who have specific experience with special populations, and the focus of the training is on the creation of performance tasks and reading passages. Performance tasks must

- match the expected student outcomes specified in the Benchmark Extensions document;
- follow the format of the template provided by the testing service provider;
- clearly link to the essence statement and be unique to the specific essence statement;

- represent freedom from bias and sensitivity;
- represent high yet attainable expectations for students with the most significant cognitive disabilities;
- include clearly defined teacher instructions and student outcomes; and
- lend themselves to use with assistive technology and other accommodations.

2.3. Item and Stimuli Review

2.3.1. Service Provider Review

Experienced testing service provider staff members, as well as content experts in the grades and subject areas for which the items (including stimuli/phenomena such as passages and scenarios) or performance tasks (for MTAS-III) 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. Minnesota's testing service provider also instructs the reviewers to consider other issues, including the appropriateness of the items and tasks to the objectives of the test, difficulty range, clarity, correctness of answer choices, and plausibility of the distractors. Minnesota's testing service provider 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 MDE for review.

2.3.1.1. Minnesota Comprehensive Assessments

Before an MCA item may be field tested, it must be reviewed and approved by the content committee and the bias and sensitivity committee. The content committee's task is to review item content and scoring rubric to ensure that each item

- is an appropriate measure of the intended content (strand, substrand, standard, and benchmark),
- is appropriate in difficulty for the grade level of the students,
- has only one correct or best answer (for MC items), and
- has an appropriate and complete scoring guideline (for TE items).

The content committees can make one of three decisions about each item: (a) approve the item and scoring rubric; (b) conditionally approve the item and scoring rubric with recommended changes or item edits to improve the fit to the strand, substrand, standard, and benchmark; or (c) reject the item and thus remove it from consideration for field testing. Each item is coded by content area and item type and presented to MDE for final review and approval before field testing. The final review encompasses graphics, artwork, and layout. The Community MCA Review Committee, whose purpose is centered around reviews of bias and sensitivity issues, reviews each item to identify language or content that might be inappropriate or offensive to students, parents, or community members or that contain stereotypical or biased references to gender, ethnicity, or culture. The committee accepts, edits, or rejects each item for use in field tests.

2.3.1.2. Minnesota Test of Academic Skills

Before an MTAS passage or item may be field tested, it must be reviewed and approved by the content committee and the bias and sensitivity 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 students, and
- has only one correct or best answer for each MC item.

The content committees can make one of three decisions about each item: (a) approve the item and scoring rubric as presented; (b) conditionally approve the item and scoring rubric with recommended changes or item edits to improve the fit to the strand, substrand, standard, and benchmark; or (c) eliminate the item from further consideration. Each 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. The Community MCA-Review Committee reviews each passage and item to identify language or content that might be inappropriate or offensive to students, parents, or community members or that contain stereotypical or biased references to gender, ethnicity, or culture. The committee accepts, edits, or rejects each item for use in field tests.

2.3.1.3. ACCESS for ELLs and WIDA Alternate ACCESS

CAL and WIDA recruit educators from across the consortium to participate in item and bias reviews. Following field testing of new items, CAL and WIDA also recruit educators from across the consortium to serve on data review panels.

2.3.2. MDE Review

MDE and Minnesota's testing service provider review all newly developed items and tasks prior to educator committee review. During this review, content assessment staff assess each item for content-to-specifications match, difficulty, cognitive demand, and plausibility of the distractors, rubrics, and sample answers and for any ethnic, gender, economic, or cultural bias. Content assessment staff from MDE and Minnesota's testing service provider discuss each MCA item, addressing any concerns during this review. Edits are made accordingly, prior to item review with educators. Similarly, assessment staff with both content and students-with-disabilities (SWD) expertise from MDE and Minnesota's testing service provider discuss each item, addressing any concerns during this review. Edits are made accordingly, prior to item review with educators.

All development and review for ACCESS and WIDA Alternate ACCESS is performed by CAL, ATLAS, and WIDA. Consortium member states do not review items as a matter of course, although they may send state education agency staff to participate in item, bias, and data reviews.

2.3.3. Item and Stimuli Committee Review

During each school year, MDE convenes committees of educators, curriculum directors, and administrators from across Minnesota to work with MDE in reviewing all newly developed test items (including passage/phenomena and performance tasks developed for use in the assessment program) and all new field test data. Approximately 40 committee meetings are convened involving Minnesota educators who represent school districts statewide.

MDE seeks recommendations for educator review committee members from best practice networks, district administrators, district curriculum specialists, and subject-area specialists in MDE's Academic Standards, Instruction and Assessment Division and other divisions in MDE. MDE selects educators to be committee members based on their expertise in a particular subject. The selection of committee members represents the regions of the state, major ethnic groups in Minnesota, and various types of school districts (e.g., urban, rural, large, and small districts).

MDE Assessment staff, along with measurement and content staff from Minnesota's testing service provider, 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 they should approve the item and scoring rubric, approve the item and scoring rubric with recommended changes, or reject the item and thus remove it from consideration from field testing. During this review, if committee members judge an item to be questionable for any reason, they may recommend the item be rejected and thus 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 groups.

2.3.3.1. Minnesota Comprehensive Assessments

Item review committees are composed of educators in a given content area who are selected so that the committee appropriately represents the state in terms of geography, ethnicity, and gender. Educators are also selected to represent ELs and special education licensures. Content area educators serving on these committees are familiar with the Minnesota Academic Standards, which they use to ensure item alignment based on subject-specific item review checklists, provided below. MDE and its testing service provider facilitate educators' discussion of the test items.

Mathematics item review checklist:

1. Is the intent of the item readily apparent and understandable as stated without having to read the answer options or re-read the item multiple times?
2. Is the item straightforward and direct with no unnecessary wordiness?
3. Is the item grammatically correct and in complete sentences whenever possible?
4. Are there any clues or clang words used within the item that may influence the student's response?
5. Is the context of the item factually correct or plausible?
6. Does each item function independently of other items?
7. Does the item clearly align to the intended benchmark?
8. Is the cognitive level (DOK) appropriate for the level of thinking required?
9. Does each MC item have only one correct answer?
10. For MC items, are all distractors plausible yet incorrect?
11. Do TE items address content in a meaningful way?
12. For TE items, are the rationales aligned to the items asked?

Reading item review checklist:

1. Is the intent of the item readily apparent and understandable as stated?
2. Is the item clearly written, and is it grammatically correct?
3. Are there any clues or clang words used which may influence the student's responses?
4. Does each MC item have only one correct answer and three plausible yet incorrect answers that are passage-based?
5. Do TE items address content in a meaningful way?
6. For TE items, are the rationales aligned to the items being asked?
7. Does the item clearly align to the intended benchmark and/or standard?
8. Is the DOK appropriate for the level of thinking required?
9. Do items address a range of standards and benchmarks for each passage set?

Science item review checklist:

1. Does the item directly relate to sense-making of the phenomenon?
2. Are the tabs necessary to answer the item?
3. Does the item contain equitable, grade-appropriate vocabulary and content?
4. Does the item clearly align to the intended benchmark?
5. Is the cognitive level (DOK) appropriate for the level of thinking required?
6. Is the intent of the item apparent and understandable to the student without having to read the answer options?
7. For MC items:
    a. Is there only one correct answer?
    b. Are all answer options homogeneous?
    c. Are all distractors plausible yet incorrect?
8. For TE items:
    a. Are rubrics aligned to the items being asked?
    b. Is the type of item appropriate?

2.3.3.2. Minnesota Test of Academic Skills

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

2.3.3.3. ACCESS for ELLs and WIDA Alternate ACCESS

Item review committees for ACCESS and WIDA Alternate ACCESS are convened by CAL and WIDA. These organizations follow industry standards when conducting item review committee meetings.

2.3.4. Bias and Sensitivity Review

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

2.4. Field Testing

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

2.4.1. Embedded Field Testing

MDE embeds field test items in multiple forms of operational tests, or, in the case of the MCA-III adaptive test, the field test items are randomly assigned to students across the state during administration to ensure that a large representative sample of responses is gathered under operational conditions for each item. Responses to most field test items are obtained from approximately 3,000–6,500 students. Research studies have shown that these procedures yield sufficient data for precise statistical evaluation of a large number of field test items in an authentic testing situation. Enough field test items are administered annually to replenish and improve the item pools.

Responses on field test items do not contribute to a student's scores on the operational tests. The specific locations of the embedded items within the assessment are not disclosed. To prevent position effects from contaminating item parameter estimates, items appear at a variety of locations randomly. 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 operational tests under standard administration procedures.

2.4.2. Stand-Alone Field Testing

When MDE implements testing at new grade levels, for new subject areas, or for revised academic standards, it is necessary to conduct a separate stand-alone field test to obtain performance data. When stand-alone 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 the 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.

2.5. Data Review

MDE convenes data review committees of qualified Minnesota educators to evaluate several statistical analyses based on classical test theory and IRT for the field test items. Significant effort goes into ensuring that these committees represent the state demographically regarding ethnicity, gender, school district size, and geographical region. These committees receive training on interpreting the psychometric data compiled for each field test item from psychometricians (typically people with an advanced degree in the application of statistical analyses to measurement), content experts (usually former educators or item writers), and group facilitators for the data review committee meetings. Data provided to the data review committee include the following:

- Numbers of students by ethnicity, gender, and EL status in each sample
- Percentage of all students choosing each response for MC items and percentage of students choosing correct, top-five incorrect, and other incorrect responses for TE items
- Low-, median-, and high-ability distributions based on performance on the overall test and that group of students' distribution choosing responses
- Item mean (*p*-value) and item-total correlation (point-biserial correlations) summarizing the relationship between each response on a particular test item and the score obtained on the total subject area test
- IRT statistical indices to describe the relative difficulty, discrimination, and guessing of each test item and Mantel-Haenszel (MH) statistical indices[4] to identify greater-than-expected differences in performance on an item associated with gender, ethnicity, and EL status

Directions are provided on the use of the statistical information and review booklets, and an outline is given to each committee member describing the field test data they will review and use to determine the quality of each item. Committee members first evaluate each test item regarding the benchmark and instructional target match, appropriateness, DOK level, level of difficulty, and bias (cultural, ethnic, gender, geographic, and economic) before recommending that the item be accepted or rejected. Items that pass all stages of development—item review before field testing, field testing, and data review—are placed in an item bank and become eligible for use on future tests. Rejected items are noted and precluded from use on any future tests.

2.5.1. Statistics Used

2.5.1.1. Classical Test Theory Statistics

Several pieces of summary statistical information are provided to the data review committee. 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 MC distractors and the most common incorrect answers for TE and FIB items.

[4] MH statistical indices requires there be at least 100 students from the focal group and 200 students from the reference group to conduct a DIF analysis.

2.5.1.2. Item Response Theory Statistics

The IRT item parameters and fit indices are provided to the data review committee. IRT, more completely described in Chapter 6: Scaling, comprises a number of related models, including Rasch-model measurement (Masters, 1982; Wright, 1977), the two-parameter and three-parameter logistic (2PL, 3PL) models (Lord and Novick, 1968), and the generalized partial-credit (GPC) model (Muraki, 1992). The IRT model must fit student responses for the scaling and equating procedures to be valid. The item's relative difficulty (b-parameter), the item's capability of separating low performers from high performers (a-parameter), and the IRT guessing parameter (c-parameter) are provided to the committee. The IRT guessing parameter represents the probability of a correct response for the extremely low performers. The review committee uses these values to identify items that might be undesirable for inclusion in the item pool.

2.5.1.3. Differential Item Functioning Analyses

Differential item functioning (DIF) analyses (i.e., item bias data) are presented during data review committees using the MH statistic and its associated chi-square significance test. The MH statistic is a log-odds ratio that investigates whether the odds of answering an item correctly is greater for one demographic group than another after matching groups by their total test scores (Holland and Thayer, 1988). When one group is much more likely to answer a particular item correctly than another across the ability strata, the item is flagged for further examination. Even though every attempt is made to write unbiased items, Minnesota conducts DIF analyses on field test items for several subgroups to identify and evaluate items that are not functioning as expected. The MH test is conducted for all field test items for the MCA-III. Only the Science MTAS included the embedded field test items in spring 2022, so DIF analyses were conducted for those items. In 2023, DIF analyses were conducted for MCA-IV and Alternate MCA embedded field test items that appeared on the MCA Science and MTAS Science and Reading assessments, respectively.

Evaluating items for DIF provides an additional piece of evidence about whether the items on the statewide assessments are displaying construct-irrelevant factors. If items show DIF and are determined to be biased according to a committee, this would lessen the validity of the assessments for any particular group of individuals. The three broad categories of groups that are evaluated for DIF are gender, race/ethnicity, and EL status.

Table 2.1 presents the comparison groups for the Minnesota tests. The gender analysis investigates whether males have greater, the same, or lower odds of a correct response in relation to females, after matching males and females on total test score. Similarly, the race/ethnicity comparison between White and Black investigates whether white students have greater, the same, or lower odds of a correct response in relation to Black students, after matching white and Black students on total test score. Lastly, the EL analysis investigates whether non-EL students have greater, the same, or lower odds of a correct response in relation to EL students, after matching non-EL and EL students on total test score.

Table 2.1. DIF Comparison Groups

Group Type	Reference Group	Focal Group
Gender	Male	Female
Race/Ethnicity	White	American Indian
Race/Ethnicity	White	Asian
Race/Ethnicity	White	Black or African American
Race/Ethnicity	White	Native Hawaiian or Pacific Islander
Race/Ethnicity	White	Hispanic or Latino
Race/Ethnicity	White	Other Indigenous Peoples
English Learner (EL)	Non-EL	EL

The MH statistic used to flag DIF is based on the widely adopted Educational Testing System (ETS) procedure for that classifies DIF (Zieky, 1993) as either A (negligible or nonsignificant DIF), B (slight-to-moderate DIF), or C (moderate-to-large DIF). The data review cards only contain information for items flagged with C-DIF, although B-DIF is also indicated but does not have an explicit flag. Items that contain C-DIF are flagged for any potential bias by the data review committee and the item will be removed from the item bank prior to operational administration if a cause for such DIF is identified.

The MH procedure used by Minnesota requires that there be at least 100 students from the focal group and 200 students from the reference group to conduct a DIF analysis. The steps used to calculate the MH DIF statistic for dichotomous items are outlined below.

The total raw score for the fixed form (i.e., Science MCA and MTAS) or scale score for adaptive tests (i.e., Mathematics and Reading MCA) is calculated for each student. These scores are used to create 10 equally sized intervals, or strata. Table 2.2 presents an example $2x\ \ 2$ frequency table that matches the reference group and focal group based on their total raw/scale score, where *j* is the number of strata used for the analysis.

Table 2.2. MH Contingency Table for Dichotomous Items

Group	Correct	Incorrect	Total
Reference	A_j	B_j	n_{R_j}
Focal	C_j	D_j	n_{F_j}
Total	m_{1_j}	m_{0_j}	T_j

A_j is the number of reference students in stratum *j* who answered correctly.
B_j is the number of reference students in stratum *j* who answered incorrectly.
C_j is the number of focal students in stratum j who answered correctly.
D_j is the number of focal students in stratum *j* who answered incorrectly.
m_{1_j} is the total number of students in stratum *j* who answered correctly.
m_{0_j} is the total number of students in stratum *j* who answered incorrectly.
n_{R_j} is the total number of reference students in stratum *j*.
n_{F_j} is the total number of focal students in stratum *j*.
T_j is the total number of students in stratum *j*.

If no DIF is present, the odds ratio, calculated as $(A_j/B_j)/(C_j/D_j)$, would be equal to 1 for all strata. This would indicate that the odds of answering correctly would be the same for both the reference and focal groups. The actual MH test estimates a common odds ratio and tests whether it is significantly different from 1.0:

$$\alpha_{MH} = [\Sigma_j A_j D_j/T_j]/[\Sigma_j C_j B_j/T_j], \tag{2.1}$$

where all terms have been defined earlier. The test statistic (α_{MH}), as well as the lower ($\alpha_{MH,lower}$) and upper ($\alpha_{MH,upper}$) limits of the 95% confidence interval, are recorded.

To improve the ease of interpretation, the test statistic (α_{MH}), as well as the ($\alpha_{MH,lower}$) and upper ($\alpha_{MH,upper}$) limits of the 95% confidence interval, are transformed to the delta metric (*MH D-DIF*) through the following three formulas (Dorans and Holland, 1992; Holland and Thayer, 1985, 1988):

$$\textit{MH D-DIF} = -2.35 \ln \alpha_{MH}$$

$$\textit{MH D-DIF} = -2.35 \ln \alpha_{MH,lower}$$

$$\textit{MH D-DIF} = -2.35 \ln \alpha_{MH,upper} \tag{2.2–2.4}$$

A positive value for *MH D-DIF* indicates that an individual item may be differentially easier for the focal group, whereas a negative value indicates that the item may be differentially easier for the reference group.

For CR items that use polytomous scoring, standardization DIF (Dorans, 2013) is used in conjunction with the MH chi-square to identify items with DIF. The MH procedure used by Minnesota requires there be at least 100 students from the focal group and 200 students from the reference group to conduct a DIF analysis.

The total raw score for each student is calculated for fixed form tests and the scale score for adaptive tests. These scores are used to create 10 equally sized intervals, or strata, between the minimum and maximum total raw scores and scale scores, based on all students.

The standardization DIF formula compares the item means of two groups after adjusting for differences in the distribution of students across the values of the matching variable (i.e., total test score) and is calculated using the following formula:

$$STDDIF = \sum_j \left(\frac{n_{j,F}}{N_F}\right) \left[E(Y|Matching_j, F\right] - \left[E(Y|Matching_j, R\right]$$

where the average difference in expected and conditional scores of the studied item (*Y*) for reference (*G* = *R*) and focal (*G* = *F*) groups matched across the *j* = 1 to *J* possible score values of a matching variable, and $n_{j,F}$ and N_F denote the focal group's conditional and overall sample sizes.

A positive *STDDIF* value means that, conditional on the total test score, the focal group has a higher mean item score than the reference group. In contrast, a negative *STDDIF* value means that, conditional on the total test score, the focal group has a lower mean item score than the reference group.

Based on the transformation of the test statistic and the lower and upper limits of the confidence interval, DIF is classified into one of three categories as summarized in Table 2.3 and Table 2.4 (Dorans and Holland, 1992; Zwick, 2012).

Table 2.3. DIF Classification Categories: Dichotomous Flagging Rules

DIF Classification Category	Criteria
A-DIF (negligible)	$\lvert \boldsymbol{MH\ D - DIF} \rvert$ is not significantly different from zero, or is less than one.
B-DIF (slight to moderate)	$\lvert \boldsymbol{MH\ D - DIF} \rvert$ is significantly different than 0.0 (but not 1.0) and is at least 1.0. OR $\lvert \boldsymbol{MH\ D - DIF} \rvert$is significantly different than 1.0 but is less than 1.5.
C-DIF (moderate to large)	$\lvert \boldsymbol{MH\ D - DIF} \rvert$is significantly greater than 1.0 and is at least 1.5.

Table 2.4. DIF Classification Categories: Polytomous Flagging Rules

DIF Classification Category	Criteria
A-DIF (negligible)	MH Chi-square p-value ≥ 0.05 OR Effect Size ≤ 0.17
B-DIF (slight to moderate)	MH Chi-square p-value < 0.05 AND 0.17 < Effect Size ≤ 0.25
C-DIF (moderate to large)	MH Chi-square p-value < 0.05 AND Effect Size > 0.25

Summaries of the DIF results of the current administration of the Mathematics, Reading, and Science MCA and the Science MTAS are provided in the *Yearbook Tables for Minnesota's Statewide Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Yearbook Tables expandable header)) under the sections entitled "C-DIF Flag Summary Reports." These summaries provide the total number of items taken to data review, the number of items taken to data review flagged with C-DIF, the number with C-DIF in each subgroup category, and the number of items rejected by the data review committee where the presence of C-DIF contributed to the committee's decision to reject the item.

2.5.2. Data Review Meetings

2.5.2.1. Minnesota Comprehensive Assessments-Series III

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

2.5.2.2. Minnesota Test of Academic Skills

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

- Developmentally Cognitively Disabled—Mild
- Developmentally Cognitively Disabled—Severe
- Autism Spectrum Disorder

2.5.2.3. ACCESS for ELLs and WIDA Alternate ACCESS

Data review committees are convened by CAL and WIDA. These organizations follow industry standards when conducting data review committee meetings.

2.6. Item Bank

Minnesota's testing service provider maintains an item bank for all tests in the Minnesota assessment program and stores each test item and its accompanying multimedia assets in an item banking system. MDE also maintains paper copies of each test item.

In addition, Minnesota's testing service provider maintains a statistical item bank that stores item data, such as a unique item number, grade level, subject, benchmark or instructional target measured, DOK, dates the item has been field tested, and item statistics. The statistical item bank also warehouses information obtained during data review that indicates whether an item is acceptable for use, acceptable with reservations, or not acceptable at all. MDE and Minnesota's testing service provider use the item statistics during the test construction process (or a simulation study for the CAT assessments) to calculate and adjust for differential test difficulty and to check and adjust the test for content coverage and balance.

The move to CAT for the grades 3–8 and 11 Mathematics MCA-III and the grades 3–8 and 10 Reading MCA-III has required that a sizable item bank be maintained for each of these subjects and grades. All operational items within the item bank are available in the item pool for the mathematics assessments. The reading assessments contain predefined testlets in which each testlet contains one or more passages, and three testlets are administered for each test administration. Testlets are constructed each year using operational items from the item bank. The CAT engine relies on algorithms that select items or testlets from these banks.

2.7. Test Construction

MDE and Minnesota's testing service provider construct test forms from the pool of items or performance tasks deemed eligible for use by the data review committees. Minnesota's testing service provider uses operational and field test data to place the item parameters on a common IRT scale. (Refer to Chapter 6: Scaling). This scaling allows for the comparison of items, in terms of item parameters, to all other items in the pool. Hence, Minnesota's testing service provider selects items within a content benchmark not only to meet sound content and test construction practices but also to maintain comparable item parameters from year to year.

The fixed-form assessments include all MCA-III paper accommodated forms, the grades 5, 8, and high school Science MCA-III, and all MTAS-III assessments for mathematics, reading, and science. To construct these tests, MDE and Minnesota's testing service provider apply 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) for mathematics. The substrand is the main organizational element for reading (e.g., informational text or literature). Each strand for mathematics and substrand for reading contains one or more standards. Each standard contains one or more benchmarks. Each year's assessment assesses items in each strand and standard but not necessarily every benchmark. 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 assessed benchmark.

For the Mathematics and Reading MCA-III CATs, Minnesota's testing service provider does not directly construct the test forms. Rather, the testing service provider performs simulation studies to determine the best parameters for the CAT algorithm to administer the test. When the simulations are completed and approved, the CAT algorithms select mathematics items or reading testlets for the student to answer that will best measure their proficiency and satisfy the test specifications on the respective tests.

For the braille and large print accommodations administered as paper accommodated forms, MDE's goal is to keep all items on an operational form. Items are replaced if they cannot be placed into a braille translation or large print mode appropriately. To date, Minnesota has met this goal in all assessments since the program began in 1997.

Chapter 3: Test Administration

3.1. Eligibility for Assessments

As a result of ESSA, all public school students enrolled in grades 3–8 and at least once in grades 9–12 must be annually assessed with a mathematics and reading or language arts assessment, while students in grades 5, 8, and high school must also be annually assessed with a science assessment. This requirement includes students who receive special education services. ESSA and Minnesota Statutes, section 124D.59, require that all EL students be assessed in grades K–12 in English language proficiency.

Public school students take the Mathematics, Reading, and Science MCA-III to fulfill their requirement for each content area. Students with IEPs who meet the eligibility criteria of the MTAS-III, as defined in the *Eligibility Requirements for Minnesota Alternate Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Minnesota Tests > Alternate Assessment Participation), are eligible to participate in the MTAS-III assessments to fulfill their requirement for each content area. ELs in grades K–12 must participate in ACCESS or WIDA Alternate ACCESS. Most ELs take the ACCESS, but students who receive special education services and meet the participation criteria in the *WIDA Alternate ACCESS Participation Decision Tree* may instead take the WIDA Alternate ACCESS.

3.2. Administration to Students

3.2.1. Minnesota Comprehensive Assessments-Series III

3.2.1.1. Mathematics

The grades 3–8 and 11 Mathematics MCA-III are administered online, with paper test materials available for eligible students. The online assessment is divided into multiple groups. For grades 3–8 and 11, the paper test books are divided into four segments, allowing districts to administer the paper version over multiple days if they choose. For the grade 11 paper test books, students may use a calculator on the entire test, and handheld calculators may be used. For the grades 3–8 paper test books, calculators are allowed for segments two, three, and four. For students in grades 3–8 taking the online assessment, only the online calculator may be used when a calculator is allowed, while grade 11 students may use the online calculator or a handheld calculator for the entire assessment.

Districts have flexibility in how online MCA administrations are scheduled, as students are not required to exit the test at the same place as other students, and the online test has functionality that prevents students from going back to items completed in previous testing sessions. Additionally, for the online assessment, students are allowed to review and change their responses to items within their current group of items in the test, but they cannot go back to review their responses to items on previous groups of the test. Each district sets their testing schedule within the state-designated testing window.

3.2.1.2. Reading

The grades 3–8 and 10 Reading MCA-III is administered online, with paper test materials available for eligible students. The online assessment is divided into groups by testlet and each testlet's associated items. The paper test is divided into four segments, allowing districts to administer the paper version over multiple days if they choose.

Districts have flexibility in how online MCA administrations are scheduled, as students are not required to exit the test at the same place as other students, and the online test has functionality that prevents students from going back to items completed in previous testing sessions. Additionally, for the online assessment, students are allowed to review and change their responses to items within their current group of items in the test, but they cannot go back to review their responses to items on previous groups of the test. Each district sets their testing schedule within the state-designated testing window.

3.2.1.3. Science

The grades 5, 8, and high school Science MCA-III are administered online, with paper test materials available for eligible students. The online assessment is divided into multiple sections. The paper test books are divided into two segments, allowing districts to administer the paper version over multiple days, if they choose. In the Science MCA, items are associated with a scenario, and each scenario is made up of multiple parts. For paper test books, since items where a calculator may be used are not included on the science test each year, the item in the test book will indicate whether a calculator can be used. For the online assessment, the online calculator or a handheld calculator is available for items on the test that require simple mathematical computations. Beginning in 2022–23, any type of handheld calculator can be used instead of or alongside the online calculator.

Districts have flexibility in how online MCA administrations are scheduled, as students are not required to exit the test at the same place as other students, and the online test has functionality that prevents students from going back to items completed in previous testing sessions. Additionally, for the online assessment, students are allowed to review and change their responses to items within their current section of items in the test, but they cannot go back to review their responses to items on previous sections of the test. Each district sets their testing schedule within the state-designated testing window.

3.2.2. Minnesota Test of Academic Skills-Series III

Any district employee who has completed the applicable MTAS test administrator training may administer the MTAS, although 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 annually prior to each test administration. The MTAS is administered to students in a one-on-one setting and scored by the test administrator. Therefore, test administrators must schedule times to administer the tasks within the state-designated testing window.

3.2.2.1. Mathematics

The Mathematics MTAS includes object lists that provide guidance on the use of objects or manipulatives for students who need this type of support. Although the MTAS is administered in a one-on-one setting, the administration of the assessment is still considered standardized. The design of the assessment and its administration are specified in the *MTAS Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to students.

3.2.2.2. Reading

For the Reading MTAS, students may interact with the passage text in one of several presentations: the passage text, a picture-supported passage, or other accommodations appropriate for students' needs. A separate-passages book is also provided that includes materials used to present Alternate MCA field test tasks to students for reading.

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, educators identify what support, if any, students had with reading the passage. This passage support was used to create the alternate ALDs and determine achievement levels in summer 2013. This level of passage support is also reported on the student report presented to parents/guardians.

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

Although the Reading MCA-III does not allow for a read-aloud accommodation, the Reading MTAS-III is used to assess a very different population. Disallowing an MTAS-III read-aloud accommodation would make assessment difficult, particularly owing to the intended population that 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-III 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. (2009) suggests that this reading passage support is appropriate:

> For each of the five options under reading and math, educators were asked to select the option that best described their students' present performance in those areas. In States 1 and 3, educators noted that over 2 percent 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 percent of the students in State 1, 12 percent in State 2, and 33 percent 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 percent, 47 percent, and 33 percent 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 percent, 14 percent, and 18 percent) 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, educators noted that 15 percent of students in State 1, 25 percent of students in State 2, and 13 percent of students in State 3 had no observable awareness of print or braille (p. 245).

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

> Our results appear consistent with those of Almond and Bechard (2005), who also found a broad range of communication skills in the students in their study (i.e., 10 percent of the students in their sample did not use words to communicate, but almost 40 percent 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 MTAS-III passages read to them. In the journal *Remedial and Special Education*, Browder et al. (2009) propose a conceptual foundation for literacy instruction for students with significant cognitive disabilities. The conceptual foundation discussed includes accessing books through listening comprehension. As Browder et al. (2009) note, "To use literature that is grade and age appropriate, books will need to be adapted, including the use of text summaries and key vocabulary. Students who do not yet read independently will need either a technological or human reader" (p. 10).

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

3.2.2.3. Science

The Science MTAS includes object lists that provide guidance on the use of objects or manipulatives for students who need this type of support. A separate phenomena book is also provided that includes materials used to present Alternate MCA field test tasks to students for science. Although the MTAS is administered in a one-on-one setting, the administration of the assessment is still considered standardized. The design of the assessment and its administration are specified in the *MTAS Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to students.

3.2.3. ACCESS for ELLs and WIDA Alternate ACCESS

ACCESS assesses four language domains (Listening, Reading, Speaking, and Writing) and is available in six grade-level clusters: K, 1, 2–3, 4–5, 6–8, and 9–12. Paper accommodations are administered for the Listening, Reading and Writing domains in the following grade clusters: grade 1, grade 2, grade 3, grades 4–5, grades 6–8, and grades 9–12. Paper accommodations for the Speaking domain are administered in the following grade-level clusters: grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. While the kindergarten ACCESS is entirely paper based, ACCESS is primarily administered online, with paper test materials available only for eligible students. For the online grades 1–12 assessments, the Listening and Reading domains are adaptive; students must answer each item to continue and may not go back to review previous responses. For the Speaking domain, once students record a response, they cannot go back. For the Writing domain, students must enter a keystroke to continue but can go back to previous items during the administration.

WIDA Alternate ACCESS also assesses four domains (Listening, Reading, Speaking, and Writing) and is available in four grade-level clusters: K–2, 3–5, 6–8, and 9–12. It is an entirely paper-based assessment and is now available for eligible kindergarten students. The domains are individually administered, and all sections are scored by the test administrator.

3.3. Secure Testing Materials

The recovery of testing materials after each administration is critical. Secure test materials, including paper test materials, must be returned to preserve the security and confidential integrity of items that will be used on future tests. While secure test materials must be returned, others can be securely disposed of at the district.

Minnesota's testing service provider assigns secure paper test materials to school districts by unique barcoded security numbers. School districts use security checklists to assist Minnesota's testing service providers in determining whether secure materials are missing. Minnesota's testing service providers scan incoming barcodes 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 service providers contact any district with unreturned secure materials. MDE's internal security procedures are documented in Appendix B of the *Procedures Manual for Minnesota's Statewide Assessments*.

3.3.1. Minnesota Comprehensive Assessments-Series III

The Mathematics, Reading, and Science MCA-III are delivered online, with paper test materials available for eligible students. For online assessments, there are no secure materials to return to Minnesota's testing service provider, but districts are required to dispose of other secure materials, such as student scratch paper and testing tickets, no more than two business days after the close of the testing window. For students taking paper tests, secure materials include paper test books in 12-, 18-, and 24-point fonts for mathematics, reading, and science; braille test books; and mathematics and science scripts. Districts enter student responses online for scoring. All secure paper materials must be returned to Minnesota's testing service provider.

3.3.2. Minnesota Test of Academic Skills

The separate passages and/or phenomena book is provided for grades containing reading and/or science to field test Alternate MCA tasks. Secure test materials for the Mathematics, Reading, and Science MTAS-III include the Task Administration Manuals, Presentation Pages, Response Option Cards, and Passages and/or Phenomena Books for applicable grades. Following administration, all used and unused Task Administration Manuals, Presentation Pages, and Passages and/or Phenomena Books must be returned to Minnesota's testing service provider. All Response Option Cards must be securely destroyed at the district.

3.3.3. ACCESS for ELLs and WIDA Alternate ACCESS

Test materials for the English language proficiency accountability assessments include the following:

- Kindergarten ACCESS: Test Administrator Script, Student Storybook, Activity Board, Cards and Card Pouch, and Student Response Booklet
- WIDA Alternate ACCESS: Test Administrator Script, Test Booklet, and Student Response Booklet
- ACCESS Online Grades 1–12: Test Administrator Script, Grades 1–3 Writing Test Booklet, Grades 4–12 Writing Response Booklet, and paper accommodations as required

Districts return **all** secure test materials—used and unused—to DRC.

3.4. Universal Supports and Accommodations

New for 2023–24, MDE has reorganized how supports and accommodations are presented for statewide assessments. This new framework replaces the former categorization (general supports, linguistic supports, and accommodations). While there are now only two main categories, additional subcategories have been used to better organize and describe available supports. Linguistic supports have been recategorized as universal supports to reduce the complexity in determining appropriate supports for all learners. It also recognizes a more inclusive population of multilingual learners (MLs), as some universal supports may be appropriate for ELs, former ELs, and/or students who have participated in dual-language education programs.

The available supports and accommodations are documented in Chapter 4 of the *Procedures Manual for Minnesota's Statewide Assessments* that is updated annually and available on the PearsonAccess Next website (PearsonAccess Next > Resources & Training > Policies and Procedures).

Universal supports are general supports available for all students that tailor the testing experience based on student needs or preferences specific to the testing environment or online features that are allowable within standardized testing. The use of a universal support may replace the need for an accommodation, depending on the student's needs; universal supports may also be provided along with accommodations. While universal supports are available to all students, some require a specific code to be applied for students to access them; these codes are noted when required. Universal supports may be provided by the school or embedded within the test and are divided into the following categories:

- Tools include classroom materials that students use in instructional or other assessment settings and features available in online tests.
- Administrative considerations include changes to or personalization of test administration procedures or practices.
- Accessibility supports include features that may be appropriate for a specific student or groups of students (such as MLs) based on needs identified by an educator (or team of educators) familiar with the student's characteristics and needs.

Accommodations are changes in the way a test is administered and are meant to ensure equal access to the assessment. Accommodations are only available to students with an IEP or who qualify for a 504 plan. All needed accommodations should be documented annually in the IEP or 504 plan prior to testing. 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. For the MTAS, any accommodation listed on a student's IEP may be used as long as the support is not specifically prohibited in the task script. Administration activities allowed for the MTAS include the following:

- Familiarizing the student with the format of the MTAS prior to administration using the MTAS released items on PearsonAccess Next
- Adapting the materials presented to meet the student's 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

3.4.1. Research Base for Supports and Accommodations

In February 2013, the Smarter Balanced Assessment Consortium (SBAC) published a review of research related to supports and accommodations provided by the consortium on its accountability assessments. In their report, Jamal Abedi and Nancy Ewers of the University of California, Davis, shared results of a compilation of expert judgments and their literature review on the key questions of whether the use of an accommodation or support by SWDs and/or English language learners is effective and whether its use alters the focal construct of the assessment (Abedi and Ewers, 2013). MDE reviewed its allowed accommodations and supports for standards-based accountability assessments against Abedi and Ewers's findings, and a summary is provided in Table 3.1. School districts may contact MDE if an IEP or 504 team wants to use an accommodation that is not on the approved list. MDE will consider allowing that accommodation for the current administration and in future administrations pending literature and research reviews.

Table 3.1. Research Base for Supports and Accommodations

Support/Accommodation	Research and Recommendations
Assistive Technology The category of assistive technology includes devices that range from very commonplace supports to sophisticated technologies. Supports available to all students include materials commonly used during instruction such as pencil grips, place markers, line guides, color and masking overlays, highlighters, low-vision aids (e.g., magnifiers, large monitor screen sizes), whisper phones, and audio amplification devices.[5] Many of these supports are provided as tools in the online testing interface. Assistive technologies identified as accommodations for students with disabilities (SWDs) include talking calculators and devices such as computer tablets that serve as calculators or for note-taking. Generally, internet access must be disabled and students' computer use must be monitored. This accommodation generally requires an individual or small-group test administration.	According to Blaskey et al. (1990); Cormier et al. (2010); Iovino et al. (1996); Johnson, Kimball, Brown, and Anderson (2001); Robinson and Conway (1990); Salend (2009); and Scarpati et al. (2011): • Although most assistive technologies have not undergone experimental research, there is no evidence these accommodations unfairly advantage students. In addition, official studies confirm that the use of assistive technologies either greatly benefits or has little to no negative impact on students. Therefore, their use is supported. • In the case of audio amplification and magnifying equipment, all students benefit. • Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.

[5] MDE considers these supports as general supports but not assistive technology.

Support/Accommodation	Research and Recommendations
ASL and signed English interpretation **Test content** IEP teams may indicate sign language interpretation of the mathematics and science scripts (see human read-aloud) for students who are deaf or hard of hearing. Interpreters may access the script up to five business days prior to test administration and are required to review it to prevent cueing test answers. Interpreters may provide direct sign language interpretation from the text-to-speech as an accommodation, but approval by MDE is required. **Test directions** Sign language interpretation of the scripted Testing Directions may be provided to students who are deaf or hard of hearing.	According to Johnson, Kimball, and Brown (2001) and Russell et al. (2009): • The further research question from the studies relates to the capabilities and qualifications of on-site sign language interpreters, especially when interpreters are unfamiliar with the tested subject and its technical terms; the inability of interpreters to gain access to and prepare for the assessment prior to testing further complicates the issue. According to Russell et al. (2009): • The obstacles and limitations presented by televised recordings of a signed test may be overcome by computer programs. According to Johnson, Kimball, and Brown (2001): • It is difficult to assess if students gain an unfair advantage, as the signing of a test is "an accommodation of an accommodation." According to Ray (1982), Sullivan (1982), and Thurlow and Bolt (2001): • Experts agree that sign language interpretation of test directions, which is used in most states, levels the playing field for deaf and hearing-impaired students. Signed test directions give these students the same opportunity to participate in and score as well on the assessments as general education students. • Research supports the effectiveness of the accommodation and recommends its use, although there are concerns about its implementation. The risk of the accommodation giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Audio presentation of mathematics and science assessments **Text-to-speech** Minnesota provides two types of text-to-speech support for online math and science assessments. Text-to-speech and other read-aloud methods are not allowed on the reading assessments. General text-to-speech is available to all students who choose to use it. Only text in the stem and answer options is typically read aloud. Tables, graphs, labels, etc., generally are not read. Accommodated text-to-speech is available as an accommodation for SWDs. All text in stems, answer options, tables, charts, graphs, labels, etc., are read aloud and positional descriptions are provided, if appropriate. **Human read-aloud** Mathematics and science scripts are available as a read-aloud accommodation for SWDs. All text in stems, answer options, tables, charts, graphs, labels, etc. are read aloud and positional descriptions are provided, if appropriate.	According to Acosta et al. (2008); Barton (2002); Bolt and Thurlow (2004); Brown (2007); Burch (2002); Calhoon et al. (2000); Castellon-Wellington (2000); Christensen et al. (2011); Cormier et al. (2010); Dolan et al. (2005); Elbaum (2007); Fuchs et al. (2000); Helwig et al. (2002); Johnson, Kimball, Brown, and Anderson (2001); Kopriva et al. (2007); Pennock-Roman and Rivera (2011); Pennock-Roman and Rivera (2012); Sato et al. (2007); Tindal et al. (1998); and Wolf et al. (2009): • Collective research provides varied conclusions as to the effectiveness of this accommodation. Although results vary across grades, subjects, disability type, and level of proficiency in a subject or skill, the overall consensus confirms SWDs benefit from this accommodation. • Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.
Braille IEP teams may select contracted or uncontracted braille test booklets for students who are blind or partially sighted and are competent braille readers. Braille materials are provided in Unified English Braille format.	According to Bennett, Rock, and Kaplan (1987); Bennett, Rock, and Novatkoski (1989); Bolt and Thurlow (2004); Coleman (1990); Thurlow and Bolt (2001); and Thurlow et al. (2000): • Although braille tests require more time to complete and may make certain types of test items more difficult, research recommends the use of the accommodation. Most, but not all, states use braille tests. • Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Extended testing time Minnesota's statewide assessments are sectioned and untimed. Testing may be split over multiple days with one or more sections completed on a given day. Taking a single test section over multiple days or sessions is allowable as a general support for all students.	According to Crawford and Tindal (2004), DiCerbo et al. (2001), Fletcher (2009), Thurlow and Bolt (2001), and Walz et al. (2000): • Research is divided on whether extending testing time over multiple days is effective. Some studies revealed that SWDs in lower grades and students with low-level reading abilities benefited. In other studies, SWDs benefited little or did not benefit at all and general education students benefited. Experts recommend the support, which is used in most states, be used thoughtfully and carefully, only when absolutely needed. • Research supports the effectiveness of the support and recommends its use. The risk of the support giving students an unfair advantage is low.
Handheld calculator for mathematics Minnesota's online math tests have built-in calculators. SWDs in grades 3–8 who require a handheld calculator must take a paper test; students taking the paper math test may use a handheld calculator where allowable.	According to Bouck and Bouck (2008), Fuchs et al. (2000), Russell (2006), and Shaftel et al. (2006): • Calculators, which often are automatically included for math tests, are widely used by all students. Although research is divided on whether the accommodation provides a significant benefit to students, the use of the accommodation is strongly supported. • Research supports the effectiveness of the accommodation and recommends its use. There is no risk the accommodation gives students an unfair advantage.
Large print test book IEP or 504 plan teams may select 18- or 24-point font test booklets for students with low vision or for SWDs who need to take a paper test and a standard font test booklet is not available.	According to Beattie et al. (1983); Bennett, Rock, and Jirele (1987); Brown (2007); Burk (1998); Grise et al. (1982); Perez (1980); Thurlow and Bolt (2001); and Wright and Wendler (1994): • Much of the research concludes that large print tests, which are used in most states, offer little benefit. However, select studies strongly indicate that students with visual impairments and specific learning disabilities significantly benefit from this accommodation. • Research supports the effectiveness of the accommodation and recommends its use. There is no risk the accommodation gives students an unfair advantage.

Support/Accommodation	Research and Recommendations
Mathematics manipulatives; abacus SWDs in grades 3–8 who use manipulatives, hundreds or multiplication tables (provided by MDE), or an abacus for mathematics take the test using paper materials.	According to Elliott et al. (2009): • Experts are uncertain of the effectiveness and fairness of mathematics manipulatives but support the accommodation's use. • Despite uncertainties, research supports the use of the accommodation. The risk of the accommodation giving students an unfair advantage is moderate.
Recording a reading test Students may record themselves reading aloud the reading test and then play it back while they take the test. This accommodation is only available for SWDs beginning in 2023–24. While previously allowed for multilingual learners (MLs), it is not a common strategy used in EL instruction.	Research findings differ regarding SWD and EL. According to Crawford and Tindal (2004), Fletcher et al. (2006), McKevitt and Elliott (2003), and Meloy et al. (2000): • Studies involving SWDs and general education students reveal that the accommodation is effective in supporting all students, but especially SWDs. One study, however, indicated the accommodation may unfairly advantage some students. • Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low. According to Acosta et al. (2008) and Kopriva et al. (2007): • Experts are uncertain of the accommodation's effectiveness and conclude that reading a passage aloud gives ELs an unfair advantage. However, reading aloud only the test items may be suitable. • Despite uncertainties, research supports the use of the accommodation. The risk of the accommodation giving ELs an unfair advantage is high.
Scribe SWDs may dictate to a scribe who enters student responses into an online or paper test form. It is also possible for students to record their responses for later transcription by a scribe.	According to Thurlow and Bolt (2001): • Experts recommend that SWDs, including students who use ASL, submit answers via computer, whenever possible, rather than relay answers to a scribe. According to Fuchs et al. (2000); Koretz and Barton (2003, 2004); Koretz and Hamilton (2000); MacArthur and Graham (1987); and Tippets and Michaels (1997):

Support/Accommodation	Research and Recommendations
	• SWD research is limited, especially regarding the impact a disability has on test taking. A body of research suggests, however, that SWDs benefit from the use of scribes. Certain factors, such as type and difficulty of test and whether other accommodations are in place, should also be considered. • Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.

3.4.2. Accommodations Use Monitoring

Minnesota monitors the assignments of accommodations on its assessments. At a state level, data are reviewed for all accommodations for students who are (a) receiving special education or identified as disabled under Section 504 of the Rehabilitation Act of 1973 and (b) ELs.

3.4.3. Data Audit

The data collection is intended to provide MDE with the information about districts' use of accommodations on state assessments. This information allows MDE to analyze the accommodation data to draw conclusions about the 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 assessments scored without accommodations. MDE continually reviews these numbers both in overall percentage and in percentage expected in specific disability categories based on past data.

Chapter 4: Reports

Several reports are provided during and after each test administration. There are two types of reports: student-level reports and summary-level reports. Student-level reports and files contain individual student assessment scores with demographics and are provided during and after each test administration. Summary reports provide aggregated results at the school, district, and state levels and are provided upon release of the final results for each test administration. 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, a brief description of each report, guidelines for proper use of scores, and cautions about misuse.

As with any large-scale assessment, the statewide 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 about student achievement. Because all tests measure a finite set of skills with a limited set of item types, individual student decisions like course placement, promotion, and retention should be based on multiple sources of information, including (but not limited to) test results. However, for ACCESS, individual student scores are used to inform EL reclassification and exiting decisions.

Information about student performance is provided on student-level 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. Reporting resources and user guides for the MCA and MTAS are available on the Individual Student Reports (ISRs) Resources page (PearsonAccess Next > Reporting Resources > Individual Student Reports (ISRs) Resources) and the Additional Reporting Resources page (PearsonAccess Next > Reporting Resources > Additional Reporting Resources). WIDA provides scoring and reporting resources for ACCESS (wida.wisc.edu/assess/access/scores-reports) and WIDA Alternate ACCESS (wida.wisc.edu/assess/alt-access/scores-reports). MDE provides resources for all statewide assessments on the District Resources page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > District Resources).

4.1. Description of Scores

4.1.1. Test Codes and Score Codes

Score codes and test codes are used for reporting purposes. While test codes are typically used to document reasons why a student did *not* participate in the statewide assessment, score codes are also used to represent whether students answered enough items to receive valid scores. For the MCA-III and MTAS-III, test codes include the following:

- Absent (ABS)
- Invalidation due to Student Action (INV-S)
- Invalidation due to Device (INV-D)
- Invalidation Other (INV-O)
- Medical Excuse (ME)
- Not Enrolled (NE)
- Refused by Parent (REF-P)
- Refused by Student (REF-S)

For the MCA-III and MTAS-III, score codes include the following:

- Valid Score (VS)
- Not Attempted (NA)
- Not Complete (NC)
- Wrong Grade (WG)

For the MCA-III, the VS score code is given to students who responded to at least 90% of the operational items and field test items. Students who start a test but do not respond to any items receive a NA score code, while students responding to one or more items but fewer than 90% receive a NC score code. Starting in the 2021–22 test administration of MTAS-III, the VS score code represents a score entry for every operational task. Students with no entered scores receive an NA score code, while students with at least one entered score but missing one or more operational scores receive a NC score code.

The INV-S test code is indicated when a student engages in an inappropriate activity or behavior during or after testing that affects the validity of the test. The INV-D test code is indicated when a student accesses a cell phone, wearable technology, or other device at any point during a test administration (including breaks), or in circumstances where test security is violated using a device (such as capturing test content or looking up answers). The INV-O test code is indicated for misadministrations or other situations, including staff actions, which compromise the validity of a student's results.

The following test codes may be indicated when a student has not taken any portion of the test. The ME (medical excuse) test code can be used when a student has a medical excuse that meets specific criteria for not taking the test. The NE (not enrolled) test code is indicated when a student was not enrolled when the test was administered in the district. The REF-S (refusal – student) test code is used when a student refuses to participate in the assessment. The REF-P (refusal – parent/guardian) test code is used when a parent/guardian refuses to allow their student to participate.

Because ACCESS and WIDA Alternate ACCESS are designed for all states in the WIDA consortium, some of the Minnesota-specific test codes can only be indicated during Posttest Editing in the Test Web Edit System (Test WES) following test administration. Not Enrolled (NE) and Medical Excuse (ME) are Minnesota-specific test codes and can only be indicated in Posttest Editing. There are four test codes available in WIDA Assessment Management System (AMS) and on the test booklets during the testing window: Absent (ABS), Invalid (INV), Declined (DEC; used for both parent/guardian and student refusal), and Special Education Deferred (SPD). Regarding the valid score rules, refer to *Procedures Manual* in Chapter 6. For the test codes and their usage, refer to *Procedures Manual* in Chapter 9 for detailed information.

4.1.2. Types of Scores

Scores represent the result of students engaging in the testing process, providing information about student performance. Three different types of scores are used on the standards-based accountability assessments and English language proficiency accountability assessments: scale scores, raw scores, and achievement levels (MCA, MTAS) or proficiency levels (ACCESS, WIDA Alternate ACCESS).

4.1.2.1. Raw Score

Raw scores are only reported for Science MCA-III and all MTAS assessments. A raw score is the sum of points earned across all items on a subject area test. In addition to total raw scores, raw scores for items that constitute a specific strand or substrand (also called performance details) 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). The total and strand scores for the Mathematics and Reading MCA-III are computed using measurement model-based pattern scoring (i.e., scores depend on the pattern of correct/incorrect responses for the 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 the Mathematics and Reading MCA-III.

4.1.2.2. Scale Score

For the MCA and MTAS, scale scores are statistical conversions of raw scores or pattern scores that maintain a consistent metric across test forms. This allows comparisons to be made over time of scores within a particular grade and subject. Because scale scores adjust for different form difficulties, they can be used to assign an achievement level, which delineates whether a student met the achievement standards in a manner that is fair across forms and administrations. Schools can also use scale scores to compare the knowledge and skills of groups of students within a grade and subject across years. These comparisons can be used to assess the impact of changes or differences in instruction or curriculum.

The scale scores for a given MCA-III subject and grade range from G01 to G99, where G is the grade tested. For each MCA-III assessment, scale score G50 is the cut score for *Meets the Standards*, and G40 is the cut score for *Partially Meets the Standards*. The MCA-III cut score for *Exceeds the Standards* can vary by grade and subject. The scale score metric for each grade and subject is determined independently, so comparisons should not be made across grades or subjects. Scale scores for the Science MCA-III are transformations of raw number-correct scores. More than one raw score point may be assigned the same scale score, except at cut scores for each achievement level or at the maximum possible scale score. Pattern scoring is used to determine scale scores for the Mathematics and Reading MCA-III.

The range of observed scale scores for the MTAS-III varies from year to year. For each MTAS-III assessment, a scale score of 200 is the cut score for *Meets the Standards*, and 190 is the cut score for *Partially Meets the Standards*. The MTAS-III cut score for *Exceeds the Standards* can vary by grade and subject. As with the MCA-III, MTAS-III scale scores from different grades and subjects are not directly comparable.

The meaning of scale scores is tied to the grade-specific content and achievement levels associated with a given set of Minnesota Academic Standards. Thus, MCA-II scores are not directly comparable to MCA-III scores because those scores reflect different content and achievement standards. Similarly, when MTAS-III assessments change the academic standards to which they are aligned (concomitant with the MCA-III in the same grade and subject), the scores from assessments based on different academic standards are not directly comparable. Refer to Chapter 6: Scaling for details about how scale scores are computed.

For ACCESS and WIDA Alternate ACCESS, the composite scale scores and scale scores within a domain are on a vertical scale and do have meaning across grade levels, so each composite scale score or each domain scale score can be compared from year to year. However, the scale scores for each composite and each domain are independent, and comparisons across composites and domains cannot be made with scale scores.

4.1.2.3. Achievement/Proficiency Levels

To help families and schools interpret scale scores, achievement levels are reported. Achievement levels provide insight into how a student's test performance relates to grade-level standards. For the MCA and MTAS, the student's scale or raw score determines each achievement level, also sometimes referred to as performance or proficiency levels. The range for an achievement level is set during the standard setting process. The first year that a new series of assessments developed to align with the updated standards is implemented, panels of Minnesota educators set the cut scores for achievement levels and created the achievement level descriptors (ALDs), located online on the Success Criteria page of the Testing 1, 2, 3 website (Testing 1, 2, 3 > Plan and Teach > Success Criteria), that explain the general knowledge, skills, and abilities from the grade-level standards demonstrated by students across each level of achievement on the MCA and MTAS. For each test, certain achievement levels are designated as demonstrating proficiency of grade-level standards.

Table 4.1 presents a summary of the achievement levels for the Minnesota statewide assessments. For individual students, the achievement levels can be used to look at performance at various grade levels to gain a general sense of progress in a subject over time. Because the standards become more complex as grade levels increase, students who move up in achievement levels from one grade to the next are demonstrating progress. However, it is difficult to make similar claims for students who remain in the *Does Not Meet* achievement level or who move down in achievement levels between grades. To gauge student progress in such cases, it is necessary to compare test results on the statewide assessment to additional evidence of academic growth (or lack thereof) in a subject. Although the academic standards are aligned across grade levels, the content on the MCA/MTAS is grade-specific. It is difficult to make claims about whether students have retained knowledge from previous grades and are improving based on statewide assessments scores alone.

Proficiency level scores for ACCESS are presented as whole numbers followed by a decimal. The whole number indicates the student's language proficiency level based on the WIDA ELD Standards. The decimal indicates the proportion within the proficiency level range that the student's scale score represents, rounded to the nearest tenth. Proficiency level scores for the WIDA Alternate ACCESS range from 1 to 5 and are unique from other ACCESS assessments. The scores indicate the student's language proficiency level based on the WIDA Alternate Proficiency Level Descriptors. The WIDA Alternate Proficiency Level Descriptors are an extension of those found in the WIDA English Language Development (ELD) Standards, 2020 Edition, and describe the continuum of language development for ELs with the most significant cognitive disabilities.

Table 4.1. Achievement Levels for the Minnesota Statewide Assessments

Test	Achievement/Proficiency Level	Proficient
MCA-III, MTAS-III	*Does Not Meet the Standards*	No
	Partially Meets the Standards	No
	Meets the Standards	Yes
	Exceeds the Standards	Yes
ACCESS	*Level 1: Entering*	*
	Level 2: Emerging	*
	Level 3: Developing	*
	Level 4: Expanding	*
	Level 5: Bridging	*
	Level 6: Reaching	*
WIDA Alternate ACCESS	*Level 1: Entering*	**
	Level 2: Emerging	**
	Level 3: Developing	**
	Level 4: Expanding	**
	Level 5: Bridging	**

*Proficiency requires an overall composite ≥4.5 with at least three of the four domains ≥ 3.5.
**MDE is currently in the process of determining proficiency for this test.

4.2. Description of Reports

Assessment reports are either public or secure. Secure online reports of student and summary data are available once final results are available to authorized district personnel from Minnesota's testing service provider and MDE. The secure summary reports available to schools and districts are not for public release because all student data are reported. MDE also releases extensive summary data for public users (subject to filtering when cell sizes fall below 10 students) through the Minnesota Report Card and Assessment Files on the MDE website (MDE website > Data Center > Data Reports and Analytics).

Authorized district personnel can access additional secure assessment data through the online reporting systems of Minnesota's testing service providers, and this data can be aggregated at the district, school, teacher, and roster levels, as well as for individual students. The online reporting system for MCA and MTAS is PearsonAccess Next, which provides dynamic data that can be used to gauge students' achievement on the statewide assessments. For ACCESS and WIDA Alternate ACCESS, the online reporting system is WIDA AMS. However, the data and reports in these systems are not to be used for official accountability purposes. MDE provides official accountability data.

4.2.1. Student-Level Reports

Student-level reports provide information on each student's overall performance for each subject and a comparison of their performance relative to other students in the school, district, and state. For many assessments, including the MCA-III, these reports provide scale scores and achievement level designations associated with the student's performance. Performance within the strand or substrand level is also reported for each student. Table 4.2 presents the available student-level reports. Sample student reports can be found in MDE's *Interpretive Guide for Statewide Assessment Reports* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > District Resources (under the Test Score Interpretation Resources expandable header)).

Table 4.2. Student-Level Test Reports

File or Report	Format	Applies to MCA-III and MTAS-III?	Applies to ACCESS and WIDA Alternate ACCESS?
On-Demand Reports	Online (PDF)	Yes (within 60 minutes after testing or data entry is completed)	No
Early Student-Level Results	Online (excel file)	No	Yes (beginning of Posttest Editing in late May)
Individual Student Report (ISR)[6] and Roster	Paper and online PDF for ISR and online PDF for roster	Yes	Yes
Video Individual Student Report (ISR)	MP4 Video accessed via URL	Yes	No
Students Result Labels (Optional)	Paper and online PDF	Yes, paper by request	Not available
District Student Results (DSR) and School Student Results (SSR)	Excel file	Yes	Yes
Student Assessment History Report	Online	Yes	Yes
Historical Student Data	Online	Yes	No

[6] In addition to the electronic versions provided in the service provider systems, paper copies of the ISRs for distribution to families are shipped to districts. Districts may elect to receive only electronic versions or the PDF copies from published reports and then post these to their student information systems or other secure electronic submission.

Student-level reports provided in PearsonAccess Next for the MCA and MTAS include the following:

- On-Demand Reports for MCA/MTAS
- ISR and Video ISR
- Student Results Label (optional)
- Rosters
- Historical Student Data

Student-level reports provided in WIDA AMS for ACCESS and WIDA Alternate ACCESS include the following:

- District Student Response File
- Individual Student Reports
- Student Roster Reports
- Translated Student Reports
- Frequency Reports

Student-level reports available as MDE assessment secure reports include the following:

- Student Assessment History Report
- District and School Student Results

4.2.1.1. On-Demand Reports/Preliminary Results

On-Demand Reports provide a student's preliminary scale score and achievement level in PearsonAccess Next within 60 minutes after testing or data entry is completed. Information on strand or substrand by reporting category and Lexile®/Quantile® scores for reading/mathematics are also available within the report. Authorized users can sign in to PearsonAccess Next to view the student's score and download printable student reports. Teachers (i.e., users with the Test Monitor/Data Entry or MTAS Score Entry user role in PearsonAccess Next) can be assigned to reporting groups to view On-Demand Reports for students in the reporting group. Results from On-Demand Reports can be exported as a data file, downloaded as a list report, or generated as PDFs (Student Detail Reports) for individual students. The results in the Student Detail Reports (SDR) in On-Demand Reports are considered preliminary, subject to verification by MDE and Pearson through a process called adjudication. Once final results are released, the preliminary on-demand results become unavailable.

In late May, MDE provides early student-level results for ACCESS to allow districts to make decisions about instruction or placement. These results are available in an Excel file that can be downloaded from Test WES. Early results are not updated during Posttest Editing, and final results are provided following Posttest Editing in Assessment Secure Reports. For 2023–24, WIDA Alternate ACCESS scores were not included in early student-level results due to additional standard setting steps needed for the new assessment. MDE used a linking method to determine whether individual scale scores on the WIDA Alternate ACCESS made a student eligible for reclassification or exit. MDE contacted the districts with students who were eligible for reclassification and exit in July 2024.

4.2.1.2. Individual Student Reports

The Individual Student Report (ISR) is the official and final record of individual student results provided for student, parent/guardian, and teacher use. The ISRs for both MCA and MTAS begin with overall results, including a description and visuals for each subject representing student achievement level on the grade-level standards. Sample MCA and MTAS ISRs are posted under Reporting Resources on the Individual Student Reports (ISRs) Resources page (PearsonAccess Next > Reporting Resources > Individual Student Reports (ISRs) Resources).

Beginning in 2022–23, a personalized video ISR resource with individual student results for MCA and/or MTAS is available for each student with at least one valid score. Video ISR resources are accessed by scanning a QR code on the ISR or by URL from a Video ISR List downloaded from PearsonAccess Next. The video ISR resources include an overview of the assessment and of the student's performance in each subject taken, as well as their performance related to the school, district, and/or state averages. In addition to English, video ISRs may be available in a home language based on student enrollment data, if specified (Amharic, Arabic, Chinese, Hmong, Karen, Oromo, Russian, Somali, Spanish, and Vietnamese). These videos are intended to be a resource to support understanding of the ISR, but the ISR itself must still be provided.

An individual student's earned scale score for each subject is presented in a bar graph along with a description of the assigned achievement level next to the graphic. School, district, and/or state average scale scores are presented on the same graphic for comparison. The MCA-III ISRs provide all three averages, whereas the MTAS-III ISRs provide the state average only, appearing below the graphic. The number of students generally included in the average drives decisions about which average scale scores are reported for a given test. For example, the number of students included in the Reading MTAS-III school-level average scale score is small for most schools, which results in large standard errors of the mean. MDE's current privacy protection rules are to not publicly report the assessment results if there are fewer than 10 students represented in the data. For the MTAS-III, the number of students is frequently quite small for school or district populations.

Next, the MCA-III ISR includes descriptions of each strand/substrand for a subject and the strand achievement level for each reported as *Below Expectations*, *At or Near Expectations*, and *Above Expectations*. The MTAS-III ISR includes descriptions of each extended standard with points earned, points possible, and total points, also known as raw scores, ending with state mean raw scores for each extended standard. The testing service provider provides a Lexile score range for the Reading MCA-III and a Quantile score range for the Mathematics MCA-III. More specifically, the upper and lower range of the predicted Lexile measure is provided, which helps match the student with literature appropriate for their reading skills. The upper and lower range of the predicted Quantile measure is also provided, which helps match the student with mathematical concepts appropriate for their mathematics skills. These allow the student's parents/guardians to actively participate in their student's educational process.

The ISRs for the grades 3–8 Mathematics and Reading MCA-III and grades 5 and 8 Science MCA-III include a table with the student's performance history. The ISRs for the grade 11 Mathematics and grade 10 Reading MCA-III include a section describing the student's MCA-III score in context of the career and college readiness (CCR) goal score for that subject. The MCA-III score and CCR goal scores are on the same scale, and an MCA-III score at or above the goal score is considered on track to demonstrate career and college readiness in the corresponding subject on a college entrance exam at the end of grade 11.

The ISRs are provided as the final and official results to the district in up to two formats: if elected by the district, a paper copy to send home to parents, and all districts have access to a PDF version posted in PearsonAccess Next for school or district use. Authorized district personnel can also access the PDF version online in PearsonAccess Next. The ISRs provided to districts reflect final official accountability results for students.

Final reports for ACCESS and WIDA Alternate ACCESS—including paper copies of ISRs, rosters, and School and District Frequency Reports—are sent to the district.[7] The ISR shows both a proficiency level and a scale score for each of the four domains and provides a snapshot of how well the student understands and can produce the language needed to access academic content and succeed in school. Starting in 2023–24, educators can access a more detailed version of the ISR for students who were administered WIDA Alternate ACCESS. This report provides data designed to support decision-making related to instructional supports and reclassification.

4.2.1.3. Student Results Label

The purpose of the Student Results Label is to provide a compact form of individual student information to record in student files. Student Results Labels for the MCA and MTAS include the test name, test date, student information, scale scores, and achievement level for each subject tested for a single test. The individual student labels are stickers that can be attached to a student's permanent paper file, if the district maintains one. Printed labels are optional and must be requested by the district in Test WES during Pretest Editing (refer to *Chapter 7 of Procedures Manual*). Starting in 2022–23, all districts can access electronic PDFs of Student Results Labels under Published Reports in PearsonAccess Next.

4.2.1.4. Rosters

Student rosters are lists of students with final test code and individual performance data. They are posted to Published Reports in PearsonAccess Next and WIDA AMS at the time the paper ISRs reach districts.

4.2.1.5. Historical Student Data

Historical Student Data for the MCA and MTAS includes the assessment history for students who have previously tested at the district and for students who are currently enrolled in the district, regardless of where they tested. This report includes a student's achievement level, scale score, performance details by strand, and test details and is available in PearsonAccess Next. Student reporting groups can be created and assigned to teachers to provide them access to these data.

4.2.1.6. Student Assessment History Report

The Student Assessment History Report (MDE website > Data Center > Secure Reports) is a secure report that allows districts to access state assessment test history for students who are currently enrolled in the district, based on student enrollment data submitted to MDE.

[7] Larger districts with more than 1,000 EL students can have reports shipped to schools.

4.2.1.7. District and School Student Results

The secure District & School Student Results (DSR & SSR) files (MDE website > Data Center > Secure Reports) provide schools and districts with final student-level data on Minnesota's statewide assessments that can be sorted and analyzed to make data-driven decisions at the school and district levels in context with other information available at the local level. These files contain student-level data for the MCA, MTAS, ACCESS, and WIDA Alternate ACCESS, including demographic information, achievement level information, and test scores.

4.2.2. Summary-Level Reports

Summary reports provide information to schools and districts that may be used for evaluating programs, curriculum, and instruction of students. For example, districts may use the MCA-III school summary reports of test results by subject as one example of evidence to consider in evaluating how well their curriculum and instruction are aligned with the Minnesota Academic Standards. Summary reports are available online to authorized district personnel from MDE's Data Center, WIDA AMS, and PearsonAccess Next.

Secure summary-level reports for ACCESS and WIDA Alternate ACCESS provided in WIDA AMS include the following:

- District and School Frequency Reports for ACCESS and WIDA Alternate ACCESS

Secure summary-level reports for MCA and MTAS provided in PearsonAccess Next include the following:

- Longitudinal reports for MCA and MTAS
- Benchmark reports for MCA

Nonsecure Summary-level reports for MCA and MTAS provided in PearsonAccess Next include the following:

- Subscore reports for MCA

Secure online reports include a wide variety of reports summarizing test results at the student, school, district, and state levels and are used to provide information to authorized school and district educators and administrators.

Secure summary-level reports available as MDE assessment reports include the following:

- Alternate Assessment Participation for MTAS
- Test Results Summary for ACCESS, WIDA Alternate ACCESS, MCA, and MTAS

Nonsecure summary reports are also available from the Data Center and PearsonAccess Next. Although individual student scores are confidential by law, reports of group (aggregated) scores are considered public information and are available for general use. MDE's current privacy protection rules are to not publicly report the assessment results if there are fewer than 10 students represented in the data.

Nonsecure summary-level reports provided on MDE's website:

- Minnesota Report Card for ACCESS, MCA, and MTAS
- Assessment Files for ACCESS, WIDA Alternate ACCESS, MCA, and MTAS

4.2.2.1. District and School Frequency Reports

High-level reports for a single grade within a school, district, or state on the number and percentage of tested students who achieved each proficiency level for each language domain and composite area.

4.2.2.2. Longitudinal Reports

Longitudinal reports for MCA and MTAS are available for authorized users in PearsonAccess Next. The longitudinal system allows users to disaggregate data by subject, grade, and specific demographics. Reports are generated for score and achievement level using aggregation criteria selected by the user. Comparison reports are also available to compare state/district/school strand performance and achievement levels from year to year. There is also an option to export longitudinal results in a PDF or Excel format.

4.2.2.3. Benchmark Reports

Benchmark reports compare MCA school- or district-level aggregate observed performance on items or benchmarks from each benchmark with expected performance, given overall student scores. A few weeks after ISRs are released, benchmark reports are available as PDFs to school- or district-level users in Published Reports (PearsonAccess Next > PearsonAccess Next). Refer to Appendix A: Benchmark Report Calculations Resource of this manual for a detailed description of the procedure to calculate the benchmark reports.

4.2.2.4. Subscore Reports

The Subscore Report (PearsonAccess Next > Reporting Resources > Subscore Report) provides the public access to school-, district-, and state-level subscore data, also known as strand/substrand achievement levels, on the MCA. The strand achievement level is determined by comparing the school (or district) performance to the state expectation at the *Meets* achievement level. The strand achievement levels are reported as *Below Expectations*, *At or Near Expectations*, and *Above Expectations*. For each grade and subject, this report includes the percentage of students in each strand achievement level for each strand that is calculated by aggregating the individual student strand achievement levels at the school, district, and state level. The functionality of subscore reports allows users to generate charts and graphs by student groups that may be used for school and district instructional decision-making and planning at the subscore or strand/substrand level.

4.2.2.5. Alternate Assessment Participation

The Alternate Assessment Participation report (MDE website > Data Center > Secure Reports) provides the district's MTAS participation rates over the last five years and includes comparison data to similar districts and statewide. It is used to assist districts in completing the Assurance, Rationale and Context (ARC) on an annual basis. Refer to the Alternate Assessment Participation page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Minnesota Tests > Alternate Assessment Participation) for more information.

4.2.2.6. Test Results Summary

The secure Test Results Summary reports (MDE website > Data Center > Secure Reports) provide schools and districts final summary data, including state summary data, for the standards-based and English language proficiency (ELP) accountability assessments.

4.2.2.7. Minnesota Report Card

The Minnesota Report Card (MDE website > Data Center > Secure Reports) allows the public to see how various groups of students across the state and within districts and schools have performed on various tests and subjects over the years.

4.2.2.8. Assessment Files

Assessment Files (MDE website > Data Center > Data Reports and Analytics) are downloadable data files that provide public summary assessment data for the state, county, districts, and schools that can be used to perform analyses.

4.3. Appropriate Assessment Results Uses

The Minnesota statewide assessments are summative assessments administered at the end of a school year, designed primarily to determine school and district accountability related to equitable implementation of the academic standards (i.e., results are designed to be used as a "system check" at a student group, school, district, and/or state level). They are also criterion-referenced, meaning they measure performance against a fixed set of criteria (i.e., the Minnesota Academic Standards or WIDA ELD Standards). The assessments provide a snapshot of a student's overall achievement, not a detailed accounting of the student's understanding of specific content areas defined by the standards. They should be considered in the context of each district's balanced and comprehensive assessment system. While data on statewide assessment results provides a useful starting point, the most robust evaluations of district and school performance, and the most useful findings for maintaining and improving that performance, occur when this information is paired with information from district, local, and classroom assessments. Resources to guide educators on appropriate and inappropriate uses of MCA and ACCESS results are available on the District Resources page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > District Resources).

Test results from the statewide assessments, when used appropriately, can provide a basis for making valid inferences about student performance. For example, student test results can be used to report results to parents of individual students. The information can help parents begin to understand their child's academic performance as related to the Minnesota Academic Standards. However, the statewide assessments are not designed to provide the same fine-grained information about student learning that classroom assessments can give. Classroom assessments provide the specific information teachers need to adjust and differentiate instruction for their students.

The following Minnesota assessment results and classroom assessment results can be used to suggest areas needing further evaluation of student performance in the classroom. Examining changes in these assessment outcomes can lead to important considerations for district-, school-, and classroom-level decision-making, including evaluation of the overall alignment of course curriculum and district or classroom assessments with the standards being measured. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the Minnesota Academic Standards or WIDA ELD Standards.

- Overall achievement/proficiency levels over time between student groups
- Percent of students by achievement/proficiency level within a subject or grade level
- Change in students moving between achievement/proficiency levels over time
- Evaluating programs, resources, and staffing patterns

Test results can also be a valuable tool for evaluating programs. For example:

- A school may use its scores to help evaluate a particular academic program or curriculum in their school or district as it relates to the Minnesota Academic Standards or WIDA ELD Standards.
- Districts can use summary assessment data, as well as accountability data provided by MDE, to look at overall performance for a given year and over time.

4.3.1. Individual Students

Individual student results show a broad overview of student learning of grade-level standards and are intended to be interpreted alongside more fine-grained information about the individual student from the classroom teacher. However, federal and state legislation requires that individual student results be reported for all statewide assessments. It is therefore important that parents/guardians are given information and guidance on the appropriate use of these results. ACCESS and WIDA Alternate ACCESS are also summative assessments, but the results are designed to measure an individual student's progress toward English language proficiency. These results can be used at the individual level as well as to evaluate EL programs at a school or district level.

Subscores provide information about student performance in more narrowly defined academic content areas. When an area of possible weakness has been identified, supplementary data should be collected to further define the student's instructional needs.

Finally, districts and families should consider the limitations of analyzing individual results as they only provide information from one point in time. Individual student test scores must be used in conjunction with other performance indicators to assist in making educational decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.

4.3.2. 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. Percent proficient 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 percent proficient for students taking the grade 8 Mathematics MCA-III in 2018 can be compared to any of the 2011–17 grade 8 MCA-III administrations. However, whenever drawing inferences from such comparisons, 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. Below are the changes that should be accounted for while making longitudinal comparisons:

- New testing programs cannot be compared to previous testing programs that assessed different academic standards. For example, MCA-III results cannot be directly compared to previous MCA-II administrations because the MCA-III assesses different academic standards. The same holds true for the grades 3–8 Mathematics MTAS-III, which beginning in 2011 has assessed new standards and thus cannot be directly compared to the grades 3–8 Mathematics MTAS-III from prior years. This is also relevant for the Science MTAS-III that was constructed to align to new academic standards in 2012. In 2013, all Reading assessments (MCA-III, MCA-Modified, and MTAS-III) were revised to align to the *2010 Minnesota K–12 Academic Standards in ELA*. Thus, 2013 results are not directly comparable to those of prior years' versions of those tests. In 2014, the grade 11 Mathematics assessments (MCA-III, MCA-Modified, and MTAS-III) were revised to align to the *2007 Minnesota K–12 Academic Standards in Mathematics*.

Table 4.3. Comparing MCA and MTAS Assessment Results from Year to Year

Test	Grades	Year Standards Last Revised	First Year Assessed on Revised Standards	Years Scores Are Comparable
Mathematics MCA and MTAS	3–8	2007	2011	2011 to present
Mathematics MCA and MTAS	11	2007	2014	2014 to present
Science MCA and MTAS	5, 8, HS	2009	2012	2012 to present
Reading MCA and MTAS	3–8, 10	2010	2013	2013 to present

- When individual student graduation stakes associated with high school MCA tests changed in 2013–14, students no longer needed to achieve proficiency to meet graduation assessment requirements. Consideration should be given to the extent to which performance changes are attributed to content mastery versus motivation.
- The 2011 administration saw the introduction of the MCA-Modified exams, which meant that some students who otherwise would have taken the MCA-III now could take MCA-Modified instead. Most of those students returned to the MCA-III in 2015, when the MCA-Modified was discontinued. Consequently, when making comparisons with past administrations it is important to consider that the population taking the MCA-III changed in 2015. Consideration must also be given to changes in test administration policies when interpreting year-to-year changes in test results.
- For Mathematics MCA-III, students taking the test in online mode in 2012 were allowed up to three administrations of the assessment and could use the highest score for accountability purposes. By contrast, in 2011, 2013, and subsequent years, students were allowed only a single Mathematics MCA-III testing opportunity.
- Due to COVID-19, there was limited data for the 2020 administration and no summary data was provided for any public or secure reports. Due to the unknown impact that the COVID-19 pandemic might have on test participation or performance, there are important considerations for using available data from the 2021 administration.

ACCESS and Alternate ACCESS were first administered in 2012 and 2013, respectively, and the test design for grades 1–12 ACCESS changed significantly in the 2015–16 school year. As a result, student results for grades 1–12 ACCESS should only be compared from 2017 to present. The 2023–24 year was the first administration of the WIDA Alternate ACCESS. Results from WIDA Alternate ACCESS should not be compared to results from Alternate ACCESS for ELLs. Table 4.4 summarizes this information.

Table 4.4. Comparing ACCESS, Alternate ACCESS, and WIDA Alternate ACCESS Assessment Results from Year to Year

Test	Grades	First Year Administered	Year Test Design Changed	Years Scores Are Comparable
ACCESS	K	2012	N/A	2012 to present
	1–12	2012	2016	2017 to present
Alternate ACCESS for ELLs	1–12	2013	2024	2013 to 2023
WIDA Alternate ACCESS	K-12	2024	N/A	2024 to present

The percentage of students in each achievement level can also be compared across administrations within the same grade, subject area, and test to provide insight into whether student performance is improving across years. For example, the percentage of students in each achievement level for the grade 8 Mathematics MCA-III in 2018 can be compared to the population of grade 8 students from 2011 to 2017, 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 *Meets the Standards* and *Exceeds the Standards* would be expected to increase.

Such year-to-year comparisons were used to show that the school or district was moving toward the previous NCLB goal of having 100% of students proficient by 2014, although Minnesota's ESEA flexibility request waived this 2014 requirement of 100% proficiency for two years. Starting with the 2018 administration, however, the ESSA regulations have applied. 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 Academic Standards. Other test scores should be used to help evaluate academic areas of relative strength or weakness. Average performance on a strand or substrand can help identify areas where further diagnosis may be warranted for a group of students.

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

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

4.4. Cautions for Score Use

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

4.4.1. Understanding Measurement Error

When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error (i.e., test scores are not infallible measures of student performance). This effect is partly due 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. For the fixed-form tests (Science MCA-III, data-entry forms of MCA-III, and MTAS-III), some score variations would be expected if the same student tested across separate occasions using equivalent forms of the test. Although all testing programs in Minnesota conduct a careful equating process (described in Chapter 7: Equating and Linking) to ensure that test scores from different forms can be compared, one form may result in a higher score for a particular student than another form. Similarly, measurement error is present for the MCA-III Mathematics and Reading tests, which are CAT-based; however, because there are no fixed forms for these assessments, the measurement error for a given student depends also on the individual items the student is administered during the assessment. Therefore, two students with the same number of items correct will likely have different scale scores and measurement errors because they are highly unlikely to have received the same items on the CAT assessment.

Similarly, starting with the 2016–17 testing year, measurement error is factored in when determining the strand/substrand achievement levels. Refer to Chapter 6: Scaling for details on the strand/substrand achievement levels.

Because measurement error tends to behave in a random fashion when aggregating over a group of students these errors in the measurement of students tend to cancel out. Therefore, the average test score for a group of students will tend to be more accurate (less measurement error) than the score for any given student. While useful inferences can be made about an individual student from the student's test score, stronger inferences can be made about a group of students based on the group's aggregated test scores. Chapter 9: Reliability describes measures that provide evidence indicating that measurement error on Minnesota statewide assessments is within a tolerable range. Nevertheless, measurement error must always be considered when making score interpretations.

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

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

4.4.3. Interpreting Score Means and Variability in Performance

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. Investigating the sources of variance in test scores can provide additional inferences drawn from test scores. For example, very different scale-score distributions (different variabilities in students' performance) 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 most students received scores lower than the cut score, while a small number of students received 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. If investigating scale-score distributions and variances for the two examples described, the variances in the first example will be higher than the second example. Therefore, the scale score mean, percentage at or above a particular scale cut score, and the variance of students' scale scores in a group should be explored and interpreted when comparing results from one administration to another.

4.4.4. Using Strand- or Substrand-Level Information

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

In addition, because the tests are equated only at the total subject-area test scale score level, year-to-year comparisons of strand- and/or substrand-level performance should be made cautiously. Significant effort is made to approximate the overall difficulty of the strands or substrands from year to year during the test construction process, but fluctuations in difficulty can occur across administrations. Observing trends in strand- and/or substrand-level performance over time, identifying patterns of performance in clusters of benchmarks testing similar skills, and comparing school or district performance to district or state performance are more appropriate uses of group strand/substrand information. For example, observing trends in the proportions of students at the three different strand achievement levels (*Below Expectations*, *At or Near Expectations*, and *Above Expectations*) at the district or school level may be a more appropriate way to evaluate performance each year and over time.

4.4.5. 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 noted in Standard 13.9 in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), “In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation” (p. 213). 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. When changes in test content occur, development of new performance standards may be required. Test scores do not solely imply student competence. Rather, the interpretation of test scores permits inferences about student competence. 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 must demonstrate to be classified into the defined achievement levels.

The MCA-III and MTAS-III have four achievement levels: *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, and *Exceeds the Standards*. ACCESS has six proficiency levels that range from one (*Entering*) to six (*Reaching*). WIDA Alternate ACCESS has five proficiency levels that range from one (*Entering*) to five (*Bridging*).

Table 5.1 presents information regarding the most recent standard setting meetings for the Minnesota statewide assessments. Standard setting for each assessment was performed in accordance with specific Minnesota standards as indicated in Table 5.1 (i.e., either the Minnesota Academic Standards in Mathematics, ELA, or Science for MCA-III or the Minnesota Alternate Academic Achievement Standards in Mathematics, ELA, or Science for MTAS-III). This chapter presents an overview of the process for establishing the achievement levels for these tests. More detailed explanations of the standard setting activities can be found in the standard setting reports located on the Technical Reports page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Standard Setting Reports expandable header)).

Table 5.1. Standard Setting Meetings

Test	Grades	Date	Location	Method	Standards
Mathematics MCA-III	11	June 18–19, 2014	Saint Paul, Minn.	Bookmark	2007
Mathematics MTAS-III	11	June 18–19, 2014	Saint Paul, Minn.	ID Matching	2007
Reading MCA-III	3–8, 10	June 24–26, 2013	Roseville, Minn.	Bookmark	2010
Reading MTAS-III	3–8	June 27–28, 2013	Roseville, Minn.	ID Matching	2010
Science MCA-III	5, 8, high school	June 25–26, 2012	Roseville, Minn.	Bookmark	2010
Science MTAS-III	5, 8, high school	June 27–28, 2012	Roseville, Minn.	ID Matching	2010
Mathematics MCA-III	3–8	June 27–29, 2011	Roseville, Minn.	Bookmark	2010
Mathematics MTAS-III	3–8	June 29–30, 2011	Roseville, Minn.	Modified Angoff	2007

5.1. Process Components

Minnesota's testing service provider, MDE, and MDE's National TAC worked together to design the standard setting activities to follow the same general procedures as the standard setting meetings for mathematics, reading, and science for MCA-III and MTAS-III. Minnesota's testing service provider facilitated the standard setting under the supervision of MDE.

The Bookmark standard setting procedure (Lewis et al., 1996) was chosen for the MCA-III assessments, and the ID Matching method (Ferrara et al., 2002) was used for most MTAS-III assessments. Although similar to the widely implemented Bookmark method, the ID Matching procedure asks panelists to indicate which of the ALDs is best matched by the knowledge and skill requirements necessary to respond successfully to each test item. Modified Angoff, a test-centered standard setting method (Jaeger, 1989), along with some features of the Reasoned Judgment method (Kingston et al., 2001), was used for the grades 3–8 Mathematics MTAS-III assessments.

5.1.1. Selecting a Method

There are a variety of standard setting methods, all of which require the judgment of education experts and possibly other stakeholders. The key differences among the various standard 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 students or groups of students. Cut scores are set based on student exemplars of different levels of competency.

Another useful dichotomy is *compensatory* versus *conjunctive* (Hambleton and Plake, 1997). Compensatory methods allow students 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 standard setting methods perform best under specific conditions and with certain item types. For example, the Modified Angoff method is often favored with selected-response items (Cizek, 2001; Hambleton and Plake, 1997), whereas the policy-capturing method was designed for complex performance assessments (Jaeger, 1995). Empirical research has repeatedly shown that different methods do not produce identical results, and many measurement experts no longer believe "true" cut scores exist (Zieky, 2001). Therefore, it is crucial that the method chosen meets the needs of the testing program and that subsequent standard 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, test developers often design the process and adapt a method to meet their specific needs.

Different methodologies also 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 people 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.

5.1.2. Panelist Selection and Training

Panelists should be subject-matter experts, understand the student 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 and Reid, 2001). It may be useful to aim for the collective panel 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 and implementing 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.

During the standard setting workshops, MDE convened separate educator panels to recommend performance standards for each assessment. Each panel was organized by grade or grade band (e.g., 3–4, 5–6, 7–8, and 10 for Reading). Each grade band had a lower grade and an upper grade for which panelists set standards. Each panel/subpanel had its own facilitator and was physically separate from the other panels. MDE invited approximately 10–30 participants from across Minnesota for each panel to set cut scores for each assessment. Details of the credentials and demographics of the participants can be found in the standard setting reports located on the Technical Reports page (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Standard Setting Reports expandable header)).

5.1.3. Table Leaders

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

5.1.4. Ordered Item Booklets

Central to both the Bookmark and ID Matching procedures is the production of an ordered item booklet (OIB) that lists items from easiest to most difficult in terms of their theta difficulty. While the OIB is often produced from only the items in the first operational test, it is rarely the case that a single operational test administration provides a comprehensive sampling of items across the range of content standards and difficulty. While recommending standards on the entire item bank is ideal in some respects, including too many items makes review of the OIB overly burdensome.

5.1.5. Feedback

Some methodologies provide student performance data to panelists for decision-making, whereas other types of feedback include consequential (impact data), rater location (panelist comparisons), process feedback, and hybrid feedback (Reckase, 2001). Experts do not agree on the amount or timing of feedback, but any feedback can influence the panelists' ratings. Reckase (2001) suggests that feedback be spread out over rounds to have an effect on the panelists. Care should be taken not to use feedback to pressure panelists into making decisions.

5.2. Standard Setting Process

Before beginning the standard setting activities, MDE and Minnesota's testing service provider 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 policymaking process of standard setting, and their final cut score recommendations might not be the final cut scores adopted by the commissioner of education. Panelists were given an overview of standard setting and were introduced to the standard setting procedure they would be using. Panelists then convened into their assigned panel group to begin the process of setting standards.

The MCA-III standard setting meetings were conducted in three rounds for mathematics and reading and two rounds for science. Panelists used the ALDs to help them generate threshold descriptors as a group. After creating the threshold descriptors and completing the standard setting training and practice activities, the committee began the process of setting standards. After the Round 1 cuts were made for mathematics and reading, psychometricians evaluated results and produced feedback forms for each table and the room as a whole. The feedback forms for each table contained summary statistics showing the median, lowest, and highest cut scores for that table, as well as all individual bookmark placements. The room feedback form contained summary statistics showing the median, lowest, and highest cut scores for each table. After completing discussions on the Round 1 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards*, followed by *Partially Meets* and *Exceeds*. After Round 2, in addition to the room form, an impact data sheet containing OIB pages and the percentage of students at or above the level for each possible cut score was provided to the facilitator for reference and discussion. After completing discussions on the Round 2 feedback, panelists again worked through the OIB, placing their cut scores for *Meets the Standards*, followed by *Partially Meets*, and *Exceeds*. The process was similar for science, except that the impact data sheet was provided following Round 1, and there was no Round 3.

The MTAS-III standard setting meeting was also conducted in a series of three or two rounds but instead used the ID Matching method (except for the grades 3–8 Mathematics MTAS-III, which used Modified Angoff in the first two rounds and Reasoned Judgment in the third round). Panelists began the standard setting process by identifying the threshold region between *Partially Meets the Standards* and *Meets the Standards*. This entailed indicating the first item in the OIB that clearly matched the *Meets the Standards* ALDs and the last page that clearly matched the *Partially Meets* ALDs. The pages in between defined the threshold region in which panelists placed their cut scores. After identifying the threshold region, panelists examined each item in the threshold region to determine the first item that more closely matches the ALDs for *Meets the Standards* than the ALDs for *Partially Meets*. Panelists marked that item as their cut score. Panelists were instructed to use the same process to determine the threshold region and cut scores for *Partially Meets* and *Exceeds*. The same feedback was given to the MTAS-III participants as was given to the MCA-III panelists.

For the grade-banded panels, Rounds 1 and 2 recommendations were first completed for the lower grade followed by Rounds 1 and 2 for the upper grade. Round 3 recommendations were made for both grades concurrently after the review of Round 2 impact data across grades.

5.2.1. Round 1

After completion of the practice activities, panelists were provided with the OIB (or task book) associated with their panel. 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 (or task book), panelists began their independent review of the items. Specifically, panelists were instructed to do the following:

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

MTAS-III panelists were also asked to identify the threshold region between *Partially Meets the Standards* and *Meets the Standards*.

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

5.2.2. Round 2

During Round 2, panelists discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 1 recommendations, including the minimum, maximum, mean, and median recommendation associated with each level. Each table was instructed to discuss their Round 1 recommendations with the goal of identifying major sources of variation among panelists. Understanding, rather than consensus, was the goal of the discussion. After the discussion, panelists again placed their bookmarks or identified their cut scores. Panelists were reminded that this was an individual activity.

5.2.3. Round 3

During Round 3 (mathematics and reading only), participants discussed their recommendations in small groups at their tables. Panelists were provided with table-level feedback on their Round 2 recommendations and impact data that were given to the facilitator. Each table discussed their Round 2 recommendations with the goal of identifying major sources of variation among panelists. Understanding, rather than consensus, was the goal of the discussion. After the discussion, panelists placed their final bookmarks or identified their final cut scores. Panelists were reminded that this was an individual activity. Table 5.2 summarizes the feedback by round.

Table 5.2. Summary of Feedback by Round

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

5.3. Standard Setting for Grade 11 Mathematics MCA-III and MTAS-III

Standard setting for the grade 11 Mathematics MCA-III and MTAS-III took place on June 18–19, 2014, in Saint Paul, Minn., using the Bookmark standard setting procedure (Lewis et al., 1996) for MCA-III and the ID Matching procedure (Ferrara et al., 2002) for MTAS-III. MDE invited approximately 12 participants from across Minnesota to set cut scores for each assessment. The standard setting activities and results are recorded in *Minnesota Assessments Summer 2014 Standard Setting: Recommended Performance Standards for Series-III Mathematics Assessment*s (MDE, 2014).

For the Mathematics MCA-III OIB, operational items common to the 2014 online and paper accommodated test form administration modes served as the base. This OIB was augmented with 21 additional operational items selected from other operational forms. These additional items were selected to complement the content distribution of the selected operational form, in terms of standards and benchmarks assessed and the item types, and to fill item difficulty gaps in the OIB. This led to an OIB that included 77 items for the Mathematics MCA-III. For the MTAS-III, Minnesota's testing service provider produced an OIB using both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS-III item bank.

5.3.1. Recommended Cut Scores

Table 5.3 presents the participant-recommended cut scores for the grade 11 Mathematics MCA-III and MTAS-III after Round 3. Cut scores are shown on the theta metric. For the MTAS-III, final cut scores were identified by selecting the observed theta score nearest to the theta value associated with the panelists' recommended page number in the OIB. The nearest observed theta in the operational test form raw-score-to-theta table is the final recommended cut. Table 5.4 presents the impact data and the percentage of students in each of the four performance categories based on the cut scores after Round 3.

Table 5.3. Participant-Recommended Cut Scores for Mathematics Grade 11

Workshop	Content Area	Grade	Cut Scores (Theta Metric): *Partially Meets*	Cut Scores (Theta Metric): *Meets*	Cut Scores (Theta Metric): *Exceeds*
MCA-III	Mathematics	11	-0.5371	0.1034	0.9989
MTAS-III	Mathematics	11	1.0260	1.6731	2.8329

Table 5.4. Impact Data Associated with Participant-Recommended Cut Scores for Mathematics Grade 11

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	Mathematics	11	28	22	31	19
MTAS-III	Mathematics	11	30	21	39	10

5.3.2. Commissioner-Approved Results

After the standard setting meeting, the Minnesota commissioner of education reviewed the recommended cut scores for overall consistency and continuity. The commissioner for the 2014 MCA-III administration approved all the panelist-recommended cut scores.

5.4. Standard Setting for Grades 3–8 and 10 Reading MCA-III and MTAS-III

Standard setting for the grades 3–8 and 10 Reading MCA-III took place on June 24–26, 2013, in Roseville, Minn., using the Bookmark standard setting procedure (Lewis et al., 1996). Standard setting for the Reading MTAS-III took place on June 27–28, 2013, in Roseville, Minn., using the ID Matching procedure (Ferrara et al., 2002). Panels were organized by grade band (3–4, 5–6, 7–8, and 10), with approximately 10 panelists per panel. The standard setting activities and results are recorded in *Minnesota Assessments Summer 2013 Standard Setting: Recommending Performance Standards for Series-III Reading Assessments* (MDE, 2013).

For the Reading MCA-III OIB, operational items from one of the 2013 test administration online fixed forms served as the base. For grades 3–8, the OIB was augmented with two additional operational passages and corresponding items selected from other operational forms. MDE selected additional passages for inclusion in the OIB that complemented the content distribution of the selected operational form, in terms of standards and benchmarks assessed and the item types, and that targeted test information gaps in the OIB. This led to OIBs that included 59–70 items across grades 3–8. At grade 10, performance standards were recommended based on the single paper form, so that the core of the OIB comprised the operational items contained in that form. The grade 10 paper OIB was augmented using one field test passage and associated items from that form, which led to a total of 57 items. For grades 5–8, an OIB was created based on one of the two forms and additional field test items administered on the 2013 tests. This led to OIBs with 47–53 items across the grades.

For the MTAS-III, Minnesota's testing service provider similarly produced an OIB using both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS-III item bank.

5.4.1. Recommended Cut Scores

Table 5.5 presents the participant-recommended cut scores for the Reading MCA-III and MTAS-III after final moderation. Cut scores are shown on the theta metric. For the MTAS-III assessments, final cut scores were identified by selecting the nearest observable theta to the theta value associated with the panelists' recommended page number in the OIB. The nearest observable theta in the operational test form raw-score-to-theta table is the final recommended cut. Table 5.6 presents the impact data, or the percentage of students in each achievement level based on the cut scores after final moderation for the Reading MCA-III and MTAS-III.

Table 5.5. Participant-Recommended Cut Scores (Final Moderation) for Reading Grades 3–10

Workshop	Content Area	Grade	Cut Scores (Theta Metric): *Partially Meets*	Cut Scores (Theta Metric): *Meets*	Cut Scores (Theta Metric): *Exceeds*
MCA-III	Reading	3	-0.6589	-0.1085	1.1921
		4	-0.8084	-0.0495	1.1556
		5	-1.1292	-0.3252	1.0237
		6	-0.8162	-0.1754	0.9008
		7	-0.6654	-0.0325	1.0741
		8	-0.6514	-0.0261	1.0228
		10	-0.9714	-0.2318	0.8172
MTAS-III	Reading	3	0.6611	1.1660	2.5183
		4	1.1928	1.6441	2.6145
		5	0.8677	1.5322	3.6884
		6	0.9286	1.7583	3.5801
		7	1.1819	2.3916	3.0936
		8	1.1021	1.9319	3.7007
		10	0.8784	1.6991	2.9514

Table 5.6. Impact Data Associated with Participant-Recommended Cut Scores (Final Moderation) for Reading Grades 3–10

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	Reading	3	25	18	44	13
		4	21	25	40	14
		5	15	22	46	18
		6	21	21	37	21
		7	26	22	37	16
		8	26	21	35	18
		10	17	22	38	23
MTAS-III	Reading	3	17	12	47	24
		4	20	12	24	44
		5	16	15	52	17
		6	17	14	39	30
		7	11	24	26	39
		8	17	18	36	29
		10	18	17	28	38

5.4.2. Vertical Articulation and Moderation

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

5.4.3. Commissioner-Approved Results

After the standard setting meeting, the Minnesota commissioner of education reviewed the recommended cut scores for overall consistency and continuity. The commissioner for the 2013 MCA-III administration approved all the panelist-recommended cut scores.

5.5. Standard Setting for Grades 5, 8, and High School Science MCA-III and MTAS-III

Standard setting for the grades 5, 8, and high school Science MCA-III took place on June 25–26, 2012, in Roseville, Minn., using the Bookmark standard setting procedure (Lewis et al., 1996). Standard setting for the Science MTAS-III took place on June 27–28, 2012, in Roseville, Minn., using the ID Matching procedure (Ferrara et al., 2002). Panels were organized by grade (5, 8, and high school), with approximately 30 panelists per panel. The standard setting activities and results are recorded in *Minnesota Assessments Summer 2012 Standard Setting: Recommended Performance Standards in Grades 5, 8, and High School Science* (MDE, 2012).

For the grades 5 and 8 Science MCA-III assessments, Minnesota's testing service provider developed an augmented OIB that was built on a proportional test blueprint that included 70 items. The high school Science MCA-III contained sufficient items, so it was not necessary to augment the OIB. For the MTAS-III, the testing service provider similarly produced an OIB using both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS-III item bank. The details of the OIB construction can be found in the report *Minnesota Assessments Summer 2012 Standard Setting: Recommended Performance Standards in Grades 5, 8, and High School Science* (MDE, 2012).

5.5.1. Recommended Cut Scores

Table 5.7 presents the participant-recommended cut scores for the Science MCA-III and MTAS-III, as taken from Round 2. Table 5.8 presents the associated impact data with the cut scores. Cut scores are shown on the theta metric.

Table 5.7. Participant-Recommended Cut Scores (Round 2) for Science Grades 5, 8, and High School

Workshop	Content Area	Grade	Cut Scores (Theta Metric): *Partially Meets*	Cut Scores (Theta Metric): *Meets*	Cut Scores (Theta Metric): *Exceeds*
MCA-III	Science	5	-0.81	-0.44	0.53
		8	-0.59	0.32	1.51
		HS	-0.69	0.07	1.04
MTAS-III	Science	5	0.82	1.64	3.68
		8	0.61	1.12	2.87
		HS	0.33	1.33	1.86

Table 5.8. Impact Data Associated with Participant-Recommended Cut Scores for Science Grades 5, 8, and HS

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	Science	5	19.2	13.7	33.4	33.7
		8	27.2	29.5	33.5	9.9
		HS	23.2	25.3	34.3	17.3
MTAS-III	Science	5	13.7	15.4	50.0	21.0
		8	13.0	9.3	49.2	28.5
		HS	13.4	21.2	20.7	44.7

5.5.2. Commissioner-Approved Results

After the standard setting meeting, the Minnesota commissioner of education reviewed the recommended cut scores for overall consistency and continuity. Table 5.9 presents the final cut scores approved by the commissioner for the 2012 Science MCA-III administration, and Table 5.10 presents the impact data associated with the final cut scores.

Table 5.9. Commissioner-Approved Cut Scores for Science Grades 5, 8, and High School

Workshop	Content Area	Grade	Cut Scores (Theta Metric): *Partially Meets*	Cut Scores (Theta Metric): *Meets*	Cut Scores (Theta Metric): *Exceeds*
MCA-III	Science	5	-0.81	-0.09	1.35
		8	-0.59	0.32	1.51
		HS	-0.69	0.07	1.04
MTAS-III	Science	5	0.82	1.64	3.68
		8	0.61	1.12	2.87
		HS	0.33	1.33	2.36

Table 5.10. Impact Data Associated with Commissioner-Approved Cut Scores for Science Grades 5, 8, and HS

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	Science	5	20.1	23.1	44.9	11.9
		8	27.2	29.5	33.5	9.9
		HS	23.2	25.3	34.3	17.3
MTAS-III	Science	5	13.7	15.4	50.0	21.0
		8	13.0	9.3	49.2	28.5
		HS	13.4	21.2	31.2	34.2

5.6. Standard Setting for Grades 3–8 Mathematics MCA-III and MTAS-III

Standard setting for the grades 3–8 Mathematics MCA-III took place on June 27–29, 2011, in Roseville, Minn., using the Bookmark standard setting procedure (Lewis et al., 1996). Standard setting for the grades 3–8 Mathematics MTAS-III took place on June 29–30, 2011, in Roseville, Minn., using the Modified Angoff method (Jaeger, 1989), along with some features of the Reasoned Judgment method (Kingston et al., 2001). Panels were organized by grade band, with approximately 12–15 panelists per panel. Approximately half of the invited panelists for MTAS-III were educators involved in special education either through academic specialty or classroom experience. The standard setting activities and results are recorded in *Standard Setting Technical Report for Minnesota Assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified* (MDE, 2011).

The Mathematics MCA-III 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 task book for MTAS-III contained all the operational tasks from the 2011 MTAS-III. The tasks were ordered in the same sequence as they appeared on the test.

5.6.1. Recommended Cut Scores

Table 5.11 presents the participant-recommended cut scores, as taken from participants' Round 3 bookmark placements. Cut scores are shown on the theta metric for the MCA-III and the raw score metric for the MTAS-III. Table 5.12 presents the associated impact data.

Table 5.11. Participant-Recommended Cut Scores (Round 3) for Mathematics Grades 3–8

Workshop	Content Area	Grade	Cut Scores: *Partially Meets*	Cut Scores: *Meets*	Cut Scores: *Exceeds*
MCA-III	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
MTAS-III	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 5.12. Impact Data Associated with Participant-Recommended Cut Scores for Mathematics Grades 3–8

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	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
MTAS-III	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

5.6.2. 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.

Minnesota's testing service provider staff provided an orientation for the stakeholders who did not participate in the grade-level standard setting activities. Standard setting methods, processes, 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 5.13 presents the cut scores from the vertical articulation, and Table 5.14 presents the associated impact data. Cut scores are shown on the theta metric for the MCA-III and the raw score metric for the MTAS-III.

Table 5.13. Vertical Articulation Panel's Smoothed Cut Scores for Mathematics Grades 3–8

Workshop	Content Area	Grade	Cut Scores: *Partially Meets*	Cut Scores: *Meets*	Cut Scores: *Exceeds*
MCA-III	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
MTAS-III	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 5.14. Impact Data Associated with Articulation Panel's Smoothed Cut Scores for Mathematics Grades 3–8

Workshop	Content Area	Grade	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
MCA-III	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
MTAS-III	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

5.6.3. Commissioner-Approved Results

After the standard setting meeting, the Minnesota commissioner of education reviewed the recommended cut scores for overall consistency and continuity. Table 5.15 presents the final cut scores approved by the commissioner for the 2011 MCA-III and MTAS-III administrations on the theta metric, and Table 5.16 presents the impact data associated with the final cut scores from 2006 for the MCA-III and from 2011 for the MTAS-III.

Table 5.15. Commissioner-Approved Cut Scores for Mathematics Grades 3–8

Workshop	Content Area	Grade	Cut Scores (Theta Metric): *Partially Meets*	Cut Scores (Theta Metric): *Meets*	Cut Scores (Theta Metric): *Exceeds*
MCA-III	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
MTAS-III	Mathematics	3	0.22	0.92	2.31
		4	0.56	1.27	2.61
		5	0.17	1.54	3.13
		6	0.19	1.60	2.74
		7	0.51	1.62	2.11
		8	0.42	1.42	2.10

Table 5.16. Impact Data Associated with Commissioner-Approved Cut Scores for Mathematics Grades 3–8

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

Chapter 6: Scaling

The MCA-III and MTAS-III are constructed to adhere rigorously to content standards defined by 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, performance standards are defined to specify how much of the content standards students must demonstrate mastery of to achieve proficiency. Developing tests that are aligned with 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. Similarly, in the case of the adaptive Mathematics and Reading MCA-III, there are no test forms constructed, so the items selected by the CAT algorithm all meet content requirements. 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 (see Chapter 7: Equating and Linking).

6.1. Rationale

Scaling is the process in which student performance is associated 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 of items answered correctly. 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 MCA-III Mathematics, do not use the raw score but instead use a model-based score (pattern scoring) as the initial score. However, tests like the MCA-III also tend to report on a scale-score metric for ease of interpretation. Because the Minnesota statewide assessments are standards-based accountability assessments, the 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 and MTAS-III, the final scaling results are a designation of *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, or *Exceeds the Standards*.

6.2. Measurement Models

Item response theory (IRT) is used to derive the scale scores for all 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 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 several related measurement models. Models under the IRT umbrella include the Rasch partial credit (RPC; Masters, 1982), the two-parameter logistic model (2PL; Lord and Novick, 1968), and the three-parameter logistic model (3PL; Lord and Novick, 1968). 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. Models designed for use with test items scored as right or wrong are called *dichotomous models*. These models are used with MC, FIB, and TE items. Models designed for use with items that allow multiple scores are called *polytomous models*. Both dichotomous and polytomous models are used for the Minnesota statewide assessments.

The models used on the statewide assessments can be grouped into two families. One family is the Rasch model, which includes the dichotomous Rasch model for MC items and the RPC model for constructed-response (CR) items. Although the dichotomous Rasch model is mathematically a special case of RPC, the models are treated separately below for expository purposes. The second family of models includes the 3PL model for item types that allow guessing, such as MC items; the 2PL model for item types where the response format precludes guessing, such as FIB items; and the IRT generalized partial-credit (GPC) IRT model (Muraki, 1992) for the Science MCA-IV 3-point CR field test items.

6.2.1. Rasch Models

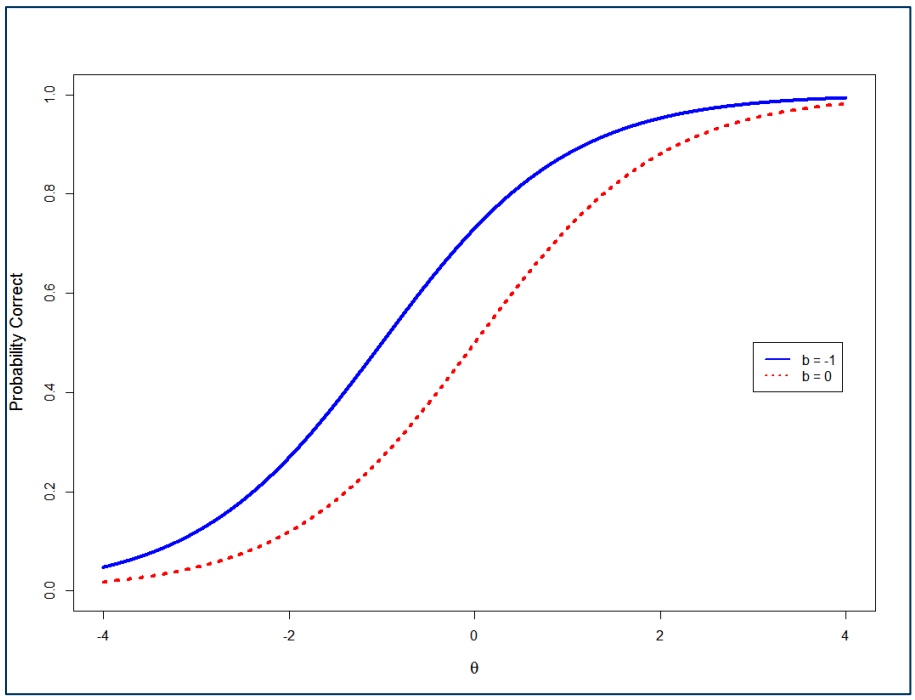
The dichotomous Rasch model can be written as the following mathematical equation, where the probability (P_{ij}) of a correct response for person *i* taking item *j* is given by:

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

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 presents the item response functions for two example items. The x-axis is the θ scale and the y-axis is the probability of a correct answer for the item. The solid curve on the left represents an item with a *b*-value of –1.0, and the dotted curve represents an item with a *b*-value of 0.0. A *b*-value of 0.0 signifies that a student of ability (that is, θ) = 0.0 has a 50% probability of correctly answering the item. The item with a *b*-value of –1.0 is an easier item, as a student with an ability (i.e., θ) of –1.0 has a 50% probability of selecting a correct answer. Students with abilities two or more theta units above the *b*-value for an item have a high probability of getting the answer correct, whereas students with abilities two or more theta units below the *b*-value for an item have a low probability of getting the answer 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 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:

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

where $x = 0, 1, 2, \ldots, m_j - 1$, and

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

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

Figure 6.2. Rasch Partial Credit Model Category Response Functions for Example Polytomous Item With b1 = –1.5, b2 = –0.3, b3 = 0.5, and b4 = 2

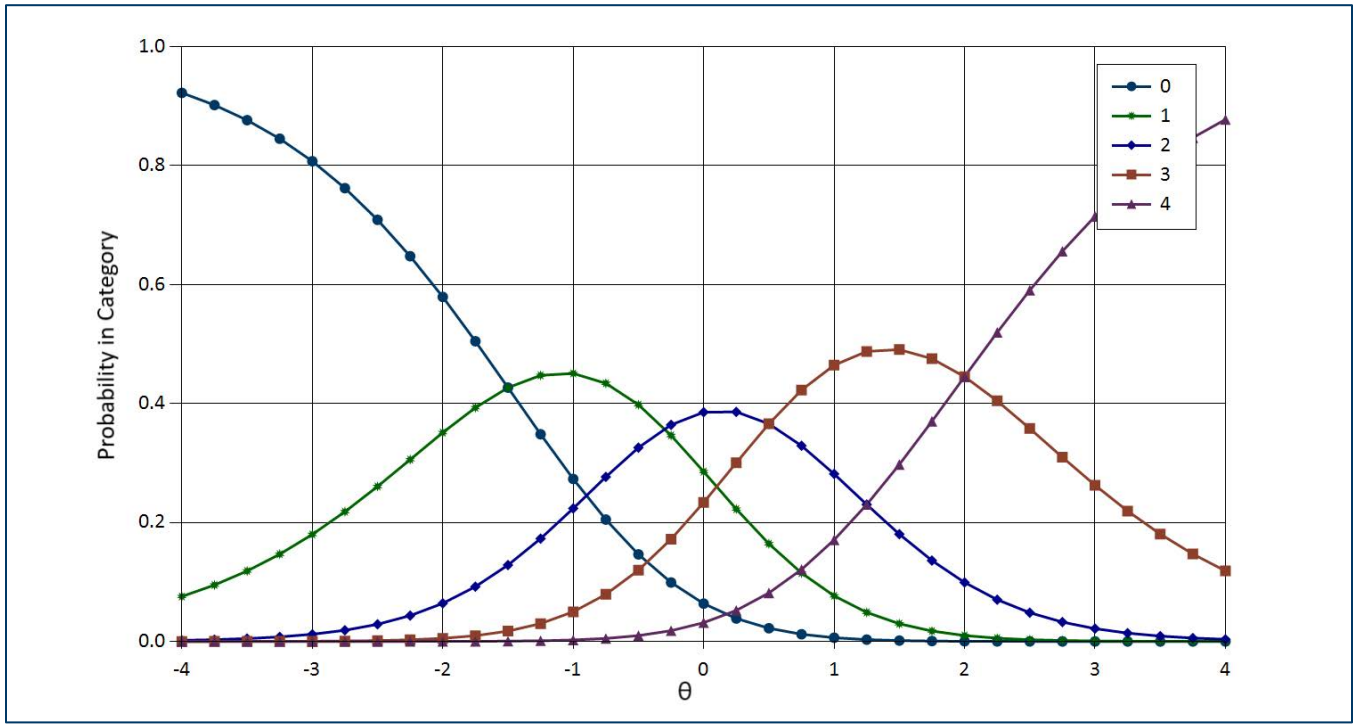
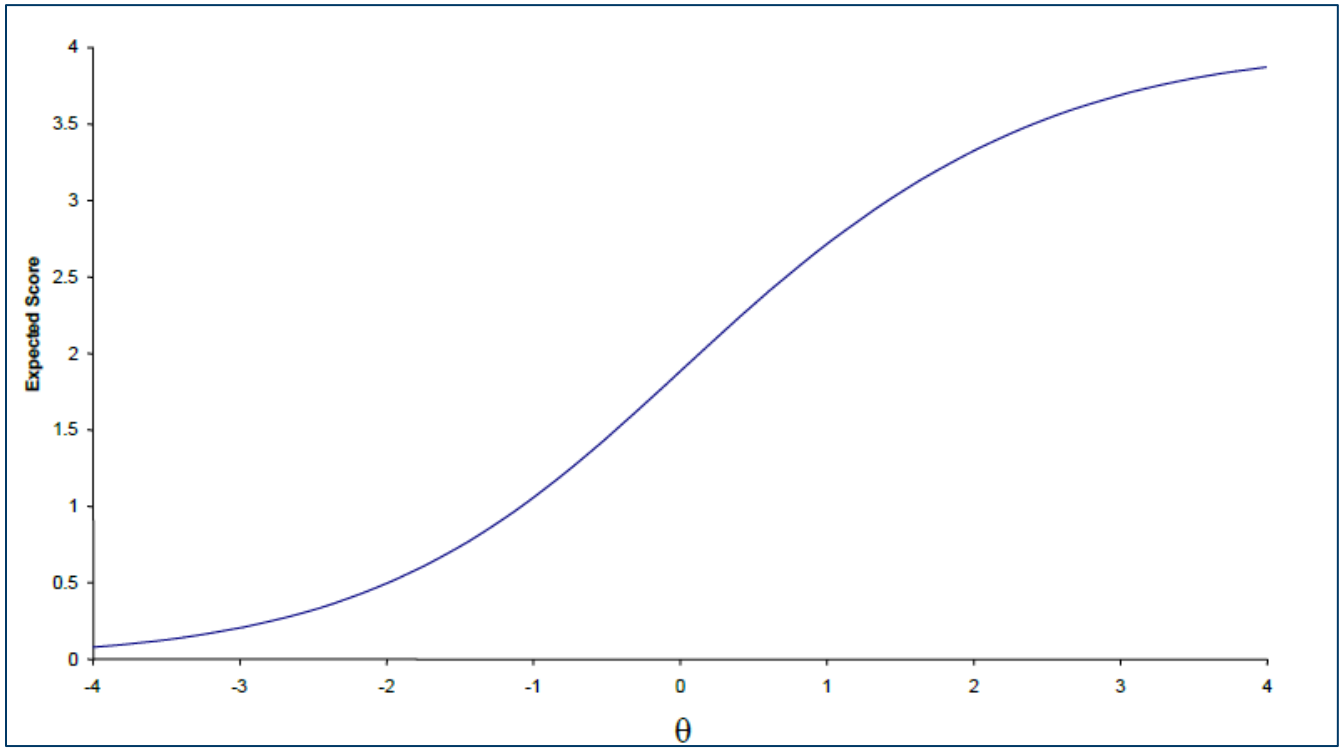


Figure 6.3 presents the average score for every ability value for the sample item given in Figure 6.2. The figure shows that students with ability θ = 0 should, on average, receive a score of "2" on the item, whereas students with ability at about 1 should average about 2.5 points on the item.

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



Calibration of items for the Rasch models is achieved using the WINSTEPS computer program (Linacre, 2006). The program estimates item difficulty for MC items and category boundary parameters for polytomously scored (e.g., CR) items. The dichotomously scored items in both the MTAS and Alternate MCA use the Rasch model, whereas the polytomously scored items use the Rasch Partial Credit model.

6.2.2. 2PL/3PL/GPC Models

This section discusses three IRT measurement models: 3PL, 2PL, and GPC. The 3PL and 2PL models are used with dichotomous items, whereas GPC is used for the Science MCA-IV 3-point CR field test items.

The 2PL/3PL/GPC models differ from the Rasch models in that the former permits 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 a student has of correctly answering an MC item. For items scored right/wrong that are not multiple choice, such as fill-in-the-blank (FIB) 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*:

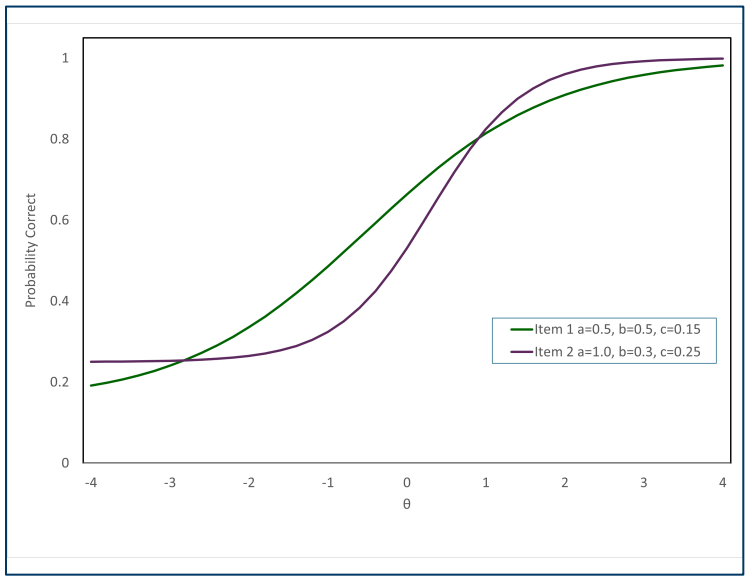
$$P_{ij} = c_j + \frac{1-c_j}{1+\exp\left(-1.7a_j(\theta_i - b_j)\right)} \quad (6.4)$$

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 above, 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.

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

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

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



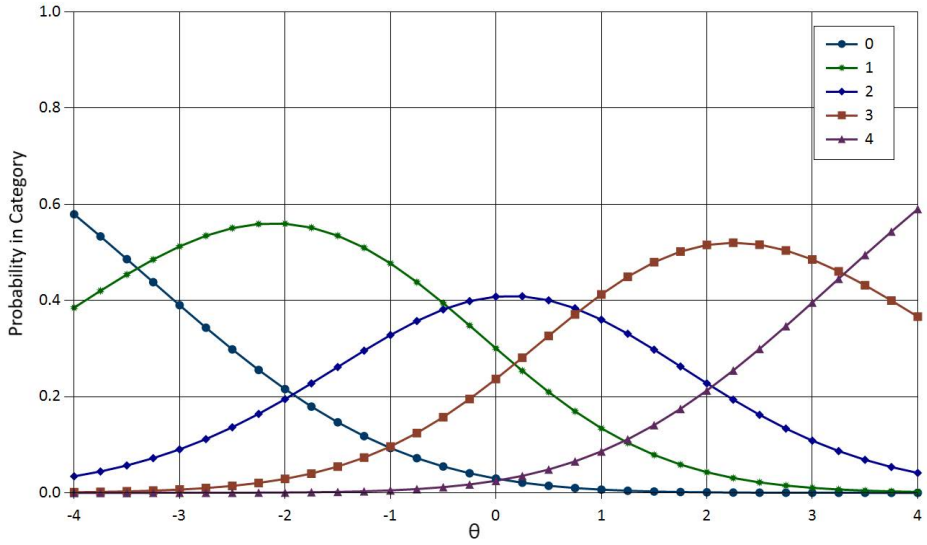
The polytomous IRT model described in this section is the GPC model. Instead of having a single probability correct, as in the 3PL model, the GPC model has a separate probability for each possible response category. The GPC model is mathematically defined as the probability of person *i* scoring in response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

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

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 that 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 d4 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 Example Polytomous Item with a=.4; b=.3; d1=0; d2=3.7; d3=.75; d4=-.5; d5=-3

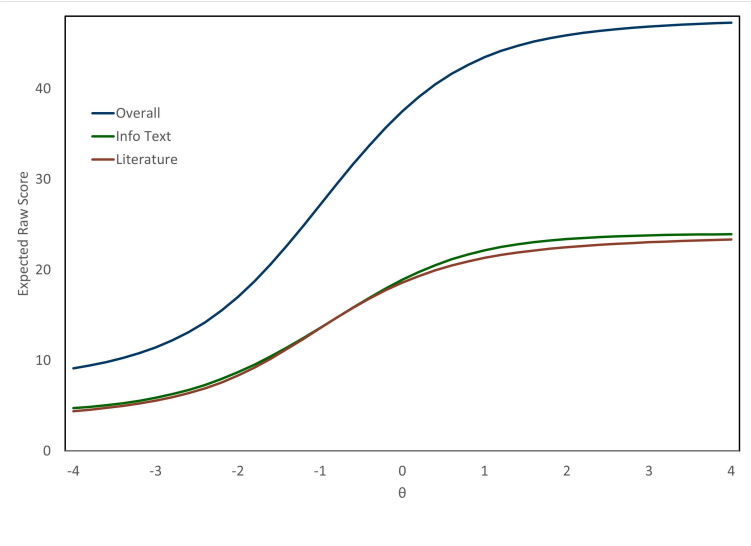


Calibration of MCA items for the 2PL/3PL/GPC models is achieved using the computer program IRTPRO 5 (Cai et al., 2011). IRTPRO estimates parameters simultaneously for dichotomous and polytomous items via a statistical procedure known as marginal maximum likelihood. Simultaneous calibration of these items automatically puts their parameter estimates on the same scale. That scale is created on the assumption that test takers have a mean ability of approximately zero and a standard deviation of approximately one.

6.2.3. Model Selection

Regardless of the 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 MCA-III. For each level of ability in the range of –4.0 to +4.0, the curve for the overall test score indicates expected performance on the number-correct scale. The graph shows that average ability students (θ = 0.0) can be expected to get a score of around 25 raw score. 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 the Minnesota tests, as will be described in the next section. In addition to the overall test score function, response functions for the two subscores are also graphed in Figure 6.6.

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



In deciding how to model responses for a particular test, measurement specialists choose from among the developed IRT models based on several considerations. Some considerations include the number and type or format of items that comprise the test, expected calibration sample size, and other general measurement theory concerns. The RPC model is well suited to model the performance task–based MTAS-III. 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, which included creating a vertical scale and linking the scales between the MCA-II and MTELL, suggested that a more complex model should be considered. After seeking the advice of the National TAC, MDE determined the 3PL and GPC models would be used for the MCA-II. The 2PL and 3PL model has been continued with the move to the MCA-III assessments.

6.3. 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 statewide assessments is derived from either the number-correct score or a measurement model-based score. These two initial scores are described below.

6.3.1. Number-Correct Scoring

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

6.3.2. Measurement Model–Based Scoring

The IRT measurement model used for Minnesota's assessments permits the use of a statistically sophisticated method that is commonly referred to as pattern scoring because the scoring procedure takes the pattern of correct and incorrect responses into account. The Mathematics and Reading MCA-III assessments make use of pattern scoring to determine student scores. Unlike number-correct scoring, where students who get the same number of dichotomously scored items correct receive the same score, pattern scoring of tests based on the 2PL or 3PL model rarely results in students receiving the same scale score even though they have the same number-correct score, because typically they differ in the items they answered correctly. Additionally, a student who gets more difficult items correct will get a higher score than a student who gets the same number of easy items correct. 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 number-correct scoring. The pattern scoring procedure used is described below.

6.3.3. Latent-Trait Estimation

For Minnesota's statewide assessments, a measurement model–based score is obtained that represents student proficiency. This is called the *latent-trait estimate* or the *theta score*. Different statewide assessments obtain the theta score in different ways. The MCA-III Mathematics and Reading assessments use a pattern scoring procedure to directly obtain the theta score from student responses of individual items. For other statewide 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.

6.3.3.1. Pattern Scoring

Pattern scoring considers the entire pattern of correct and incorrect student responses. Unlike number-correct scoring, where students who get the same number of dichotomously scored items correct receive the same score, such students in pattern scoring rarely receive the same score, as even students getting the same number correct typically differ in the items they got correct or incorrect. Because pattern scoring uses information from the entire student response pattern, this type of scoring produces more reliable scores than does number-correct scoring.

Students taking the MCA-III Mathematics and Reading assessments are assigned maximum likelihood scores that are based on the items the student answers correctly and the difficulty of those items. The Minnesota statewide assessments include multiple item types, much as MC and TE items. The likelihood for scoring using a generalized IRT model based on a mixture of item types can be written as follows:

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR} \tag{6.6}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N} \left[c_i + \frac{1-c_i}{1+exp[-Da_i(\theta-b_i)]}\right]^{x_i} \left[1 - c_i + \frac{1-c_i}{1+exp[-Da_i(\theta-b_i)]}\right]^{1-x_i} \tag{6.7}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N} \frac{exp \sum_{k=0}^{x_i} (\theta_i - b_{jk})}{\sum_{v=0}^{m_j-1} exp \sum_{k=0}^{v} (\theta_i - b_{jk})} \tag{6.8}$$

where *N* is the number of items and all other terms have been previously mentioned.

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

6.3.3.2. 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 follows:

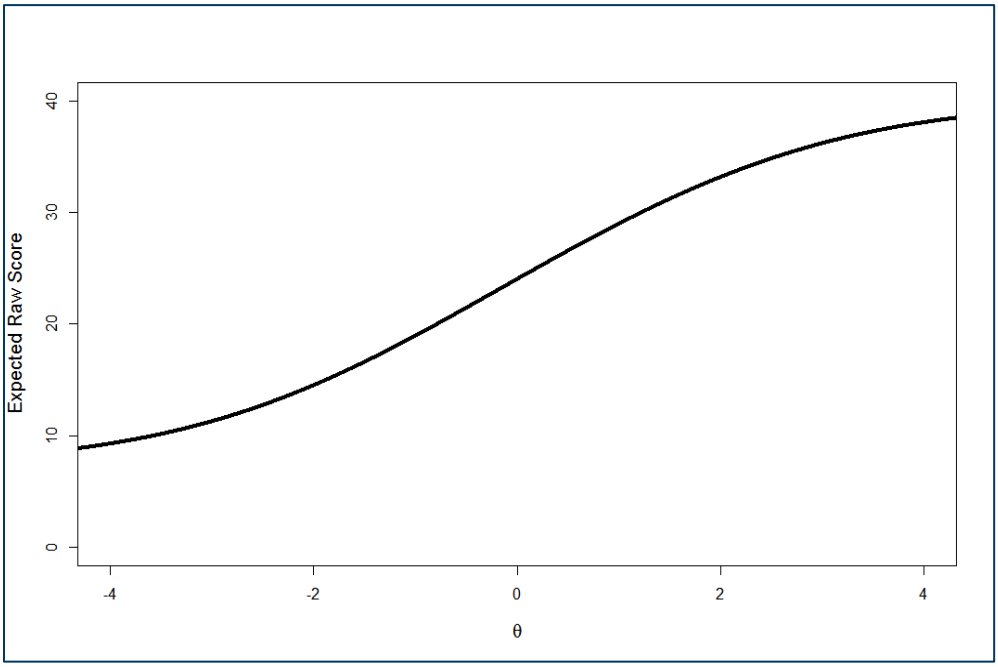
$$\text{TCF}(\theta) = \sum_{j=1}^{N} \sum_{K=0}^{m-1} kP_{ik}(\theta) \tag{6.9}$$

where *j* is an index of the *N* items on the test, *k* is an index of the *m* score categories for an item, and $P_{ik}(\theta)$ is the item response model probability correct for the item. The test characteristic function is the expected raw score given the person proficiency value θ and the item-parameter values of the IRT model.

Figure 6.7 presents the test characteristic function for a hypothetical 40-item MC test. For example, based on this figure, people with θ proficiency equal to 2.0 would, on average, have a raw score of 33. Consequently, using reverse table lookup, a raw score of 33 would be assigned an estimated theta score of 1.0.

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

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



6.4. MCA-III Scaling

To simplify comparison of student scores across years, the equated student ability estimates are transformed mathematically to a more convenient metric. For the 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.

6.4.1. Transformation

The general transformation formula used to obtain scale scores for the MCA-III is the following:

$$Scale\ Score = (\theta - \theta_{Std2}) * Spread + Center + Grade * 100 \quad (6.10)$$

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

For the MCA-III, the transformation formula uses cut scores on the θ scale. For the Mathematics and Reading MCA-III, the commissioner of education approved cut scores that were already on the θ scale. For the Science MCA-III, the cut scores on the proficiency scale were obtained by using the test response function to find the θ values that corresponded to the approved raw score cuts.

One goal for the scale transformation was to make the proficiency level scale score cuts as consistent as possible across grades. Using a linear transformation-like equation (6.8) allows two of the three scale cut scores to be fixed. As stated above, the cut score for *Meets the Standards* was set 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 set to equal g40. The *Spread* constant for each grade per subject combination was selected to force the first scale cut score to be equal to g40. The formula used to find the *Spread* is as follows:

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})} \tag{6.11}$$

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

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

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

For Mathematics and Reading MCA-III, all students are assigned a θ score by the scoring algorithm, so no further manipulation of the score is necessary. However, Science MCA-III scoring is based on raw scores, and when using IRT, special consideration is also necessary for scaling very low number-correct scores. For a test containing MC 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 items correct. Consequently, 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-correct scores below the chance level is not defined.

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

6.4.2. Progress Score

6.4.2.1. Prior to 2016

Prior to 2016, a vertical (or growth) scale linked tests in the same subject area across grade levels. With a vertical scale, the gain in knowledge from one year to the next could be measured for each student. An accurate measure of student growth is valuable information for users of test scores. The underlying assumption in using such a linked scale is that, for example, one year's grade 3 form and the following year's grade 3 form will 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. This problem can be mitigated to some degree by using common items in adjacent grades and linking grades in a stepwise fashion.

From 2012 to 2015, a vertical scale was reported for the grades 3–8 Mathematics MCA-III. Beginning in 2014, a vertical scale was reported for the grades 3–8 Reading MCA-III. This scale is called the *progress score*. Linking across grades using common items in adjacent grades formed the progress score scale. 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 allows student performance on the MCA-III to be tracked as the student progresses from grade to grade. The actual linking process used to form the IRT vertical scale is described in Chapter 7: Equating and Linking.

6.4.2.2. 2016 through 2019

In the 2016–2019 operational administrations, a direct theta-to-progress-score transformation was used for obtaining progress scores. Information regarding the scoring process for progress scores can be found in the *2017–18 Minnesota Career and College Readiness (CCR) Summary Report for the Minnesota Comprehensive Assessment (MCA)* (MDE, 2018).

6.4.2.3. 2020 and Later

Beginning with the spring 2020 operational administration, progress score reporting has been removed for grades 3–8 Reading and Mathematics MCA due to the removal of this requirement by law.

6.5. MTAS-III Scaling

The general transformation formula used to obtain scale scores for the MTAS-III is as follows:

$$Scale\ Score = (\theta - \theta_{Std2}) * Spread + Center \tag{6.16}$$

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

Chapter 5: Performance Standards describes the process of setting the standards for the MTAS-III, a procedure culminating in the commissioner of education approving the cut scores. The ability cut scores corresponding to the commissioner of education–approved raw score cuts were used to set the MTAS-III scales.

As with the MCA-III, the aim was to make the proficiency-level scale score cuts as consistent as possible across grades. Using a linear transformation-like equation (6.10) allows two of the three scale cut scores to be fixed. For all grades and subjects of the MTAS-III, 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 set to be equal to 190. Note that the 2007 MTAS-III value was 195, but beginning in 2008, the cut was changed to 190. The increase in score points for the revised MTAS-III justified a corresponding increase in scale score values between the *Partially Meets* and the *Meets* scale score cuts. The *Spread* constant for each grade and subject combination of the MTAS-III was selected to force the first scale cut score to be equal to 190. The formula used to find the *Spread* is:

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})}, \quad (6.17)$$

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.

6.6. Subscores

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

6.6.1. Subscale Reporting

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

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

$$\text{Subscale Score} = 5 + \text{Round}(2\theta) \tag{6.18}$$

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

$$\text{Subscale SEM} = \text{Round}(2*\text{SEM}(\theta)) \tag{6.19}$$

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

6.6.2. Strand and Substrand Achievement Levels

Beginning in 2016, MDE has reported strand-level ALDs for the MCA-III in the ISRs. Strand or substrand performance is reported as either Below Expectations, At or Near Expectations, or Above Expectations. Because there is measurement error in any student score estimate, there must be bounds placed around the student score estimate to assign a given student to an achievement level. Both the student strand or substrand ability estimate $\hat{\theta}$ and conditional standard error of measurement (CSEM) are used to calculate this range. Additionally, the Score Target is identical to the theta cuts score at the "Meets the Achievement Standards" level. Both $\hat{\theta}$ and CSEM($\hat{\theta}$) are described in greater detail in Chapter 9: Reliability. The computation of the lower and upper limit is made to the fourth decimal place as follows:

$$Lower\ Limit = \hat{\theta} - CSEM(\hat{\theta}) \tag{6.11}$$

$$UpperLimit = \hat{\theta} + CSEM(\hat{\theta}). \tag{6.12}$$

The lower and upper limit are then applied to the following formulas to assign a student achievement level:

$$Below\ Expectations\ (B): Upper\ Limit < Score\ Target \tag{6.13}$$

$$Above\ Expectations\ (A): Lower\ Limit > Score\ Target \tag{6.14}$$

$$At\ or\ Near\ Expectations\ (N): Lower\ Limit \leq Score\ Target \leq Upper\ Limit. \tag{6.15}$$

Table 6.1. Score Targets of Strand Achievement Levels for Mathematics MCA-III

Grade	Strand	Score Targets
3	NOPS	-0.5200
3	ALGS	-0.5200
3	GMS	-0.5200
3	DANS	-0.5200
4	NOPS	-0.4400
4	ALGS	-0.4400
4	GMS	-0.4400
4	DANS	-0.4400
5	NOPS	-0.0400
5	ALGS	-0.0400
5	GMS	-0.0400
5	DANS	-0.0400
6	NOPS	0.0300
6	ALGS	0.0300
6	GMS	0.0300
6	DAPS	0.0300
7	NOPS	0.0300
7	ALGS	0.0300
7	GMS	0.0300
7	DAPS	0.0300
8	NOPS	-0.0300
8	ALGS	-0.0300
8	GMS	-0.0300
8	DAPS	-0.0300
11	ALGS	0.1034
11	GMS	0.1034
11	DAPS	0.1034

Table 6.2. Score Targets of Strand Achievement Levels for Reading MCA-III

Grade	Strand	Score Targets
3	LSS	-0.1085
3	INFS	-0.1085
4	LSS	-0.0495
4	INFS	-0.0495
5	LSS	-0.3252
5	INFS	-0.3252
6	LSS	-0.1754
6	INFS	-0.1754
7	LSS	-0.0325
7	INFS	-0.0325
8	LSS	-0.0261
8	INFS	-0.0261
10	LSS	-0.2318
10	INFS	-0.2318

Table 6.3. Score Targets of Strand and Substrand Achievement Levels for Science MCA-III

Grade	Strand/Substrand	Score Targets
5	NSE	-0.0900
5	PSCS	-0.0900
5	ESS	-0.0900
5	LIFS	-0.0900
8	NSE	0.3200
8	PSCS	0.3200
8	ESS	0.3200
8	LIFS	0.3200
HS	NSE	0.07
HS	LIFS	0.07
HS	POSS	0.07
HS	POES	0.07
HS	INTS	0.07
HS	SFLS	0.07
HS	IALS	0.07
HS	EILS	0.07
HS	HILS	0.07

As can be seen from the above computations of the lower and upper limits, two students can differ in their achievement level even if they have the same student estimate. The upper and lower limits for a student depend on both the estimate and the CSEM from the items administered. Since students in mathematics and reading are administered different items, their CSEMs may differ, thus leading to different student achievement levels.

Caution is always required when interpreting subscale scores. Because some subscale scores are based on a few items, the error of measurement increases. That is, individual subscale scores may not be stable or consistent. As a consequence, differences in student performance across subscales may not be of practical importance. Thus, caution is required when interpreting differences between subscale scores for a student. Using the subscores at the student level is not an effective use of the data; local district and classroom assessment information is much better suited for student-level discussions of learning or understanding.

6.7. ACCESS for ELLs and WIDA Alternate ACCESS Scaling

The scaling procedures for ACCESS are available online in the following report on the WIDA website (wida.wisc.edu/resources): *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 601, 2022–2023 Administration.* The scaling procedures for WIDA Alternate ACCESS will be provided online at WIDA > Resource Library when available.

6.8. Scale Score Interpretations and Limitations for MCA and MTAS

Because the on-grade scale scores associated with the 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 to be equivalent across subjects or grades.

As indicated by equations (6.8) and (6.11), the scale score difference associated with a theta score difference of 1.0 will depend upon the *Spread* parameter. Therefore, 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 grades or subjects. The scale scores can be used to direct students who need remediation (that is, students falling below *Meets the Standards*), but scale score gain comparisons between individual students are not appropriate.

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

The scale score's primary function is to determine how far students are from the various proficiency levels without depending on the changing raw scores. Scale scores do not change in their representation of proficiency across years, whereas raw scores do not maintain their proficiency level meaning across years. Schools may also use the scale scores for purposes of program evaluation across years. For example, it is appropriate to compare the average grade 5 scale score in reading for this year to the grade 5 average for last year (if the test series has not changed). Explanations for why the differences exist will depend on factors specific to individual schools.

Beyond the information provided by the overall test scores, the strand- and substrand-level scores and descriptors provide additional information about student proficiency in various content areas within each subject. Strand scores are given in a range of 1 to 9 based on a transformation of the underlying measurement (theta) scale. Because these scores are transformations of interval-level theta scores, users are justified in treating these scores as having close to interval-level scale properties, and thus the practice of carrying out arithmetic operations on strand scores when calculating averages or summary scores is defensible.

However, users should employ extreme caution when making interpretations based on scale scores and descriptors at the strand or substrand level. These scores and descriptors are based on subsets of items administered to the students, with as few as six or seven items depending on strand or substrand, grade, and subject area. Further, these scores are distilled down into a compressed scale score range of 1–9. The strand-level ALDs, considering both the overall subject-level performance expectations and the measurement error present in the strand score, are probably the best basis for making limited instructional decisions for individual students.

Aggregations of the strand scale scores across schools or districts can serve as a guidance tool to identify possible gaps in instructional content that staff may find relevant and important. These gaps should confirm what is already being seen in classroom evidence of student learning. Different strands or substrands have different numbers of items, and the range for the number of items is defined in the testing specifications. For strands based on relatively few items, small fluctuations in the performance of relatively few students within the school or district may have disproportionately large effects on their aggregated (i.e., mean or median) scale scores. For strands based on comparatively many items, much larger changes in a greater number of students' performances would be required for an effect of equivalent magnitude to be observed. For this reason, strand-to-strand comparisons within or (especially) across grades and/or subjects are usually not appropriate and may lead to incorrect conclusions.

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

6.9. Conversion Tables, Frequency Distributions, and Descriptive Statistics

The *Yearbook Tables for Minnesota's Statewide Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Yearbook Tables expandable header)) provides tables for converting raw scores to derived scale scores for the fixed-form assessments (MCA-III Science and all MTAS-III assessments) and tables of frequency distributions and summary statistics for scale scores by grade and subject under the sections entitled "Frequency Distribution Reports."

Chapter 7: Equating and Linking

Equating and linking are procedures that allow test scores to be compared across years. The procedures are generally thought of as statistical processes applied to the results of a test. However, 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 the MCA and MTAS.

The equating and linking procedures for ACCESS are available online in the following report (wida.wisc.edu/resources): *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 601, 2022–2023 Administration*. The equating and linking procedures for WIDA Alternate ACCESS will be provide online at WIDA > Resource Library when available.

7.1. Rationale

To maintain the same performance standards across different administrations of a particular test for linear, fixed-form tests, 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 a subsequent administration of a test to the same scale as the scores of the first 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.

Historically, a “two-stage with pre- and post-equating” design was used to maintain comparable difficulty across administrations for the MCA-II and with the large-scale paper form administrations of the Mathematics and Reading MCA-III. Both “pre-equated” and the “two-stage pre- and post-equating” designs are commonly used in state testing. In the pre-equating stage of a “two-stage pre- and post-equating” design, item-parameter estimates from prior administrations (either field test or operational) are used to construct a form with a difficulty level similar to that of previous administrations. This is possible because of the embedded field test design that allows for linking 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 for which 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.

MDE now uses a pre-equating design for the MCA and MTAS. A benefit of online testing is on-demand reporting. When moving to online assessments, MDE decided to use a pre-equating design to allow for immediate score results reporting (On-Demand Reports). In a pre-equated design, all items are placed on the base scale prior to an operational administration and the banked item parameters are used for scoring. The pre-equating design is fully described in the sections that follow.

7.2. 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 calibrates all new field test items to the base scale. This results in a bank of items used for scoring student responses, which are all on the same base scale. In this way, each item is placed on the same metric as that of the prior years so the metric is maintained across years. Each new MCA-III assessment is constructed from a pool of items for which parameters have been equated to the base scale. The base scales were established in 2011 for grades 3–8 mathematics, in 2012 for science, in 2013 for reading, and in 2014 for grade 11 mathematics. New items are equated to the base scale during field testing.

A major advantage of pre-equating is that once item parameter estimates are determined, those estimates can be used for scoring in situations where post-equating would be difficult or ill-advised. For example, the COVID-19 pandemic raised concerns for testing programs in other states that required post-equating procedures for the 2021 administration, as it was unknown what the pandemic impact might be on test participation or performance. Because Minnesota uses pre-equated parameter estimates for its assessments, these pre-equated parameter estimates could be used for the 2021 administration to provide scores (e.g., achievement level and scale score) that can be validly compared to previous years.

7.2.1. Test Construction and Review

7.2.1.1. Fixed-Form Assessments

Test construction for MCA-III Science fixed-form assessments begins by selecting the operational items for an administration. Using the items available in the item pool, psychometricians and content specialists from Minnesota's testing service provider and MDE construct new forms by selecting items that meet the content specifications of the subject tested and targeted psychometric properties. Psychometric properties targeted include test difficulty, precision, and reliability. The construction process is an iterative one, involving Minnesota's testing service provider and MDE staff. Because the IRT item parameters for each item in the item bank are maintained on the same scale, direct comparisons of test characteristic functions and test information functions can be made to ascertain whether the test has similar psychometric properties (e.g., difficulty) to those of other years. Having all items on the same scale allows the psychometricians to create raw score-to-scale-score lookup tables to be used for scoring purposes.

Psychometricians and content staff review the newly constructed test to ensure that 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 reexamines 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 correct 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 an item in the item bank and the review process begins again. This process for reviewing each newly constructed test form helps ensure each test will be of the highest possible quality.

7.2.1.2. Simulations for Adaptive Assessments

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

The adaptive algorithm and a complex blueprint have many adjustable parameters, such as balancing the weight given to one strand versus other strands or item type constraints. More specific details regarding the WPM and CRM algorithms used for the simulation can be found in Chapter 1: Background. The optimal values for the algorithm parameters vary depending on the item pool, specifics of the blueprints, and their interaction. Some of the most important variables, which are fine-tuned during the simulation stage prior to administration, are the (a) weights assigned to the strand for mathematics (substrand for reading), (b) weights assigned to the standard, (c) starting theta, (d) weights assigned to the item information relative to the test specifications, (e) theta range, and (f) number of items selected per range. Prior to each operational testing window, simulations are conducted for each grade separately for each subject to evaluate the quality of the adaptive item selection for mathematics, and testlet-selection for reading algorithm. Simulations enable key blueprint and configuration parameters to be manipulated to match the blueprint, minimize measurement error, and control item exposure.

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

Reading and Mathematics MCA simulations differ because, unlike for mathematics, the reading simulations are given in stages, each of which is a testlet with one or more passages and their associated items depending on the grade. Students encounter three operational testlets during the test. The item order within each testlet is fixed, and the items always appear in the same order. The testlets are built to stage specific blueprints such that any combination of a stage 1, stage 2, and stage 3 testlet will always meet the overall test blueprint, and constraints are imposed so that each student will always be administered exactly one testlet from each of the three stages. In this way, all students receive a test that complies with the overall test blueprint. The assessment contains between four and seven passages for grades 3–8 and between four and eight passages for grade 10.

7.2.2. MCA Field Test Items

When a newly developed item has survived committee reviews (passage review for reading, scenario review for science, and new item review and bias and sensitivity review for mathematics, reading, and science), the item is ready for field testing. For the Mathematics MCA-III, field test items are randomly placed in the test at pre-selected positions where for grades 3–8 the calculator items and non-calculator items are placed within their respective sections. The field test items are arranged in blocks and each student is administered only one set of items. For the reading test, the items appear in testlets along with their respective passages, which are placed at pre-selected positions within the test. For science, the field test items are embedded in a test form among the operational test items. For example, in a particular grade's Science MCA-III administration, there may be 15 different forms containing the same operational test items; however, each form would also contain one or more unique field test scenarios and corresponding unique field test items. The field test items do not count toward an individual student's score.

In online administrations of fixed forms, forms are assigned randomly to students. For example, for grade 5 science, with a statewide enrollment of approximately 63,000, approximately 3,700 students would respond to each of 17 field test forms. In online adaptive tests, field test items are assigned at random to students in designated slots during the administration. This 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, 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. For the paper accommodated forms of the MCA-III and MTAS-III data-entry forms, there is one operational form that does not contain field test items.

7.2.2.1. 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 required. Some student data, however, are excluded from the pre-equating calibration of items. In addition, the responses of home-schooled and private school students are excluded from the calibration data set. Home-schooled and private school students are not required to take the MCA-III and are not included in statewide summary statistics or in ESSA calculations. Their test scores are reported to students, parents, and schools, similar to students at public schools. If the number of items a student attempts does not meet the minimum attemptedness criterion, then data from that student are excluded from the calibration data set. For the MCA-III and other online assessments, students must respond to at least 90% of the items on the test to be classified as "attempted."

7.2.2.2. Pre-equating Quality Checks

When the statewide data file has been edited for exclusions, a statistical review of all operational and field test items is conducted before beginning IRT calibration. Items are evaluated for printing or processing problems. A key check analysis is conducted for the MC items, which entails an evaluation of the mean score, percentage of students who gave each possible response, and the item-total correlation. Items where an unusually attractive incorrect option appears on any one form, which differs substantially from all other forms or items with a low mean score or low point-biserial, are flagged for review. An adjudication analysis is conducted for all non-MC items, which involves the content specialists analyzing for correctness every response string not checked in a previous administration. The key check and adjudication process occurs at least three times during the testing window for operational items (once near the beginning of testing, once toward the middle of testing, and once

following the end of testing). The key check and adjudication are only conducted at the end of the testing window for field test items. Minnesota's testing service provider and MDE content staff review each flagged item, as administered, to ensure that the item was correctly printed (for paper accommodated forms) or displayed (for online forms) and to certify the key is the correct answer.

7.2.2.3. Field Test Item Equating Procedures

For all MCA-III assessments, the commercial software IRTPRO is used for item calibrations. For reading, all MC items are calibrated as three parameter items, and non-MC (TE) items are calibrated as two parameter items. For mathematics, all FIB items are calibrated as two-parameter items; all other types of items (including all TE items) are calibrated as three-parameter items. For science, all items are calibrated as three-parameter items, except for 3-point CR items that are calibrated using the GPC model. These models are described in Chapter 6: Scaling.

7.2.2.4. Evaluation of Operational Item Parameter Drift

In IRT, item parameter drift is a change in the item parameters over time, of which particular concern is placed on the change in the difficulty of an item over time. To determine the final set of anchor items, drift analysis is conducted to flag items that may have moved from their bank parameters (on the base scale). The purpose of the drift analyses is to identify items that may have shifted in difficulty relative to the bank as a whole. This might occur because of changing emphases in instruction throughout the state from year to year, exposure of the item, or for many other reasons. Because sample sizes are too small on the paper accommodated forms for these analyses to be effective, they are only conducted on the online forms.

When assessments are scored using pre-equated item parameters, there is no post-administration calibration and equating of operational items. However, items must still be examined for signs of misfit or drift. Beginning with the spring 2014 administration of the MCA-III Mathematics grades 3–8 online, the MCA-III Reading grades 3–8 and 10 online, and MCA-III Science assessments, items have been evaluated for parameter drift. The general approach to evaluating goodness of fit involves the comparison between observed and model-predicted frequencies for various ability (theta) subgroups using d-square (Wells et al., 2014) and robust z based on chi-square fit statistics methods. The item-fit statistics employ a pseudo-observed theta distribution as proposed in Stone (2000). MDE, the testing service provider, and Minnesota's quality assurance service provider conduct the initial drift analyses, and items with large (i.e., exceeding predefined critical values) fit statistics are flagged. For mathematics, flagged items are evaluated by content staff for possible release and are removed from the operational item bank. For reading, items flagged in a single year are identified as "potentially flagged" and are monitored for drift again in the next following administration; items flagged in two consecutive years are either recalibrated or removed from the bank.

Items that are identified as having drifted in the operational pools of the computer adaptive mathematics and reading tests will either be recalibrated or removed from future administration. To be eligible for recalibration, items must meet certain n-count requirements and must have a distribution that is reasonably similar to that of field test items. Items not meeting these requirements must be removed from the operational pools because such items' parameters would be estimated based on a sample of students from a restricted range of ability and would not yield an accurate estimate of the item parameters. Such items in the mathematics and reading adaptive assessments must be removed from the item bank or be re-field tested. However, science items are administered operationally on fixed forms to students of all ability levels, so items identified as drifted can be recalibrated. Therefore, for science when an item has been identified as drifted, the item is recalibrated and that updated set of item parameters is used operationally thereafter.

7.2.2.5. Field Test Calibration

Historically, when the MCA-III was only administered on paper accommodated forms, the Stocking-Lord procedure was used to equate the field test items for the non-accommodated paper forms. However, because there were no field test items on the paper accommodated forms of the MCA-III, no equating is conducted.

For the computer-adaptive Reading and Mathematics MCA-III, non-drifted operational items and field test items are calibrated together in IRTPRO with the non-drifted operational items being fixed to their bank parameters and field test items being freely estimated. By anchoring the non-drifted operational items to their bank value, the field test items are automatically placed on the base scale (Stocking and Lord, 1983).

For fixed-form online Science MCA-III, field test item calibration has alternated between using the Stocking-Lord (1983) and fixed operational-item parameter approaches, depending on the testing service provider, with the Stocking-Lord approach currently being applied. The program STUIRT (Kim and Kolen, 2004) is used to find scale transformation constants, slope A and intercept B, to place the item parameter estimates of the new (N) items onto the operational (O) scale. Transformations for IRT parameters of items using item index *j* are given as follows:

$$a_{jO} = \frac{a_{jN}}{A} \tag{7.1}$$

$$b_{jO} = A \times b_{jN} + B \tag{7.2}$$

$$c_{jO} = c_{jN} \tag{7.3}$$

$$d_{jvO} = A \times d_{jvN} \tag{7.4}$$

7.3. MTAS Equating

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

Because field test items were administered in the current operational year for the Science MTAS-III, field test calibration and equating were needed to put the field test items in the same scale as the base year. Historically, equating to the base year was accomplished using conceptually similar procedures to those used with the MCA-III. For MTAS-III, a simultaneous calibration of operational and field test tasks was performed by grade and subject. Specifically, this was accomplished by using fixed anchor calibration, where the field test items were calibrated by anchoring on all operational items. The fit of field test tasks to the model was scrutinized to ensure that a poorly fitting field test task did not compromise the calibration of the operational tasks. In addition, linearity was checked by plotting linking task IRT difficulty values against those from the base year. Linking tasks were 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 was then added to the IRT difficulty values of the linking tasks. The equated IRT parameters were then compared with the base-year values. Differences between equated and base-year values are called *displacement value*s. Displacement values were scrutinized, and tasks with displacements greater than 0.3 were considered for removal from the equating. After dropping any linking task that failed the stability check, another WINSTEPS calibration was performed for all tasks with linking task parameters fixed to their base-year values. The task parameter values from the second calibration were considered the final parameter values for purposes of scale score calculation and item banking.

7.4. Item Pool Maintenance

The next step is to update the item pool with the new statistical information. Item statistics and parameter estimates for the field test items and recalibrated operational items due to drift issues for fixed forms are added to the item pool database. Since pre-equating has been used, new parameter estimates are not obtained for operational items, except for the recalibrated operational items due to drift issues for fixed forms.

7.5. 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, linking has been used for two purposes: (a) scaling across grades with the progress score (mathematics and reading) prior to 2016 and (b) linking the Reading MCA-III to the Lexile® Reading scale and the Mathematics MCA-III scale to the Quantile® Mathematics scale. 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, linking is used. 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).

7.5.1. Linking Grades 3–8 to the Progress Score (Prior to 2016)

Prior to 2016, the Mathematics and Reading MCA-III each used a vertical scale called the *progress score*, which contained linking items. These MC items from an adjacent grade's test were used to link grades on the progress score. The grades 10 and 11 tests did not use vertical linking items because no progress score was reported for these grades. Vertical scales, such as the progress score prior to 2016, were designed to help evaluate how much students improved from one year to the next. Linking for the progress score was accomplished by using common items on adjacent grades on the 2011 Mathematics MCA-III administration and the 2013 Reading MCA-III administration. The linking of off-grade items did not count toward a student's final score. The linking design was such that no student responded to both upper-grade items and lower-grade items. For example, some fourth-grade students responded to a linking set of third-grade items, and some fourth graders responded to a linking set of fifth-grade items. The determination of which students responded to the linking sets was done by random assignment.

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

The vertical scale was replaced with a progress score interpreted in relation to the grade in which a student belonged by using a direct theta-estimate-to-progress-score transformation where off-grade items were linked to the current grade metric through the previously derived vertical equating relationships. Vertical scaling transformation constants (slope and intercept) were used to derive the discrimination and difficulty parameters to place the off-grade items on each grade's respective scale. The transformed item parameters were used for item selection during the MCA-III adaptive assessments. However, off-grade items were post-equated following administration and new parameters estimated for these parameters. The post-equated parameters were used for student scoring purposes in 2016, but after review of the results from the post-equating, it was decided that they would not be used in subsequent administrations (the original transformed item parameters will instead be used for both item selection and scoring in future administrations).

Beginning with the spring 2020 operational administration, progress score reporting is removed for grades 3–8 Reading and Mathematics MCA.

7.5.2. Linking Reading MCA-III to the Lexile® Scale

MetaMetrics typically uses a common-person design to develop linkages between statewide assessments and their proprietary Lexile scale. For example, to link the previous MCA-II Reading scale to the Lexile scale, MetaMetrics administered a stand-alone test to students in a sample of districts following regular administration of the MCA-II Reading assessment in 2010. However, the common-person design used to establish the initial linkage had significant disadvantages. The independent assessment, although administered concurrently with the accountability assessment, was voluntary and carried no stakes for students or schools. Consequently, motivation for high performance may have been diminished. Motivation effects may have been more pronounced for older students, especially grade 10 students. Younger students may not have readily made distinctions between high-stakes and low-stakes testing situations and thus have treated the assessments in the same manner. Perhaps more important, this testing design placed a substantial additional assessment burden on participating schools and students (as well as an increased burden on MDE for recruiting sampled schools), requiring approximately the same amount of testing time used for the MCA Reading assessment.

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

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

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

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

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

7.5.3. Linking Mathematics MCA-III to the Quantile® Scale

For the MCA-III Mathematics assessments, MetaMetrics agreed to allow MDE to embed Quantile® items in field test slots of the spring 2018 accountability test administration. Embedding Quantile items on the MCA-III Mathematics assessment allowed MDE to administer Quantile items under operational testing conditions and confine the burden of field testing to the standard administration of the accountability assessment, eliminating much of the cost and burden of a stand-alone field test.

Subject matter experts selected pools of Quantile linking items that best aligned with the grade-level *Minnesota K–12 Academic Standards in Mathematics* based on both content and difficulty. To achieve this, the percentage of Minnesota Academic Standards represented on each grade level of the Mathematics MCA-III assessment were reviewed and aligned with the content strands of MetaMetrics Quantile Framework. The Quantile linking item pool contained 31 items in each of grades 3–8 and 36 items in grade 11. Each grade-level set included items from an adjacent grade to provide connectivity for the linking analysis. One or two items were administered to each student during the Mathematics MCA-III administration. Each linking item was evaluated for use in the linking study based on potential alternate answer choices being more attractive than the correct answer choice (i.e., low point-measure correlation). A total of 21 items across all grades were removed because they had low point-measures or misfit criteria outside the acceptable range. With Quantile items administered alongside MCA-III Mathematics items, the items were calibrated concurrently, allowing all items to be placed on a common scale.

Three steps were performed prior to the linking analysis. First, a concurrent calibration of all Mathematics MCA-III assessment items and Quantile linking items was conducted to evaluate the appropriateness of scaling both Quantile and MCA-III items on the same scale. Second, a concurrent calibration of the Mathematics MCA-III items with the Quantile linking items anchored to their theoretical Quantile values was conducted to place the Mathematics MCA-III items on the Quantile scale. Finally, a scoring run using only the Mathematics MCA-III items on the Quantile scale was conducted to express student results from the Mathematics MCA-III assessment in the Quantile metric. These three steps were performed separately for each grade.

During the initial concurrent calibration for each grade, data for all students were submitted to a WINSTEPS (Linacre, 2006) analysis using a logit convergence criterion of 0.0001. Student records were removed from further analysis if the data did not fit the Rasch model, indicated by an infit statistic greater than 1.5 and outfit statistic greater than 2.0. Approximately 96.54% of the initial sample remained in the final sample for the Mathematics MCA-III link.

The LEGS (*L*inking with *E*quivalent *G*roups or *S*ingle Group Design) software program used for calculating equivalent scores using equipercentile methods was employed to conduct an equipercentile linking of the Mathematics MCA-III assessment unrounded scale scores and the Mathematics MCA-III calibrated Quantile measures for grades 3–8 and 11 (Brennan, 2004). Equipercentile linking functions were constructed relating the Mathematics MCA-III scale scores and Mathematics MCA-III calibrated Quantile measures for all students in the sample, by grade level. Conversion tables were developed for all grade levels to express the Mathematics MCA-III scale scores in the Quantile metric. Student Quantile measures were reported with a +/–100 Quantile measure of upper and lower bounds.

Chapter 8: Validity

The term *validity* refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014). Validation can be described as the process of collecting evidence to support inferences from assessment results. A primary consideration in validating test scores is determining whether the test measures what it purports to measure: the construct. 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 can be an example of 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 that this is a reasonable and valid use of the scores. During the process of evaluating whether the test measures the construct of interest, several threats to validity must be considered. For example, the test may be differentially more or less difficult for a particular demographic group relative to another group, test scores may have lower than desirable levels of reliability, students may not be properly motivated to perform on the test, or the test content may not span the entire range of the construct to be measured. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring the test is measuring what it is supposed to measure, it is also important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Section 4.4: Cautions for Score Use and Section 6.8: Scale Score Interpretations and Limitations for MCA and MTAS.

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought, and debate in the educational measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result, the ways in which test validity and validation are defined has evolved. This chapter summarizes validity evidence for MCA-III assessments and is based on the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) refers to "types of validity evidence, rather than distinct types of validity." The four broad categories of validity evidence mentioned in the *Standards* that are relevant to the Minnesota statewide assessments are (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, and (d) evidence based on relationships with other variables. Taken together, a combination of these types of validity evidence can be used to create a validity argument (Cronbach, 1988). It is important to note that the types of validity evidence selected for a given assessment must be relevant to the selected measure, so not every form of validity evidence applies to every assessment.

8.1. Evidence Based on Test Content

Content validity evidence addresses whether the test adequately samples the relevant domain of material it purports to measure. 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 high content validity. For example, a content-valid test of mathematical ability should be composed of items that allow students to demonstrate their mathematical ability. One way to evaluate the content validity of an assessment such as the MCA-III is to evaluate the alignment of the standards with test content.

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

Evaluating content validity is a subjective process that is based on rational arguments. Generally, experts make judgments about agreement between the parts of the test and construct. This process often involves experts assigning test items to one of the major content areas being measured in the assessment. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity speaks only to the validity of the test itself, not to decisions made based on the test scores. For example, a poor score on a highly content-valid mathematics test indicates that the student did not demonstrate mathematical ability. But from this alone, one cannot conclusively conclude the student has low mathematical ability. This conclusion can only be reached if it can be shown or argued that the student put forth his or her best effort, the student was not distracted during the test, and the test did not include content-irrelevant elements that prevented the student from scoring well.

To ensure that the content is aligned with the construct, the development of the items is based on test specifications (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Test Specifications) for each subject and grade that is being assessed. Rigorous processes have been put in place to align items and test forms with the standards while developing items for Minnesota statewide assessments. As a result, each Minnesota assessment is developed with content-related validity evidence in mind.

Panels consisting of members of the Minnesota Academic Standards Committee and educators were convened to develop the test specifications for each test, subject, and grade. Many of the educators were recommended to the MDE by various education organizations, school districts, and other stakeholder groups. The starting point for development of the test specifications was revision of the Minnesota Academic Standards for the relevant subject and grade (for mathematics, the 2007 version, and for reading, the 2010 version of the language arts standards, and 2009 version of the science standards). These panels developed the test specifications, and their decisions regarding the specific subject matter that was to be assessed were made with reference to these standards. Therefore, the administered test is based on the Minnesota Academic Standards (which detail what should be taught from each subject at each grade level), informed by the expertise of selected Minnesota educators.

The current test specifications identify eligible test content and provide item count targets for various item properties such as content strands or substrands, standards, domains, item types, and DOK levels. These targets are codified into a test *blueprint*, which provides direction to item writers, psychometricians, content specialists from Minnesota's testing service provider, and MDE so that all relevant content is sufficiently covered by the assessment. This coverage is one piece of evidence for the content validity of the test.

Both the testing service provider and MDE are involved in item development. The items are developed based on the test specifications. The items are rigorously scrutinized during the content review, which involves all members of the assessment team. This review checks for the appropriateness of test items, difficulty, clarity, correctness of answer choices, plausibility of the distractors, and fairness of the items and tasks. Then the items must be reviewed and approved by the content review committees, which assure that each item appropriately measures the intended content, is appropriate in difficulty, contains only one correct (or best) answer for MC items, and has an appropriate and complete scoring guideline for TE items. Next, a bias and sensitivity committee must approve the items, reviewing each item for language or content that may be inappropriate or offensive to students, parents, or community members, or that contains stereotypical or biased references to gender, ethnicity, or culture.

A set of separate alignment studies for each subject was conducted for the MCA-III and MTAS-III tests. An external independent service provider conducted these studies to provide evidence that a given assessment was aligned with its respective set of test specifications. Specific areas of interest included both how much and what type of content was covered by the assessment, as well as whether students were being asked to demonstrate knowledge at the same level of rigor as stipulated by the content standards. Each alignment study identified weaknesses in the assessments and provided recommendations to strengthen the alignment between the assessments and the Minnesota Academic Standards, as codified within the test specifications in future assessment years. Information from these studies was used to modify the pool to account for areas that are lacking.

Following the first administration of new assessments aligning with revised academic standards, MDE also develops ALDs, which provide a description of typical grade-level performance for the achievement levels. The ALDs are descriptions of the knowledge and skills demonstrated by students in each performance category. Higher scores translate to a greater level of knowledge and skills demonstrated. There is a link between the ALDs and the knowledge and skills required to meet proficiency according to the standards. To ensure the ALDs have high validity, the ALDs are developed by content area experts and stakeholders.

Content experts and stakeholders participate in standard setting, a process setting the levels of performance on the assessment that are reported to students, parents, and schools. This committee sets the cut scores that delineate the four levels of achievement reported in Minnesota (*Exceeds the Achievement Standards*, *Meets the Achievement Standards*, *Partially Meets the Achievement Standards*, and *Does Not Meet the Achievement Standards*). The Achievement Level Descriptors (ALD)s (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Achievement Level Descriptors) define the grade-level student performance in each level of achievement based on the assessment results.

Also important for content validity is the control of random measurement error. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Reliability and the standard error of measurement (SEM) are discussed in Chapter 9: Reliability. The *Yearbook* has tables reporting the SEM (for fixed-form tests) and the coefficient alpha reliabilities for raw scores for fixed-form tests, broken down by demographic groups. Additionally, the *Yearbook* also reports tables comparing the CSEM of the MCA adaptive tests with previous administrations. These measures show the Minnesota statewide assessments to be reliable.

8.2. Evidence Based on Response Processes

Validity evidence based on response processes involves explicit assumptions about the cognitive processes engaged in by the test takers. Analyses of the response processes of test takers provides evidence concerning the fit between the construct and the nature of the performance or response required of the test takers (AERA et al., 2014). Put another way, if an assessment is designed to measure mathematical reasoning, the assessment should be evaluated to determine if students are actually evaluating the mathematical questions as planned. Generally, this type of evidence is inferred through analysis of individual responses to determine the methods and strategies that students employ when answering a given item. This evidence is most frequently obtained through a cognitive lab, a response process study in which test takers from different groups are monitored to determine the process they go through to answer a given item.

The test specifications discussed previously include the number of items targets for each of first three DOK levels for mathematics and reading. DOK, or cognitive complexity, refers to the cognitive demand associated with an item. The level of cognitive demand focuses on the type and level of thinking and reasoning required of the student when interacting with a particular item. Levels of cognitive complexity for MCA-III are based on Norman L. Webb's (Webb, 1999) DOK levels:

- Level 1 (recall) items require the recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula. A well-defined and straight algorithmic procedure is at this level. A Level 1 item specifies the operation or method of solution, and the student is required to carry it out.
- Level 2 (skill/concept) items call for the engagement of some mental processing beyond a habitual response, with students required to make some decisions as to how to approach a problem or activity. Interpreting information from a simple graph and requiring reading information from the graph is a Level 2. An item that requires students to choose the operation or method of solution and then solve the problem is a Level 2. Level 2 items are often similar to examples used in textbooks.
- Level 3 (strategic thinking) items require students to reason, plan, or use evidence to solve the problem. In most instances, requiring students to explain their thinking is a Level 3. A Level 3 item may be solved using routine skills, but the student is not cued or prompted as to which skills to use.
- Level 4 (extended thinking) items require complex reasoning, planning, developing, and thinking, most likely over an extended period of time. Level 4 items are best assessed in the classroom, where the constraints of standardized testing are not a factor.

Response process validity evidence is most frequently provided through conducting cognitive labs with students who are interacting with an item. Cognitive labs are not conducted for the Minnesota tests. Instead, each item is developed to strictly adhere to one of the first three DOK levels and is reviewed internally by both the content

teams of the testing service provider and MDE. Qualified educators and community members who interact with students in the classroom review and verify the DOK levels of each field test item. These committees act as a proxy for the students by considering the process that students follow while responding to a given item. Of particular concern during item development is the development of items that contain no irrelevant information that may interfere with how the item is interpreted or scored. The test specification review committees, who have experience working with students and their cognitive processes on a daily basis, determine what proportions of the test that should be devoted to items at each of the first three levels of DOK.

8.3. Evidence Based on Internal Structure

Internal structure validity evidence shows the degree to which items and test components conform to the construct on which the proposed test score interpretations are based (AERA et al., 2014). For instance, a mathematics test may be broken into several strands such as data analysis, algebra, geometry and measurement, and number and operation. Internal structure validity evidence identifies the degree to which the item relationships conform to the individual subscales and overall mathematics scale.

One type of evidence for internal structure that is provided for the fixed-form MCA-III Mathematics, Reading, and Science and all MTAS-III assessments is dimensionality analysis, which is often referred to as factor analysis. The dimensionality analysis identifies several components that best explain the relationships among the items. It is common for educational assessments to measure more than one dimension, but generally these tests measure a strong major dimension and several minor or less important factors. Each MCA-III and MTAS-III assessment is designed to measure a multifaceted composite of knowledge and skills appropriate for the subject and grade. This composite of knowledge and skills is expected to be composed of separate, but highly correlated, components such that the measured composite can be considered as a unidimensional construct, thus permitting the use of unidimensional IRT models.

A principal component analysis (PCA) is annually conducted on Minnesota's statewide assessments, and results for all grades and subjects of the MCA-III and MTAS-III can be found in the *Yearbook Tables for Minnesota's Statewide Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Yearbook Tables expandable header)) under the section heading "Dimensionality Reports."

Dimensionality results reported in the *Yearbook* include the ratio of the first to the second eigenvalue and the proportion of variance accounted for by the first eigenvalue. Various rules of thumb have been proposed in the research literature to help interpret these measures. Various authors (e.g., Gorsuch, 1983; Morizot et al., 2007) give the rule that if the ratio of the first to second eigenvalue exceeds a value of three, unidimensionality is indicated. As shown in analyses reported in the *Yearbook*, MCA-III and MTAS-III eigenvalue ratios generally always exceed this criterion, implying the tests are unidimensional. Regarding the percent of variance accounted for by the first factor, since the first principal component explains the maximum variance, then the percentage of total variance explained by the first principal component is often regarded as an index of essential unidimensionality. The higher percentage of total variance the first principal component accounts for, the closer the test is to essential unidimensionality.

Reckase (1979) found that good unidimensional ability estimates could be obtained even if the first factor accounts for less than 10% of the variance. However, the rule of thumb he gave for essential unidimensionality was for the first factor to exceed 20% of the total variance because he found that item calibration results could be unstable when the variance accounted for was less than 20%. MCA-III tests generally show the first eigenvalue accounting for 10% to 20% of the total variance, while MTAS-III tests generally show much higher values for percent of total variance accounted for (greater than 30%). Although the MCA-III tests do not always meet Reckase's 20% rule of thumb, they do show that the first factor accounts for a substantial proportion of the variance, and the IRT item drift analyses conducted every year show the MCA-III item calibration results to be stable. Therefore, both the ratio of eigenvalues and the proportion of variance accounted for analyses reported in the *Yearbook* provide support that MCA tests measure an essentially unidimensional composite.

In addition to the PCA, the unidimensional composite for the fixed-form assessments can be investigated at the item level through the point-biserial correlation. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called point-biserial correlation) is the correlation between an item and the total test score. Conceptually, if an item has a high item-total correlation (i.e., 0.30 or above), it indicates that students who performed well on the test tended to answer the item correctly and students who performed poorly on the test tended to answer the item incorrectly; that is, the item discriminated well between high-ability and low-ability students. Assuming the total test score represents the extent to which a student possesses the skills or knowledge being measured by the test, high item-total correlations indicate the items on the test require proficiency in these skills or knowledge to be answered correctly. The *Yearbook* presents item-total correlations in tables under the section heading "Item Statistics Reports," located on the MDE website. For Minnesota's statewide assessments, mean item-total correlations are generally high.

To provide further evidence of the internal structure of the test, correlations among the total test score and subscales are provided. These correlations quantify the relationships among strands (for mathematics and science) and substrands (for reading) and the overall test score. The overall test score is represented by the total scale score for the MCA-III assessment and the total raw score for the MTAS-III assessment. These correlations demonstrate that the factors (strands and substrands) composing the overall test are highly related (as demonstrated through high correlations) to the overall test while also distinct in the factors they are measuring. Put another way, high correlations indicate that the assessment is measuring one underlying construct. As can be referenced in the correlation tables in the *Yearbook*, there are high correlations between the scale score, or the raw score for fixed-form tests, and the strand scores (substrand for reading) for each of the grades, while there are moderate-to-high correlations among the strand (or substrand) scores. The correlation tables are provided in the *Yearbook* for MCA-III Mathematics, Reading, and Science, and MTAS-III Mathematics, Reading, and Science assessments under the section heading "Internal Consistency Reports," located on the MDE website.

The dimensionality analysis examines the number of factors measured by the items, the item-total correlations investigate the consistency of students' performance on an item to their overall test scores, and the correlations among the total scale score (or raw score for fixed-form tests) and the strand (substrand for reading) provide evidence that the strand (or substrand for reading) scores are highly related to the total test score, but less related to each other. Together, these three pieces of evidence collectively demonstrate the structure of the test can be measured using a unidimensional composite.

8.4. Evidence for Different Student Populations

In addition, internal structure evidence should show that individual items are functioning similarly for different demographic subgroups within the population being measured. Minnesota's statewide assessments measure the statewide content standards that are taught to all students. In other words, the tests have the same content validity for all students because what is taught to all students is measured for all students. In addition, all the tests are given under standardized conditions. Great care has been taken to ensure the items in the statewide assessments are fair and representative of the content domain expressed in the content standards. Special attention is given to find evidence that construct-irrelevant content has not been inadvertently included in the test, as such content could result in an unfair advantage for one group versus another. For instance, a test item could contain language or address topics that are more familiar to male students compared to female students. Both judgmental and statistical methods are used to identify and remove such items from use to mitigate their impact on any of the demographic subgroups that make up the population of the state of Minnesota.

As described in Chapter 2: Test Development, this process begins with item writers trained on how to avoid economic, regional, cultural, and ethnic bias when writing items. After items have been written, they are reviewed by a bias and sensitivity committee, which evaluates each item to identify language or content that might be inappropriate or offensive to students, parents, or other community members or that contain stereotypical or biased references to gender, ethnic, or cultural groups. The Community MCA Review Committee accepts, edits, or rejects each item for use prior to the item's initial (field test) administration.

DIF analyses are conducted for the purpose of identifying items that are differentially difficult for different subpopulations of individuals. Refer to Chapter 2: Test Development for more details about DIF and the method used to flag items that function differently. Though DIF analyses flag items as being differentially difficult for one group as compared to another, being flagged through this analysis does not solely provide sufficient evidence for removing the item from use. Flagged items are examined during data review meetings that take place after the initial (field test) administration of each item. Items are removed from use only when the data review committee identifies a concrete reason for having been flagged by the DIF, such as bias or sensitive content.

These multiple reviews are a critical component of the item and test development process. They support the validity of the test for all the diverse populations that make up the state of Minnesota.

8.5. Evidence Based on Relations to Other Variables

The *Standards of Educational and Psychological Measurement* (AERA et al., 2014) highlight that often, the interpretation or use of a particular measure can be validated by comparison to other measures of the same or a related construct. Because the Minnesota tests have been developed with a specific set of standards to be assessed, it is challenging to identify tests that measure the same construct with the same content.

Both convergent and divergent evidence fall under this category. Convergent validity evidence provides validity evidence through high correlations between test scores on the measure of interest and other measures that measure the same, or similar, constructs. Divergent validity provides validity evidence by showing lower correlations between the test score and different constructs.

To provide convergent validity evidence, Minnesota MTAS-III tests are administered by test administrators. These tests are given concurrently with the Learner Classification Inventory for Alternate Assessments on Alternate Achievement Standards (LCI), which collects student LCI data and assistive technology information for entry into the data-entry interface (Kearns et al., 2006). The collection of this data allows for a correlation to be calculated between student factors and their scores on the MTAS-III assessment. If the relationships are convergent, the performance of students on two measures should be highly correlated if they measure a similar construct. In particular, the raw scores of the MTAS-III Mathematics, Reading, and Science assessments were correlated with the LCI Mathematics and LCI Reading variables, which are items on the inventory that summarize the degree to which students can apply mathematical and reading skills. More specifically, the LCI Reading variable measures the degree to which a student is aware of text, can use text, read text, and understand the text. The LCI Mathematics variable measures the degree to which students have an awareness of numbers, can use those numbers, and conduct and apply computations with those numbers. High positive correlations between the LCI Mathematics and LCI Reading variables represent the congruence between the skills measured by the MTAS-III and the observed skills the test administrator observed. These correlations can be found in the *Yearbook* under the section "Correlation of LCI Variables with MTAS-III Scale Scores."

Also, correlations between MTAS-III student scores and two LCI variables, expressive communication and receptive language, are calculated. Expressive communication, as measured by the LCI, represents the degree to which symbolic language and expression are used in communication, while receptive language, as measured by the LCI, represents the degree to which students can follow directions and are alert to sensory stimuli from others. Correlations between the MTAS-III Mathematics, Reading, and Science scores and the LCI variables expressive communication and receptive language provide one further piece of evidence that the scores from the students who take the MTAS-III are correlated to the skills of those students. These correlations can be found in the *Yearbook* under the section "Correlation of LCI Variables with MTAS-III Scale Scores."

In additional support of the validity of the MTAS-III, the relationships among the MTAS-III content areas (mathematics, reading, and science) were investigated. The validity evidence provided by this analysis is derived by comparing the observed relationships to the expected relationships. For instance, validity evidence can be provided if the observed relationships between Mathematics MTAS-III and Reading MTAS-III or Science MTAS-III are consistent with expectations. Results from these analyses are provided in the *Yearbook* under the section "Correlation of LCI Variables with MTAS-III Scale Scores."

8.6. Criterion Validity

Criterion validity relies upon the demonstration of a relationship between the test and an external criterion measure. If the test is intended to measure mathematical ability, for example, then scores from the test should correlate substantially with measures that require mathematical ability to achieve a high score. 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 and the external criterion. For the observed relationship between the assessment and the criterion to provide evidence of a strong relationship, the criterion should measure the same or a similar construct of the assessment. Criterion validity evidence is typically expressed in terms of the product-moment correlation between the test and criterion scores.

There are two types of criterion-related evidence: *concurrent* and *predictive*. The difference between them relates to 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 the relationship between scores from a district-wide assessment (the measure being validated) to those from a nationally recognized college entrance exam (the criterion). In this example, if the results from the district-wide assessment and the nationally recognized college entrance exam were collected in the same semester of the school year and were highly correlated, this would provide concurrent criterion-related evidence of the validity of the district-wide assessment. On the other hand, predictive evidence is collected at different times; typically, the criterion information is obtained subsequent to the administration of the measure being validated. For example, if results from a nationally recognized college entrance exam were being used to predict success in the first year of college, the nationally recognized college entrance exam results would be obtained in the junior or senior year of high school, whereas the criterion—college grade point average (GPA)—would not be available until a year or two later. The correlation of the two would then be a measure of the validity of the exam with respect to its use in predicting first-year college success.

In ideal situations, the criterion-validity approach can provide convincing evidence of a test's validity. However, there are two important obstacles to implementing the approach. First, a suitable criterion must be found. A standards-based test such as the 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 Minnesota Academic Standards may be hard to do without creating yet another test. It would be possible to correlate performance on the MCA-III with other types of assessments, such as a nationally recognized college entrance exam or school assessments, but even though strong correlations with a variety of other assessments would provide some evidence of validity for the MCA-III, the evidence is less compelling if the criterion measures are only indirectly related to the standards. The same can be said of the MTAS-III. Finding a criterion representing proficiency on this assessment is difficult because the sample of students who take this assessment do not typically do take other large-scale assessments that measure ability.

A second obstacle to the demonstration of criterion validity is that the criterion may require validation 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.

Correlational analyses were conducted to investigate the relationship between the MCA-III for high school students and a nationally recognized college entrance examination. Results indicated that there is a strong positive relationship between the MCA-III Mathematics assessment for those in high school and a nationally recognized college entrance examination (r = 0.86); similarly, there was a strong positive relationship between the MCA-III Reading assessment for those in high school and the nationally recognized college entrance examination (r = 0.76). In addition, there was a strong and positive correlation (r = 0.78) between grade 8 MCA-III Mathematics scale scores and the nationally recognized "precollege" entrance examination; similar findings were observed for reading, where the correlation between the grade 8 MCA-III Reading scale scores and nationally recognized "precollege" entrance exam was positive and strong (r = 0.70). Additional criterion-related validity evidence on Minnesota's statewide assessments will be collected and reported on an ongoing basis. 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.

8.7. Additional Validity Evidence

8.7.1. Scoring Validity Evidence

Scoring validity evidence can be divided into two sections: (a) the evidence for the scoring of performance items and (b) the evidence for the fit of items to the model.

8.7.2. Scoring of MTAS-III Items

The auditing of the Minnesota Test of Academic Skills (MTAS-III) administrations and task ratings supplies validity evidence for the scoring of these performance tasks. The auditing procedure is described in Chapter 9: Reliability and results of the audits are provided in the *Yearbook*.

8.7.3. Model Fit and Scaling

IRT models provide a basis for Minnesota's statewide 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 scrutinized before a decision is made to put it on the test. However, most items show adequate item fit. Justification for the scaling procedures used for Minnesota's statewide 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 items that could invalidate scores. One such factor relevant for the MCA-III assessments is the mode of administration. Prior to 2015, the MCA-III was administered either online or on paper, depending on the choice made by the school district. Thus, it was important to evaluate whether mode effects between the two versions of the test could raise validity concerns for the test scores. Since the 2014–15 testing year, the MCA-III has moved to online testing for all mathematics, reading, and science tests; however, students who are eligible can still be administered the paper accommodated version of the test. In spring 2011, a mode-comparability study was conducted using a matched group study design to compare the results of students taking one of the online operational test forms with the results of student taking a similar form given on paper for the MCA-III Mathematics grades 3–8.

The results of the comparability study suggested that, although testing mode was found to affect certain items in common between the online and paper versions, this effect could be mitigated by essentially treating the online and paper versions of the items as distinct items with mode-specific item parameters. The online and paper parameters were scaled to a common metric by using a set of linking items. Because the online and paper administrations are pre-equated and the paper form has been fixed over time, the mode specific parameter estimates are still applicable for the current assessment. The complete MCA-III Mathematics MCA-III Comparability Study Report (2012) can be found on the MDE website (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Other Studies expandable header))

In spring 2013, a mode-comparability study was conducted using matched samples to compare student performance on the MCA-III Reading online and paper assessment modes. The results of the comparability study suggested that there was a mode effect but that it could be resolved by applying the results of a Stocking-Lord equating to place the scores on the same scale. The complete MCA-III Reading comparability report is available upon request from MDE.

In addition, in spring 2014, a mode-comparability study was conducted using matched samples to compare student performance on the MCA-III Mathematics grade 11 online and paper assessment modes. The results of the comparability study suggested that there was a mode effect, but that it could be resolved by applying the results of a Stocking-Lord equating to place the scores on the same scale. The complete MCA-III Mathematics grade 11 comparability report is available upon request from MDE.

Chapter 9: Reliability

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) note that the term *reliability* is typically used in one of two different ways within the field of measurement. The first is within the term *reliability coefficient*, which refers to the "reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form." The second term refers to reliability/precision and refers to the "more general notion of consistency of the scores across instances of the testing procedures."

The *Standards* mention that reliability can be quantified as standard errors, reliability coefficients, generalizability coefficients, error/tolerance ratios, IRT information functions, and various indices of classification consistency as appropriate to the assessment for which the reliability is being measured. When a score is reported for a student, there is an expectation that if the student had taken a different but equivalent version of the test, a similar score would have been achieved. A test that does not meet this expectation (that is, a test that does not measure student ability and knowledge consistently) has little or no value. Furthermore, the ability to measure consistently is a prerequisite to making appropriate interpretations of scores on the measure (i.e., showing evidence of valid use of the results). However, a reliable test is not necessarily a valid test, and a reliable and valid test 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. Therefore, reliability is a necessary, but not sufficient, condition for validity. The concept of test validity was discussed earlier in this document in Chapter 8: Validity.

9.1. 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 follows:

$$Observed\ Score = True\ Score + Error \tag{9.1}$$

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 student 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 test scores are likely to be systematically lower for the entire group of students taking the test than they would have been had the lighting been at adequate levels.

Reliability can be quantified in many ways. One common representation is as the proportion of true score variance relative to observed score variance, that is, the variance of the students' true scores divided by the variance of their observed scores follows:

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2} \tag{9.2}$$

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 variance component in the denominator of the ratio increases and the reliability decreases.

Using assumptions from classical test theory (Equation 9.1 and random error assumptions), an alternative formulation can be derived. This formulation is more closely aligned to the reliability coefficient discussed earlier. Reliability, the ratio of true variance to observed variance, can be shown to equal the correlation coefficient between observed scores on two *parallel* tests. The term *parallel* has a specific meaning: The two tests meet the standard classical test theory assumptions, as well as yield equivalent true scores and error variances. The proportion of true variance formulation and the parallel test correlation formulation can be used to derive sample reliability estimates.

9.2. Estimating Reliability

There are three broad categories of reliability coefficients in classical test theory: (a) test-retest, (b) alternate forms, and (c) internal consistency methods. The test-retest and alternate forms methods both rely on testing students multiple times while internal consistency reliability assesses the degree of reliability through a single administration of a test.

Reliability can vary from one sample to another. As discussed in Chapter 8: Validity, Minnesota has taken a multifaceted approach to providing validity evidence of their assessments to ensure that the assessments are equally valid for different samples of students. This is conducted through DIF analysis and expert knowledge from a committee of individuals representing a diverse cultural knowledge base, as well as content committees and bias and sensitivity committees whose members are familiar with the diversity of cultures within Minnesota. Because different samples can vary in their reliability/precision, separate estimates of reliability are provided for several subgroups of students to which Minnesota administers assessments. Therefore, separate estimates of reliability are provided for the total (overall) group of students, female, male, Asian, Black/African American, Hispanic, American Indian/Alaska native, multi-race, Native Hawaiian/Pacific Islander, and white groups for the current year.

9.2.1. Test-Retest Reliability Estimation

Reliability can be estimated by calculating the correlation coefficient between scores from a test administered on one occasion with scores from the same test administered on another occasion to the same students. Essentially, the test is acting as its own parallel form and the reliability estimate is representing the consistency over replications of the testing procedure. 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's statewide assessments.

9.2.2. 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 in terms of the general distribution of content, item formats, administrative procedures, and population score means and standard deviations. For Minnesota's statewide assessments, alternate forms reliability estimation is not possible because no student takes more than one form of the test during any test administration. Reducing the frequency of testing provides more time for the students in the classroom as well as limits the item pool usage per grade, meaning fewer items must be developed and maintained.

9.2.3. Internal Consistency Reliability Estimation

Internal consistency methods use a single administration to estimate test score reliability. As stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), "[I]t should not be inferred, however, that alternate-form or test-retest coefficients based on test administrations several days or weeks apart are always preferable to internal-consistency coefficients." The reason for this is that invariance over occasions is a reasonable assumption if there is a strong theoretically argument for this to be true. Specifically, Minnesota's statewide assessments are developed to control many of the factors that can influence students test scores. Therefore, internal consistency is the primary reliability estimate used for the fixed-form assessments administered in Minnesota. In addition, 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. The most frequently used internal consistency reliability estimate is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the essentially tau-equivalent measurement model and the formula is as follows:

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} S_{Y_i}^2}{S_X^2}\right) \tag{9.3}$$

where *N* is the number of items on the test, S_Y^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 a point estimate of score reliability, and it may be important to consider the precision of that estimate, particularly when it is based on a small number of data points and/or restriction of range.

Coefficient alpha is calculated for all paper accommodated form assessments, including the Mathematics and Reading MCA-III, as well as the MTAS-III data-entry assessments for mathematics, reading, and science, which are all data-entry forms. 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. Results from the dimensionality analysis are discussed in Chapter 8: Validity. Dimensionally analysis results are provided in the *Yearbook Tables for Minnesota's Statewide Assessments* (MDE website > Districts, Schools and Educators > Teaching and Learning > Statewide Testing > Technical Reports (under the Yearbook Tables expandable header)) under the section "Dimensionality Reports."

Given the results of the dimensionality analysis, it was assumed to be reasonable that the tests were homogeneous and that it was appropriate to compute coefficient alpha. Alpha is based on the total sample of test takers that take the same set of items. Because not all students see the same set of items during the administration of the MCA-III Mathematics and Reading online adaptive assessments, standard measures of reliability are not appropriate. Therefore, coefficient alpha is used as reliability evidence for only the fixed-form assessments, including the MTAS-III assessments, as well as the data-entry forms for the MCA-III Mathematics and Reading. The MCA-III Science assessment, similar to the MCA-III Mathematics and Reading assessments, calculated marginal reliability in lieu of coefficient alpha. Marginal reliability is discussed later in this chapter.

The data-entry forms calculate coefficient alpha for two primary reasons. First, prior to year 2014–15, districts could choose between online or paper forms assessments. Year 2014–15 was the first operational year where all students were required to take the MCA-III assessment online except for those who were eligible to take the data-entry accommodated forms. Because of this change in policy, students taking the data-entry forms for MCA-III tend to have lower scale scores than students taking the online test. Also, the number of students taking the data-entry form version of the MCA-III tends to be quite low in relation to the online test. Because of these reasons, coefficient alpha is an appropriate index to calculate the reliability for the data-entry forms of Mathematics and Reading MCA-III. Second, coefficient alpha tends to be higher for assessments with more items in general, given they are well-written items. In contrast, the coefficient alpha for a single strand level may not be as high as the overall assessment because there are a limited number of items measuring a given strand. The distributional information for the MCA-III online and data-entry forms, as well as the coefficient alpha for the data-entry forms, can be referenced in the *Yearbook* under the section "Summary Statistic Reports." Similarly, for MTAS-III, coefficient alpha is calculated because of the limited number of students in relation to the MCA-III online. Similar to the MCA-III, the coefficient alpha statistics for each grade and subject can be found in the *Yearbook* under the section "Summary Statistics Reports."

Additionally, item-total correlations are computed. A reliable measure should contain items that correlate with the sum of the other items on the measure. An item-total correlation is simply the correlation between each item and the total-raw score after removing the item of interest. The item-total correlations are calculated for the paper accommodated form of the MCA-III Mathematics grades 3–8 and 11; MCA-III Reading grades 3–8 and 10; all fixed-form assessments that include MCA-III Science grades 5, 8, and high school; and all MTAS-III assessments for mathematics, reading, and science. Item-total correlations for each grade and subject are provided in the *Yearbook* under the section "Item Statistics Reports" for both MCA-III and MTAS-III assessments. Item-total correlations are also calculated during key-check processes where item-total correlations are calculated for all MC items for all CAT and linear form assessments. Here the item-total correlation for CAT tests represents the correlation between the item and the total raw score with that item removed (subtracted from the total score); however, for the CAT test the items that comprise the total raw score will vary by student. This statistic represents the relationship between how someone does on the overall test and how they performed on an individual item. By scrutinizing the relationship between individual items on a student's performance on the test, the items included on the test will be more likely to function similarly.

9.2.4. IRT–Based Reliability

Instead of reporting coefficient alpha for the MCA-III online assessments for mathematics, reading, and science, estimates of reliability based on IRT are given. IRT provides a means of estimating reliability that operates on both the individual pattern of responses to items given by students and the statistical characteristics associated with those items. The IRT analogue to classical reliability is called marginal reliability and is calculated using the variance of the theta (ability) scores and the average of the expected error variance. Similar to the decomposition of an observed score in classical test theory, one can decompose the estimated IRT ability into the true ability plus error:

$$\hat{\theta} = \theta + \epsilon, \tag{9.4}$$

where θ is the estimated ability, θ is the true ability, and $\in$ is the error associated with the estimate.
The reliability can then be expressed as follows

$$R_\theta = \frac{var(\theta)}{var(\hat{\theta})} = \frac{var(\hat{\theta}) - var(\epsilon)}{var(\hat{\theta})}. \tag{9.5}$$

The marginal reliability (Green et al., 1984; Thissen and Wainer, 2001) of the reported scale score can then be expressed as follows:

$$Marginal\ Reliability = \frac{\sigma^2_{theta} - \overline{\mathrm{SE}_{\mathrm{theta_i}}^{\ 2}}}{\sigma^2_{theta}}, \tag{9.6}$$

where σ^2_{theta} is the variance of ability scores for the population of interest and SE_{theta_i} is the standard error of the ability estimate of student *i*. Marginal reliability can be calculated by subtracting the average of the squared CSEM (error variance) for each student from the estimated variance of IRT ability scores and dividing by the estimated variance of IRT ability scores. In the case of MCA-III online strand and substrand scores for mathematics, reading, and science, where EAP methods are used to estimate scores, an alternative formula, described by Bock and Mislevy (1982), is used to estimate score reliability that is based on the assumption the ability distribution is distributed normally, *N*(0,1):

$$\mathrm{EAP}\ Marginal\ Reliability = 1 - \overline{PSD^2}, \tag{9.7}$$

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

$$\mathrm{PSD} = \mathrm{Var}(\theta|\mathbf{u}) = \frac{\sum_{k=1}^{q} (\mathrm{X_k} - \theta)^2\, L(\mathrm{X_k}) W(\mathrm{X_k})}{L(\mathrm{X_k}) W(\mathrm{X_k})} \tag{9.8}$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with the quadrature point, and $L(X_k)$ is the likelihood function conditioned at that quadrature point. PSD is equal to the standard error of the EAP estimate for each student in the student data file.

For the MCA-III online assessments in mathematics, reading, and science, the marginal reliability is given for the overall scale score. For these assessments, standardized integer scale scores are reported for the strands, based on a linear transformation of the estimated strand theta (= 5.0 + 2θ_{est}), in place of raw scores. For MCA-III online strand scores for mathematics, reading, and science, the marginal reliability is calculated for the estimated theta score. This result is multiplied by the square of the correlation between strand theta estimate and the reported standardized scale score to reflect the impact resulting from transformation of the theta score to integer scale score values constrained to a one-to-nine range. The modified EAP Marginal Reliability used for reporting purposes is:

$$EAP\ Marginal\ Reliability = (1 - \overline{PSD^2}) * (r^2_{(\theta * ScaleScore)}) \quad (9.9)$$

where $r^2_{(\theta * ScaleScore)}$ is the squared correlation between the student theta estimates and student scale score estimates. Subscore reliability will generally be lower than total score reliability because reliability is influenced by the number of items (as well as their covariation). In some cases, the number of items associated with a subscore is small (10 or fewer). Results involving subscores (and subscore differences in particular) must be interpreted carefully, because these measures have lower reliability associated with them compared to total scores. Marginal reliabilities are provided in the *Yearbook* for the MCA-III online assessments in mathematics, reading, and science under the section "Summary Statistics Reports." Marginal reliabilities provided in the *Yearbook* for the online assessments include reliabilities for the overall population, as well as reliabilities broken down by gender, ethnicity, and accommodated/non-accommodated status. For the purposes of reliability and summary statistic calculations, students taking an online accommodation are grouped together regardless of which accommodation they took because most accommodation types have very small sample sizes.

Reliability information provided in the *Yearbook*, such as for the MCA-III online assessments in mathematics, reading, and science under the section "Summary Statistics Reports," shows that Minnesota's statewide assessments in the current year of administration were reliable measures. The reliabilities of the current year of administration tests are consistent with previous administrations.

9.3. 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. Because Minnesota students are only tested at one point during the testing window each academic year, it is not possible to estimate the standard error through multiple measures. Instead, the SEM can represent a lack of score consistency for the sample of students. 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:

$$SEM = s_x\sqrt{1 - \rho_{xx}}, \quad (9.10)$$

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

9.3.1. 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 SEM band placed around the student's scale score would result in a range of values most likely to contain a student's observed score upon replication. For example, if a student has an observed scale score of 350 on a test having score reliability of 0.84 and a standard deviation of the observed score of 10.0, the SEM would be as follows:

$$SEM = 10.00\sqrt{(1.00 - 0.84)} = 4.00 \quad (9.11)$$

Placing a one-SEM band around this scale score would result in a score range of 346 to 354 (that is, 350 ± 1×4.00). It should be noted that scale scores are rounded to the nearest integer. Furthermore, in the case of unbiased scores and if measurement error is normally distributed, then the true scores for approximately 68% of test takers would fall in the interval band created by adding and subtracting one SEM from their reported score. Thus, 68% of students with an observed score of 350 and SEM = 4 would have a true score within the interval 346–354. This interval is called a *confidence interval* or *confidence band*. By increasing the range of the confidence interval, one improves the likelihood the confidence interval includes the observed score. For example, an interval of 1.96 times the SEMs around the scale score is referred to as a 95% confidence interval. It should be noted that the above interpretation of the likelihood of having a score within a range is only approximate because the confidence interval is constructed around a point estimate and does not have an associated direct probability statement. While it is common practice to use a frequentist confidence band around the observed score and treat such as a probability statement, only Bayesian methods allow for such an interpretation because the score has a probability distribution due to the use of a prior distribution. Here, a score based on a Bayesian procedure (such as the process used with EAP-based strand scores) would have what is denoted as a posterior distribution (e.g., a set of plausible test scores) and *credible* intervals that represent direct probability statements about a score given the observed data.

The SEM for the subscales and total raw score is reported for the Mathematics, Reading, and Science MTAS-III in the *Yearbook* under the section "Summary Statistics Reports" for each respective grade and subject. The overall SEM for each MCA-III test can be calculated with data provided in the *Yearbook*. However, given the use of IRT for all Minnesota's assessments, the conditional SEM (discussed in the next section) is the primary reporting measure of precision associated with each scale score.

9.3.2. 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. Put another way, the true score can be conceptualized as the average of an infinite number of testing replications. Hypothetically, if a student was exposed to an infinite number of testing replications the error in measurement would be normally distributed with a mean equal to the true score and a variance equal to the standard error. Therefore, the SEM for a particular true score is defined as the standard deviation of the observed scores of students with that true score. This standard deviation of the observed score 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 because of measurement error. The conditional standard deviation defines that 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, the IRT model allows for the CSEM to be estimated for any test where the IRT model holds.

For assessments scored by a transformation of raw score-to-scale score table, such as Science MCA-III or Mathematics, Reading, and Science MTAS-III, the equation of CSEM for the test level scale score is as follows:

$$CSEM(O_X|\theta) = \sqrt{[\sum_{X=0}^{Max\ X} O_X^2 p(X|\theta)] - [\sum_{X=0}^{Max\ X} O_X p(X|\theta)]^2}, \quad (9.12)$$

where 0_x is the observed (scaled) score for a particular number-correct score X, θ is the IRT ability scale value conditioned on, and $P^{(*)}$ is the probability function.p^(X|θ) is computed using a recursive algorithm given by Thissen et al. (1995). Their algorithm is a polytomous generalization of the algorithm for dichotomous items given by Lord and Wingersky (1984). The values of θused 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 MCA-III for Science, the strand level CSEM in the theta scale is calculated as follows:

$$CSEM(\theta|x) = \sqrt{\frac{\sum(Q-EAP(\theta|x))^2 L_x(Q)\phi(Q)}{\sum L_x(Q)\phi(Q)}}, \quad (9.13)$$

where *Q* is a quadrature distribution, $EAP(\theta|x)$ is the EAP strand theta for which the CSEM is being estimated, x is the summed score, $L_x(Q)$ is the likelihood for summed score *x* (via the Lord-Wingersky [1984] recursion) for quadrature point *Q*, and $\boldsymbol{\phi}(Q)$is the standard normal distribution for the quadrature distribution *Q*, normalized to sum to one. For this application, the quadrature distribution *Q* ranges from −5.0 to 5.0, with intervals of 0.1. The CSEM that results from this calculation is in the theta metric, and thus should be multiplied by two to place it in the strand scale score metric as shown in the following equation:

$$CSEM(Scale_i) = 2 * CSEM(\theta|x). \quad (9.14)$$

MTAS-III scale scores are reported in their raw score metric and no CSEMs are reported.

For the Mathematics and Reading MCA-III, which employ pattern scoring based on the 3PL measurement model, the CSEM of student *i*'s scale score for the CSEMs of the on-grade test-level scale score, strand/substrand scale score, and progress score is calculated from the CSEM of the obtained θ_i estimate:

$$CSEM(Scale_i) = Spread * CSEM(\theta_i). \quad (9.15)$$

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

$$CSEM(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{9.16}$$

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

$$I(\theta_i) = \sum_{j=1}^{N} \frac{P'_{ij}(\theta_i)^2}{P_{ij}(\theta_i)(1-P_{ij}(\theta_i)} \tag{9.17}$$

where *N* is the number of items on the assessment, P_{ij} is the probability of student *i* answering item *j* correctly, and $P'_{ij}(\theta)$ is the first derivative of P_{ij} with respect to θ. Note that the calculation depends both upon the unique set of items each student answers and his or her estimated ability level (θ). Therefore, different students will likely have different CSEM values even if they have the same raw score and/or theta estimate. Each item contains a unique amount of information for a given θ, which depends on each item's discrimination, difficulty, and pseudo-guessing parameters. Therefore, the IRT estimate of CSEM depends on the specific set of items administered during the assessment. The more information across items for a given θ, the lower the conditional standard error of measurement will be.

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

Mean CSEMs are provided in the *Yearbook* for each subject and grade under the section "Frequency Distribution Reports." These values are reported for each scale score for MCA-III Mathematics and Reading and are reported by each raw score and scale score for MCA-III Science as well as MTAS-III Mathematics, Reading, and Science. 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.

Additionally, under the section "Measurement Precision Comparison with Previous Administration," the *Yearbook* provides a comparison of the measurement precision of MCA assessments from the current administration to that of the previous administration. To make this comparison, CSEM values are averaged within each decile of proficiency level, with the distribution of proficiency based on the student population. The comparison with the previous administration allows one to gauge how well the measurement precision across the scale is maintained across years. MCA assessments generally show similar measurement precision levels across the proficiency scale, or slight improvements due to item pools becoming larger and more robust across time. For MCA-III Mathematics and Reading, an additional comparison is presented showing CSEM values from when the main operational test was administered as fixed form. The *Yearbook* CSEM comparisons show that transitioning the MCA-III Mathematics and Reading assessments to CAT resulted in improved measurement precision.

9.3.3. 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 are. Much of the error owing to systematic factors (i.e., bias) can be avoided with a well-designed assessment instrument that is administered under appropriate and standardized conditions. The remaining random error present in any assessment cannot be fully eliminated, but for groups of students, random error is apt to cancel out (i.e., average to zero). Some students score a little higher than their true score, while others score a little lower. The larger the number in the group, the more the canceling of errors tends to occur. The degree of confidence in the average score of a group is generally greater than for an individual score.

9.3.4. 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 because of 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. 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 follows:

$$s_e = \frac{s_x}{\sqrt{N}}, \tag{9.18}$$

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 the following:

$$s_e = \frac{10}{\sqrt{20}} = 2.2 \tag{9.19}$$

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

9.4. Auditing of MTAS-III Administrations and Task Ratings

Reliability evidence primarily focuses on the amount of error involved in measurement. In an assessment such as the MTAS-III, where the test administrator scores performance tasks, an additional source of measurement error can come from the test administrator. To minimize the measurement error in the MTAS-III, Minnesota test administrators strictly adhere to the procedures for administering and scoring the assessment. Because many students taking the MTAS-III have unique communication styles that require significant familiarity with the student to understand their intended communication, the MTAS-III performance tasks are prepared, administered, and scored by educators familiar with the student. To show that the test administrators are correctly following the standardized guidelines for test administration and scoring, rater agreement can be used as one form of reliability evidence. Minnesota conducts rater audits on test administrators for the MTAS-III. The

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-III and visited several randomly selected schools to observe the test administration and scoring of actual assessments. The auditors also interviewed the local educators to get their opinions on the ease of preparing and administering the test. The auditors' agreement percentages between their own ratings and those of the test administrator as well as counts of the number of audits are provided in the *Yearbook* under the section "Field Auditor Results."

9.5. Classification Consistency

Every test administration will result in some error in classifying students. The *Standards for Educational and Psychological Testing* (AERA et al.) recommends reporting *decision accuracy*, or the "extent to which observed classification of students based on the results of a single replication would agree with their true classification status." The concept of the 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, because of 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 Section 9.3: Standard Error of Measurement, 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 borderline students whose true scores lie close to achievement level cut scores.

For the MCA-III and the MTAS-III assessments, the levels of achievement are *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, and *Exceeds the Standards*. The analysis of classification consistency is described below.

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

The method followed is based on the work of Rudner (2005). An assumption is made that for a given (true) ability score θ, the observed score $\hat{\theta}$ is normally distributed with a mean of θ and a standard deviation of $SE(\theta)$ (i.e., the CSEM at θ). Using this information, the expected proportion of students with true scores in any 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:

$$P(Level_k) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{SE(\theta)}\right) - \phi\left(\frac{a-\theta}{SE(\theta)}\right) \right) f(\theta), \tag{9.20}$$

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, the observed distribution of theta estimates (and observed achievement levels) is used to represent the true theta score (and achievement level) distribution. Based on that distribution, Equation 9.20 is used to estimate the proportion of students at each achievement level who are expected to be assigned to each possible achievement level. To compute classification consistency, the percentages are computed for all cells of a true-versus-expected achievement level cross-classification table. The diagonal entries within the table represent agreement between true and expected classifications of students. The sum of the diagonal entries represents the decision consistency of classification for the test.

Table 9.1 presents 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 9.20. In this example, total decision consistency is 81.0% (sum of diagonal elements), while the cell showing 9.9% shows the percentage of students who were correctly classified as *Does Not Meet the Standards*. Similarly, 1.3% of students were incorrectly classified as *Does Not Meet the Standards*, when their true score indicates *Partially Meets the Standards*.

Table 9.1. Example Classification Table

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

Note. D = *Does Not Meet the Standards*, P = *Partially Meets the Standards*, M = *Meets the Standards*, E = *Exceeds the Standards*

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 the accountability purposes. To compute decision consistency in this case, the table is dichotomized by combining cells associated with *Does Not Meet the Standards* with *Partially Meets the Standards* and combining *Meets the Standards* with *Exceeds the Standards*. For the example table above, this results in a classification accuracy of 92.2%. The percentage of students incorrectly classified as *Partially Meets the Standards* or lower, when their true score indicates *Meets the Standards* or above, is 2.3%.

The *Yearbook* contains tables with the overall classification accuracy for each grade and subject of MCA-III and MTAS-III under the section *Classification Accuracy Reports*.

Chapter 10: 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 service provider has developed and refined a set of quality procedures to help ensure that all MDE's testing requirements are met or exceeded. These quality-control procedures are detailed in the paragraphs that follow. In general, Minnesota's testing service provider's commitment to quality is evidenced by initiatives in two major areas: (a) task-specific quality standards integrated into individual processing functions and services and (b) a network of systems and procedures that coordinates quality across processing functions and services.

10.1. Quality Control for Test Construction

Following the test construction process described in Chapter 2: Test Development, items are selected and placed on a particular pre-equated test form to provide a strictly parallel form across years 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. 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 service provider and MDE agree to a final pre-equated form. Similarly, the baseline raw score-to-scale-score tables are used as the target tables that the pre-equated test form (under construction) should match. This form is provided to Minnesota's testing service provider for form construction and typesetting.

10.2. Quality-Control Non-scannable Documents

Minnesota's testing service provider follows a meticulous set of internal quality standards to ensure high-quality printed products. Specific areas of responsibility for staff involved in materials production include monitoring all materials-production schedules to meet contract commitments, overseeing the production of test materials, coordinating detailed printing and post-printing specifications, outlining specific quality control requirements for all materials, and conducting print reviews and quality checks. The quality production and printing processes follow printers' reviews and quality checks. Project Management and Print Procurement staff work closely with the printers during the print production phase. Press proofs are checked to ensure high-quality printing and to verify adherence to printing specifications. The printing staff randomly pulls documents throughout the print run for quality control inspections.

10.3. Quality Control for Online Test Delivery Components

Each release of every Online Test Delivery goes through a complete testing cycle, including regression and performance testing. The system goes through user acceptance testing (UAT). During UAT, Minnesota tests that will be administered on that particular release will be used.

The testing service provider also conducts production validation testing in which they publish the Minnesota tests in a production environment and recommend test scenarios. The tests are completed and scoring deliverables are generated during this period, including preliminary Student Detail Reports and the student data files. The validation process includes confirmation of the tests published and the scoring deliverables. Approvals are required at the close of the production validation period prior to the opening of the testing window.

For changes required during the testing window, a patch build is implemented. Release notes are provided that include the fixes made and/or system upgrades. The patch is tested and approved before it is scheduled to be deployed to the field. Only patch builds that are relevant to Minnesota are applied to its pipeline. The deployments are scheduled outside of the regular testing window timeframes.

10.4. 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 service provider, Minnesota's quality-control vendor, and MDE's Division of Academic Standards, Instruction and Assessment use several quality-control procedures to ensure that this equating is accurate:

- Minnesota's testing service provider and MDE perform a "key check" analysis for the MC item type to ensure the appropriate scoring key is being used. The content staff at both the service provider and MDE review the flagged items. If there are any miskeys for the operational and field test items, the correct keys and students score will be updated.
- Minnesota's testing service provider performs an "adjudication" analysis for the TE item types. The adjudication process includes a check of all responses given by students in the current administration to ensure all possible responses are scored appropriately and functionalities of the TE items perform correctly.
- For all assessments, a drift analysis is conducted by Minnesota's testing service provider, Minnesota's quality-control vendor, and MDE's Division of Academic Standards, Instruction and Assessment to determine whether the IRT item parameters have shifted over time. Mathematics and Reading items that have shifted are investigated and a resolution whether to keep or remove an item is made, whereas the drifted Science items are recalibrated.
- The field test analyses are conducted by Minnesota's testing service provider, Minnesota's quality control vendor, and MDE's Division of Academic Standards, Instruction and Assessment to bring the field test items onto the MCA-III measurement scale.

Glossary of Terms

Achievement Level Descriptors (ALDs). ALDs provide descriptive information of what typical students at each achievement level are expected to know of the Minnesota Academic Standards. ALDs appear as Performance Level Descriptors on the Individual Student Reports (ISRs).

Achievement Levels. MCA-III has four achievement levels: *Exceeds the Standards* (proficient), *Meets the Standards* (proficient), *Partially Meets the Standards* (not proficient), and *Does Not Meet the Standards* (not proficient). Students are assigned an achievement level based on their scale score. MTAS-III also has four achievement levels: *Exceeds the Alternate Achievement Standards* (proficient), *Meets the Alternate Achievement Standards* (proficient), *Partially Meets the Alternate Achievement Standards* (not proficient), and *Does Not Meet the Alternate Achievement Standards* (not proficient).

Adequate Yearly Progress (AYP). The amount of progress required by schools each year to meet established federal standards-based accountability goals. The specific progress required is negotiated by the state.

Assessment. The process of collecting information to support decisions about students, educators, programs, and curricula.

Career and College Readiness (CCR). For the high school Reading and Mathematics MCA-III, a graphical representation of a student's "progress" score compared to the CCR goal score. CCR goal scores are identified by directly linking scale scores on these tests to scores on the corresponding subject-level subtests from a nationally recognized college entrance exam. At each grade, CCR goal scores are indicators that performance is on track to demonstrate career and college readiness on a college entrance exam at the end of grade 11. A high school student's MCA-III scale score for a subject is on the same scale as the CCR goal score for that subject and can be interpreted for performance comparison. If a student's MCA-III scale score is at or above the CCR goal score, he or she is expected to be able to successfully complete credit-bearing coursework at a two- or four-year college or university or other credit-bearing postsecondary program without any need for remediation. Student scores below the CCR goal score may indicate that the student's performance is not on track to meet career and college readiness, and the student may benefit from remediation. CCR goal scores are not reported for Science.

Classification Accuracy. The degree to which the assessment accurately classifies students 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.

Computer Adaptive Testing (CAT). A mode of test delivery where each item (or testlet) is adaptively selected for administration on a test based on a test taker's currently estimated ability level, estimated from the prior items in the test.

Content Standards. Content standards describe the goals for individual student achievement, specify what students should know, and specify what students should be able to do in identified disciplines or subject areas.

Content Validity. Evidence that the test items represent the content domain of interest.

Differential Item Functioning (DIF). A term applied to investigations of test fairness. Explicitly defined as difference in performance on an item or task between a designated minority and majority group, usually after controlling for differences in group achievement or ability level.

Elementary and Secondary Education Act (ESEA). Originally, an act of 1965, amended by the No Child Left Behind Act (NCLB) in 2002, which increased accountability and statewide assessment requirements. Recently, ESEA has granted flexibility to some of the specific requirements of this act to Minnesota in exchange for a comprehensive plan detailing a commitment to implementing higher standards, a plan for improved state and district accountability and support for all students, and a plan to support effective instruction and leadership.

Every Student Succeeds Act (ESSA). In December 2015, the Every Student Succeeds Act (ESSA) was signed into law, which replaced No Child Left Behind (NCLB) and changed many portions of Elementary and Secondary Education Act (ESEA). MDE will work closely with the U.S. Department of Education to ensure Minnesota's students, educators, schools, and districts experience a clear and orderly transition to the new law. The 2018–19 school year was the first full year of ESSA implementation.

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.

Lexile® Measure. The predicted Lexile measure of the student's reading ability, and the upper and lower range that helps match the student with literature appropriate for his or her reading skills. Available for Reading MCA-III only.

Longitudinal Reports. Longitudinal reports allow districts to analyze trends and patterns over time and provide an analysis of results from a specific administration, from multiple administrations within a year, or from year to year. Longitudinal reports are available only in PearsonAccess Next.

Modifications. Changes made to the content and performance expectations for students.

MTAS-III Scoring Rubric. The 0–3 rubric used by the test administrator administering the test to score MTAS-III tasks.

No Child Left Behind (NCLB). Federal law enacted in 2001 that requires school districts to be held accountable to receive federal funding. Under this law, every state was required to create a plan that involved setting performance targets so that all students would be academically proficient by the year 2013–14.

On-Demand Reports. On-Demand Reports are preliminary test results that are available within 60 minutes after testing is completed. On-Demand Reports are available for all online assessments and student responses from paper accommodated test materials entered into data-entry forms in TestNav for MCA-III, but they are not available for MTAS-III. On-Demand Reports are available in PearsonAccess Next.

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.

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 items correct receive the same score, in pattern scoring students rarely receive the same score, as even students getting the same number correct typically differ in the particular items they get correct or incorrect. Because pattern scoring uses information from the entire student response pattern, this type of scoring produces more reliable scores than does number-correct scoring.

***P*-Value.** A classic item-difficulty index that indicates the proportion of students who answered an item correctly.

Quantile® Measure. The predicted Quantile measure of the student's mathematical ability, and the upper and lower range that helps match the student with mathematical concepts appropriate for his or her mathematics skills. Available for Mathematics MCA-III only.

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.

Scale Score. For MCA-III: A score that takes the student's item response pattern (Reading and Mathematics MCA-III) or raw score (Science MCA-III) and adjusts it for possible differences in test difficulty from one year to the next. For MTAS-III: A score that takes the student's raw score and adjusts it for possible differences in test difficulty from one year to the next.

Standards. The MCA-III and MTAS-III are based on the most recent academic content standards in Mathematics, Reading, and Science. The MCA-III and MTAS-III assessments are the statewide tests that help districts measure student progress toward Minnesota's academic standards. The academic standards are revised according to a schedule set forth by statute. Two or three years after standards are revised and adopted, a new series of assessments is ready for operational administration.

Standard Error of Measurement (SEM). 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.

Student Progress Score. A student scale score is converted to a student progress score that translates across grade levels.

Test-Centered Standard Setting Methods. A type of process used to establish performance standards that focus on the content of the test itself. A more general classification of some judgmental standard setting procedures.

Testlet. On the online MCA-III Reading assessment, a testlet is defined as a group of one or more passages and associated items. Each testlet is an adaptive stage in the test, where the adaptive algorithm selects the next testlet to administer based on how the student performed on the item from previous testlets.

Test-Retest Reliability Estimate. A statistic that represents the correlation between scores obtained from one measure when compared with scores obtained from the same measure on another occasion.

Test Specifications. Specific rules and characteristics guide the development of a test's content and format. They indicate which strands, substrands, standards, and benchmarks will be assessed on the test and in what proportions. The test specifications are a helpful tool for developing tests and documenting content-related validity evidence.

True Score. The piece of an observed student score that is not influenced by error of measurement. The true score is used for convenience in explaining the concept of reliability and is unknown 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

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* (AERA et al., 2014) is one source of professional standards. As evidence of our dedication to fair testing practices, the table of contents for this manual is annotated below, outlining the manual's alignment with the *Standards*.

References

Abedi, J., and Ewers, N. (2013). *Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Prepared for SBAC by the University of California, Davis.

Acosta, B. D., Rivera, C., and Shafer Willner, L. (2008). *Best practices in state assessment policies for accommodating English language learners: A Delphi Study*. The George Washington University Center for Equity and Excellence in Education.

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

Barton, K. (2002). Stability of constructs across groups of students with different disabilities on a reading assessment under standard and accommodated administrations [Doctoral dissertation, University of South Carolina, 2001]. *Dissertation Abstracts International*, 62/12, 4136.

Beattie, S., Grise, P., and Algozzine, B. (1983). Test modifications and minimum competency test performance of learning disabled students. *Learning Disability Quarterly*, *6*, 75–77.

Bennett, R., Rock, D., and Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and non-handicapped groups. *Journal of Special Education*, *21*(3), 9–21.

Bennett, R. E., Rock, D. A., and Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement*, *24*(1), 41–55.

Bennett, R. E., Rock, D. A., and Novatkoski, I. (1989). Differential item functioning on the SAT-M braille edition. *Journal of Educational Measurement*, *26*(1), 67–79.

Blaskey, P., Scheiman, M., Parisi, M., Ciner, E., Gallaway, M., and Selznick R. (1990). The effectiveness of Irlen filters for improving reading performance: A pilot study. *Journal of Learning Disability*, *23*(10), 604–612.

Bock, R. D., and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Bolt, S. K., and Thurlow, M. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education*, *25*(3), 141–154.

Bouck, E., and Bouck, M. (2008). Does it add up? Calculators as accommodations for sixth grade students with disabilities. *Journal of Special Education Technology*, *23*(2), 17–32.

Brennan, R. L. (2004). *Manual for LEGS , version 2.0*. Center for Advanced Studies in Measurement and Assessment (CASMA) Research Report No. 3. education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-research-report-3.pdf

Browder, D. M., Gibbs, S., Ahlgrim-Delzell, L., Courtade, G., Mraz, M., and Flowers, C. (2009). Literacy for students with significant cognitive disabilities: What should we teach and what should we hope to achieve? *Remedial and Special Education*, *30*, 269–282.

Brown, D. W. (2007). *The role of reading in science: Validating graphics in large-scale science assessment*. Unpublished dissertation.

Burch, M. (2002). Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities [Doctoral dissertation, Vanderbilt University, 2002]. *Dissertation Abstracts Internationa*l, 63/03, 902.

Burk, M. (1998, October). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities* [Paper presentation]. Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century, Office of Special Education Program Cross Project Meeting, US Department of Education, Washington, DC.

Cai, L., Thissen, D., and du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Scientific Software International.

Calhoon, M., Fuchs, L., and Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, *23*, 271–282.

Castellon-Wellington, M. (2000). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests*. (CSE Technical Report No. 524). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Christensen, L. L., Braam, M., Scullin, S., and Thurlow, M. L. (2011). *2009 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 83). University of Minnesota, National Center on Educational Outcomes.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Erlbaum.

Coleman, P. J. (1990). Exploring visually handicapped children's understanding of length (math concepts). [Doctoral dissertation, The Florida State University, 1990]. *Dissertation Abstracts International*, 51, 0071.

Cormier, D. C., Altman, J. R., Shyyan, V., and Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007–2008* (Technical Report 56). University of Minnesota, National Center on Educational Outcomes.

Crawford, L., and Tindal, G. (2004). Effects of a student read-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality*, *12*(2), 71–88.

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 and H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.

DiCerbo, K., Stanley, E., Roberts, M., and Blanchard, J. (2001, April). *Attention and standardized reading test performance: Implications for accommodation* [Paper presentation]. Annual Meeting of the National Association of School Psychologists, Washington, DC.

Dolan, R., Hall, T., Banerjee, M., Chun, E., and Strangman, N. (2005). Universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment*, *3*(7). napoleon.bc.edu/ojs/index.php/jtla/article/view/1660

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test and score fairness (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04). Princeton, NJ: Educational Testing Service.

Dorans, N. J., and Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series, 1992*(1), i–40.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335–337.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*, 218–229.

Elliot, S., Kratochwill, T., McKevitt, B., and Malecki, C. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly*, *24*(4), 224–239.

Ferrara, S., Perie, M., and Johnson, E. (2002). *Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure* [Invited colloquium]. Board on Testing and Assessment of the National Research Council, Washington, DC.

Fletcher, J., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., and Vaughn, S. (2006). Effect of accommodations on high-stakes testing for students with reading disabilities. *Council for Exceptional Children*, *72*(2), 136–150.

Fletcher, J., Francis, D., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C., Kalinowski, S., Young V., and Vaughn, S. (2009). Effects of a Bundled Accommodations package on high-stakes testing for middle school students with reading disabilities. *Exceptional Children*, *75*(4), 447–463.

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., and Karns, K. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, *29*(1), 65–85.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Grise, P., Beattie, S., and Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research*, *76*(1), 35–40.

Hambleton, R., and Plake, B. (1997). *An anchor-based procedure for setting standards on performance assessments* [Paper presentation]. Annual Meeting of the American Educational Research Association, Chicago, IL.

Hambleton, R. K., and Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.

Helwig, R., Rozek-Tedesco, M. A., and Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *Journal of Special Education*, *36*(1), 39–47.

Holland, P. W., and Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (RR-85-43). Educational Testing Service

Holland, P. W., and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Brown (Eds.), *Test validity* (pp. 129–145). Erlbaum.

Iovino, I., Fletcher, J., Breitmeyer, B., and Foorman, B. (1996). Colored overlays for visual perceptual deficits in children with reading disability and attention deficit/hyperactivity disorder: Are they differentially effective? *Journal of Clinical and Experimental Neuropsychology*, *20*(6), 791–806.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). 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*, *14*(4), 16–20.

Johnson, E. S., Kimball, K., and Brown, S. (2001). American Sign Language as an accommodation during standards-based assessments. *Assessment for Effective Intervention*, *26*(2), 39–47.

Johnson, E. S., Kimball, K., Brown, S., and Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high-stakes assessments. *Exceptional Children*, *67*(2), 251–264.

Kearns, J., Kleinert, H., Kleinert, J., and Towles-Reeves, E. (2006). *Learner characteristics inventory*. University of Kentucky, National Alternate Assessment Center.

Kim, S., and Kolen, M. (2004). *STUIRT* [Computer software]. Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., and Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Erlbaum.

Kolen, M. J. (2004). *POLYEQUATE* [Computer software]. Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Kolen, M. J., and Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer-Verlag.

Kopriva, R., Emick, J., Hipolito-Delgado, C., and Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improvised decision making on scores for English language learners*. Educational Measurement: Issues and Practice*, *26*(3), 11–20.

Koretz, D., and Barton, K. (2003, 2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment*, *9*(1and2), 29–60.

Koretz, D., and Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity*. Educational Evaluation and Policy Analysis*, *22*(3), 255–272.

Lewis, D. M., Mitzel, H. C., and Green, D. R. (1996, June). *Standard setting: A bookmark approach* [Paper presentation]. D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring* [Symposium]. Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linacre, J. M. (2006). *A user's guide to WINSTEPS*. MESA Press.

Linn, R. L. (1993). Linking results in distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

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

Lord, F. M., and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *8*, 453–461.

MacArthur, C. A., and Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *Journal of Special Education, 21*(3), 22–42.

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

McKevitt, B. C., and Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review*, *32*(4), 583–600.

Meloy, L., Deville, C., and Frisbie, D. (2000, April). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA.

Minnesota Department of Education (MDE). (2011). *Standard setting technical report for Minnesota assessments: Mathematics MCA-III, Mathematics MCA-Modified, Mathematics MTAS, Reading MCA-Modified*. Pearson. education.mn.gov/MDE/dse/test/Tech/

Minnesota Department of Education (MDE). (2012). *Minnesota assessments summer 2012 standard setting: Recommended performance standards in grades 5, 8, and high school science*. American Institutes for Research (AIR). education.mn.gov/MDE/dse/test/Tech/

Minnesota Department of Education (MDE). (2013). *Minnesota assessments summer 2013 standard setting: Recommended performance standards for series-III reading assessments*. American Institutes for Research (AIR). education.mn.gov/MDE/dse/test/Tech/

Minnesota Department of Education (MDE). (2014). *Minnesota assessments summer 2014 standard setting: Recommended performance standards for series-III mathematics assessments*. American Institutes for Research (AIR). education.mn.gov/MDE/dse/test/Tech/

Minnesota Department of Education (MDE). (2018). *Minnesota career and college readiness (CCR) summary report for the Minnesota Comprehensive Assessment (MCA)*. Pearson.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Educational Testing Service, Policy Information Center.

Morizot, J., Ainsworth, A. T., and Reise, S. P. (2007), Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, and R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). Guilford Press.

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 and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). Springer-Verlag.

Pennock-Roman, M., and Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*(3), 10–28.

Pennock-Roman, M., and Rivera, C. (2012). *Smarter Balance Assessment Consortium: Summary of literature on empirical studies of the validity and effectiveness of test accommodations for ELLs: 2005–2012*. Prepared for Measured Progress by The George Washington University Center for Equity and Excellence in Education.

Perez, J. V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation [Doctoral dissertation, University of South Florida, 1980]. *Dissertation Abstracts International*, 41, 0206.

Ray, S. R. (1982). Adapting the WISC-R for deaf children. *Diagnostique*, *7*, 147–157.

Raymond, M., and Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–158). Erlbaum.

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

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Erlbaum.

Robinson, G., and 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, 621–626.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research and Evaluation*, *10*(13).

Russell, M. (2006). *Technology and Assessment: The Tale of Two Interpretations*. Information Age.

Russell, M., Kavanaugh, M., Masters, J., Higgins, J., and Hoffmann, T. (2009). Computer based signing accommodations: Comparing a recorded human with an avatar. *Journal of Applied Testing Technology*,*10*(3).

Salend, S. (2009). Using technology to create and administer tests. *Teaching Exceptional Children*, *41*(3), 40–51.

Sato, E., Rabinowitz, S., Worth, P., Gallagher, C., Lagunoff, R., and McKeag, H. (2007). *Guidelines for ensuring the technical quality of assessments affecting English language learners and students with disabilities: Development and implementation of regulations* (Assessment and Accountability Comprehensive Center Report). WestEd.

Scarpati, S., Wells, C., Lewis, C., and Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *Journal of Special Education*, *45*(1), 54–62.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., and Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, *11*(2), 105–126.

Shin, C. D., and Chien, Y. (2017). Conditional randomesque method for item exposure control in CAT. *International Journal of Intelligent Technologies and Applied Statistics*, *10*(3), 144–155.

Shin, C. D., Chien, Y., Way, W. D., and Swanson, L. (2009). Weighted penalty model for content balancing in CATS. *Pearson Papers*, 1–17.

Sireci, S. G., Li, S., and Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). School of Education, University of Massachusetts, Amherst.

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

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, 37, 58–75.

Sullivan, P. M. (1982). Administration modifications on the WISC-R Performance Scale with different categories of deaf children. *American Annals of the Deaf*, *127*(6), 780–788.

Thissen, D., Pommerich, M., Billeaud, K., and Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

Thissen, D., and Wainer, H. (Eds.). (2001). *Test scoring*. Erlbaum.

Thurlow, M., and Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41). National Center on Educational Outcomes, University of Minnesota.

Thurlow, M., House, A., Boys, C., Scott, D., and Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 Update* (Synthesis Report 33). National Center on Educational Outcomes, University of Minnesota.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., and Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children*, *64*(4), 439–450.

Tippets, E., and Michaels, H. (1997, April). *Factor structure invariance of accommodated and non-accommodated performance assessments* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Towles-Reeves, E., Kearns, J., Kleinert, H., and Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *Journal of Special Education*, *42*(4), 241–254.

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

Walz, L., Albus, D., Thompson, S., and Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students* (Minnesota Report 34). University of Minnesota, National Center on Educational Outcomes.

Webb, N. L. (1999) *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). University of Wisconsin–Madison, National Institute for Science Education.

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., and Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*(3), 214–231.

WIDA. (2012). *2012 amplification of the English language development standards, kindergarten–grade 12*. Board of Regents of the University of Wisconsin System, on behalf of WIDA.

Wolf, M. K., Kim, J., Kao, J. C., and Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116.

Wright, N., and Wendler, C. (1994, April). *Establishing timing limits for the new SAT for students with disabilities* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.

Zieky, M. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Erlbaum.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i–30.

Appendix A: Benchmark Report Calculations Resource

This appendix serves as a supplement to the Benchmark Report Interpretation Guide, Benchmark Report "How To" Quick Guides (by subject), and the Understanding the Benchmark Report Video. The purpose of this appendix is to provide detail about the benchmark performance calculation methods. Benchmark reports are provided for public schools and districts in the state for each measured benchmark on the Reading, Mathematics, and Science MCA assessments. Benchmark performance within a school or district is reported by comparing the average benchmark performance for students within the organization to the benchmark performance expected of students who perform at the *Meets* achievement level.

Table A.1 presents the indicator symbols used on the benchmark reports. Schools or districts with fewer than 20 student responses on a particular benchmark are not given a performance indicator for that benchmark due to the small sample size.

Table A.1. Performance Indicator Symbols Used on Benchmark Reports

Performance Indicator on a Benchmark	Symbol
Less than *Meets*	
Similar to *Meets*	
Greater than *Meets*	
There were fewer than 20 student responses for a benchmark and results are not available.	*

Note. Benchmark performance indicators and symbols do not correspond to overall achievement (i.e., *Does Not Meet*, *Partially Meets*, *Meets*, or *Exceeds the Standards*), and the color/shape of each marker does not reflect benchmark difficulty.

A.1. Performance Indicator Calculations

Calculations used to determine benchmark performance indicators are described below. All calculations are performed separately by grade, subject, and organization.

A.2. Student Data

Student test data from the current administration of Reading, Mathematics, and Science MCA assessments are included in benchmark indicator calculations. The calculations use data from public school students with valid test scores. Student data from both online and paper (data-entry) tests are included in the calculations. Benchmark reports are provided to all schools and districts regardless of the number of students within the organization. However, the school/district must have at least 20 student responses for items in a particular benchmark to calculate the performance indicator on a benchmark. If there are fewer than 20 student responses to items within a benchmark, the school/district receives an asterisk (*) for their performance indicator on that benchmark (see Table A.1).

A.3. Observed Performance Measure

School or district benchmark performance is measured by finding the observed average probability correct (*p*-value) for all students in organization *o* across all items measuring a particular benchmark *b*. The calculation to find the observed *p*-value (OBS) for organization *o* on benchmark *b* is made as follows:

$$OBS_{ob} = \frac{\sum_{i \in b} \sum_{s} u_{is}}{\sum_{i \in b} N_{io}}, \tag{A.1}$$

where N_{io} is the number of students administered item *i* in organization *o*, u_{is} is the item score (0, 1) for student *s* on item *i*, the summations in the numerator are across all students in organization *o* and across all items measuring *b*, and the summation in the denominator is across all items measuring benchmark *b*.

A.4. Expected Performance Measure

The actual items administered to students may vary from school to school or district to district. This is particularly true for the online Reading and Mathematics MCA CAT assessments. Therefore, the observed performance measure of an organization needs to be compared to a level of performance that would be expected based on the actual items administered to that organization. A range, called the expected *Meets* range, is calculated based on the expectation of how students performing at the *Meets* achievement level would perform on the items that were administered to the school or district.

The first step in finding the expected range for a given benchmark and organization is to find the lower bound and upper bound of the expected *Meets* range for each item *i* in the pool. These are found using the following formulas:

$$\text{LB}_i = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i (\theta_M - CSEM - b_i)}}, \tag{A.2}$$

$$\text{UB}_i = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i (\theta_M + CSEM - b_i)}}, \tag{A.3}$$

where LB is the lower bound, UB is the upper bound, θ_M is the theta cut score for the *Meets* achievement level on the ability scale for that grade and subject, and a_i, b_i, and c_i are the item parameters from the 3PL model for item *i* and *e* is the base of the natural logarithm (*e* = 2.71828…). The CSEM value for the grade and subject is calculated by averaging the empirical theta scale conditional standard error of measurement of all public school students with valid test scores from the current administration who scored exactly at the *Meets* scale score cut. This average value is rounded to four decimal places and multiplied by 2.0 to obtain the CSEM used for that grade and subject in the formulas above.

To find the expected *Meets* range for a given organization (school or district) *o* on benchmark *b*, the LB and UB need to be summed and averaged based on the number of times each item from benchmark *b* was administered to the students of organization *o*. The following formulas describe how the lower (ELB_{ob}) and upper (EUB_{ob}) bounds of the expected *Meets* range are calculated for benchmark *b* and organization *o*:

$$ELB_{ob} = \frac{\sum_{i \in b} N_{io} \times LB_i}{\sum_{i \in b} N_{io}}, \quad \text{(A.4)}$$

$$EUB_{ob} = \frac{\sum_{i \in b} N_{io} \times UB_i}{\sum_{i \in b} N_{io}}, \quad \text{(A.5)}$$

where N_{io} is the number of students administered item *i* in organization *o*, and the summations in the numerator and denominator are across all items measuring benchmark *b*.

A.5. Indicator Determination

Once the observed and expected performance measures for a benchmark are derived for an organization, the benchmark indicator is found by determining if the observed value is less than, within, or greater than the expected range. If OBS_{ob} is less than ELB_{ob}, then the indicator reported is **Less than** *Meets*. If OBS_{ob} is greater than EUB_{ob}, then the indicator reported is **Greater than** *Meets*. If OBS_{ob} is greater than or equal to ELB_{ob} and less than or equal to EUB_{ob}, the indicator reported is **Similar to** *Meets*.

A.6. Resources

View the subject-specific benchmark report "How To" Quick Guides for reading, mathematics, and science on the Additional Reporting Resources page (PearsonAccess Next > Reporting Resources > Additional Reporting Resources) for information about how you can use the reports in your district or school.

View the Benchmark Report Interpretive Guide (PearsonAccess Next > Reporting Resources > Additional Reporting Resources) for a comprehensive overview of the Reading, Mathematics, and Science MCA Benchmark Reports, along with information about understanding and using the data in your district or school.

View the Understanding the MCA Benchmark Report Video (PearsonAccess Next > Reporting Resources > Additional Reporting Resources) for an overview of benchmarks within the Minnesota Academic Standards and a walk-through of each section of the report.