DEPARTMENT OF EDUCATION

2024-25 Technical Manual for Minnesota's Statewide Assessments

Prepared by Pearson | May 2026

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
IEA................. Pearson’s Intelligent Essay Assessor
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
MRE .... Math Reasoning Engine
MTAS .... Minnesota Test of Academic Skills
MTELL .... Mathematics Test for English Language Learners
NCLB .... No Child Left Behind
OIS .... ordered item set
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, 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. Students are also 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 (MDE website > Districts, Schools and Educators > Teaching and Learning > Academic Standards [K-12]). To meet this requirement, Minnesota developed the Reading, Mathematics, and Science Minnesota Comprehensive Assessments (MCAs), the Reading and Mathematics Minnesota Test of Academic Skills (MTAS), and the Science Alternate Minnesota Comprehensive Assessment (Alt MCA). The academic content standards were approved for reading, mathematics, and science in 2010, 2007, and 2019, respectively. Under ESEA and Minnesota Statutes 2024, section 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.

MDE is transitioning to new academic content standards and updating the MCA series of assessments to align to the new content standards. This year is the first year of the new Science MCA-IV and Science Alt MCA assessments aligned to the revised 2019 Minnesota Academic Standards for science. The MCA-IV assessments will replace the MCA-III assessments in each subject, beginning with science in spring 2025, reading in spring 2026, and mathematics in spring 2028. The Alt MCA assessments are replacing the MTAS-III assessments according to the same schedule. In MDE's official documents and test data, the series number is added to indicate the academic content standards that are being measured (for example, MCA-III designates the science academic content standards adopted in 2009, while the MCA-IV designates the science academic content standards adopted in 2019).

The reading, mathematics, and science standards-based accountability tests are primarily given online. The accommodated versions are available for eligible students who are unable to take the main online test due to a disability. For paper tests to be scored and reported, the district must enter the student responses online during the testing window. Reading and Mathematics MCA-III are administered online adaptively, while Science MCA-IV is administered online as a fixed form. Minnesota also administers the Reading and Mathematics MTAS-III and Science Alt MCA each spring. The alternate assessments are available in every grade in which the MCA is available. The alternate assessment is administered in a one-on-one setting using paper materials for MTAS-III and in an online, paper, or hybrid mode for Alt MCA.

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. The first online operational administration of ACCESS was 2015-16. The WIDA Alternate ACCESS, first administered in the 2023-24 school year, is aligned to the WIDA English Language Development (ELD) Standards Framework, 2020 Edition and replaced the Alternate ACCESS.

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.

1.1.1. History Summary by Year

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]). 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. Refer to Minnesota Statutes 2024, 120B.30 for more information. 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 was 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 due to the removal of this requirement by law. 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 began development of 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*, was 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*, was updated to school year 2025-26. Additionally, the timelines for the first administration of the Alternate MCA were set to 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.

2024

In spring 2024, field testing continued for Science Alt MCA and Reading Alt MCA. The 2024 administration also 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 English language skills for accountability purposes and identifying continuation of EL services, as well as to inform continuing eligibility and exit decisions for students receiving English learner services. In addition, revisions to Minnesota Statutes 2024, 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. Additionally, this is the final year that subscore scale scores were reported for the MCAs.

2025

This year marked the first operational administration of the Science MCA-IV assessment aligned to the revised *2019 Minnesota Academic Standards in Science*, including the first operational administration of the CR items. This was also the first administration of the Science Alt MCA aligned to extended benchmarks derived from the *2019 Minnesota Academic Standards in Science*, administered in online, paper, or hybrid (any combination of online and of paper) mode as determined by the student's IEP team. Standard setting meetings took place for both the Science MCA-IV and Science Alt MCA in July 2025 to establish new performance standards for the assessments beginning with the spring 2025 administration. Reading MCA-IV and Alt MCA field testing continued in spring 2025 and was the first year of field testing for the Mathematics Alt MCA assessment. In addition, the alternate assessment eligibility requirements were revised to incorporate details about the transition to the Alt MCA and to clarify eligibility criteria. The decision-making tool was also revised to assist with assessment decisions for students. It is now titled *Eligibility Requirements for Minnesota Alternate Assessments*. Additionally, this is the final year that Lexile and Quantile measures were reported for the Reading and Mathematics MCA.

During this year, WIDA also held a voluntary Kindergarten ACCESS Field Test. Districts that applied and were selected received instructions from WIDA and DRC. The Kindergarten ACCESS Field Test did not serve as a substitute for the required ACCESS testing.

1.1.2. Timeline

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
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

Year	Event
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
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

Year	Event
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
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)

Year	Event
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 • 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 • Last year subscore scale scores were reported for Reading, Mathematics, and Science MCA
2024-25	• First operational administration of Science MCA-IV and Science Alt MCA • First year of Science Alt MCA administered in online, paper, or hybrid (any combination of online and of paper) mode • Standard setting for the Science MCA-IV and Science Alt MCA in July 2025 to establish cut scores for the new assessments • Mathematics Alt MCA-IV field test items administered for the first time in spring 2025 • Alternate assessment eligibility requirements were revised • Administration of WIDA's Kindergarten ACCESS Field Test • Last year Lexile and Quantile measures were reported for Reading and Mathematics MCA

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. 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 high-level aspects of the statewide assessment program, including approving testing calendars for future years and providing feedback and guidance on public messaging. 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).

Minnesota educators also served on standard setting and alignment study committees. During standard setting, experienced Minnesota educators collaboratively determine cut scores that define the expectations for student learning. Alignment studies are conducted to evaluate the alignment of the grade-level assessments and the Minnesota Academic Standards.

These educators helped to ensure the assessments reflected what students truly need to know and be able to do by applying their deep understanding of the academic content standards and their instructional experience with Minnesota students.

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 performance levels on the Minnesota Academic Standards.

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 . 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	Former Indiana Assessment Director
Dr. Gregory J. Cizek	Guy and Phillips Distinguished Professor of Educational Measurement and Evaluation, Emeritus	University of North Carolina at Chapel Hill
Dr. Claudia Flowers	Professor Emeritus of Educational Research, Measurement and Evaluation	University of North Carolina at Charlotte
Dr. Mark Reckase	Professor of Measurement and Quantitative Methods, Emeritus	Michigan State University
Dr. Michael Rodriguez	Dean, College of Education and 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)/performance level descriptors (PLDs), 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 Reading, Mathematics, 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

Assessment	Grades
Mathematics MCA-III	3-8, 11
Reading MCA-III	3-8, 10
Science MCA-IV	5, 8, high school[1]
Mathematics MTAS-III	3-8, 11
Reading MTAS-III	3-8, 10
Science Alt MCA	5, 8, high school[1]
ACCESS	K-12
WIDA Alternate ACCESS	K-12

1.3.1. Mathematics MCA-III

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

[1] The grade 5 Science MCA-IV and Alt MCA assess the grades 3-5 standards, and grade 8 assesses the grades 6-8 standards. Students in grades 9–12 are expected to take the high school assessment during the year in which they are enrolled in a life science or biology course and/or when they have received instruction on all 9-12 life science benchmarks.

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.

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 non-calculator sections of the assessment for grades 3-8 to ensure that students are not allowed to use a calculator in the non-calculator section. Students are administered four non-calculator items in the adaptive online assessments for grades 3-8 mathematics. Grade 11 mathematics does not contain non-calculator benchmarks, so a calculator is available for all items.

1.3.2. Reading MCA-III

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.

1.3.3. Science MCA-IV

The Science MCA-IV is aligned to the *2019 Minnesota Academic Standards in Science* and administered in grades 5 and 8 and once in high school. It is a phenomenon-based, fixed-form assessment with MC, TE, and short answer CR items. Minnesota revised its academic standards in science in 2019 and implemented them in September 2024. The standards are based on current science education principles found in A Framework for K-12 Science Education (NRC, 2012) that emphasize the inclusion of three dimensions within science standards, curriculum, and instruction: Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs). The standards are broken down further into benchmarks that incorporate aspects of the Framework's three dimensions. The items on the assessment are multi-dimensional in that each item assesses the SEP and DCI of the benchmark, and most also assess the CCC. In grades 5 and 8, students answer items about concepts and skills in the following strands:

- Practices in Physical Science
- Practices in Earth and Space Science
- Practices in Life Science

In high school, students answer items about concepts and skills in the following Life Science (LS) strands:

- Practices in LS1: From molecules to organisms: Structure and processes
- Practices in LS2: Ecosystems: Interactions, energy, and dynamics
- Practices in LS3: Heredity: Inheritance and variation of traits
- Practices in LS4: Biological Evolution: Unity and diversity

1.3.4. Mathematics MTAS-III

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.

1.3.5. Reading MTAS-III

The Reading MTAS-III is aligned to the *Minnesota K-12 Academic Standards in ELA* and administered in grades 3-8 and 10. The tasks in reading are designed to assess comprehension of the MTAS-III passages. 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.

1.3.6. Science Alt MCA

The Science Alt MCA is aligned to extended benchmarks derived from the *2019 Minnesota Academic Standards in Science* but reduced in depth, breadth, and complexity. Each extended benchmark has three Access Points (A, B, and C) that provide different levels of complexity for students to access the benchmark. The 2019 Minnesota Academic Standards in Science are based on current science education principles found in *A Framework for K-12 Science Education* (NRC, 2012) that emphasize the inclusion of three dimensions within science standards, curriculum, and instruction: SEPs, CCCs, and DCIs. The extended benchmarks incorporate at least two of the three dimensions (SEPs and DCIs). Test items are aligned to a specific SEP, DCI, and CCC as written in the extended benchmark's Access Points. Items on the assessment may be one- or two-dimensional.

The Science Alt MCA is a phenomenon-based, fixed-form assessment with MC items only. Within the grade-level assessment, each extended benchmark is assessed through a task that includes a stimulus (i.e., phenomenon) and three items based on the stimulus. Each grade-level assessment will include nine tasks with three items each for a total of 27 raw score points. The tasks are developed to multiple levels of cognitive complexity as described by the Access Points, and each item in the task is aligned to one of the three Access Points. The items also range from the most to least support. The operational test also includes embedded field test items that do not contribute to a student's score.

The alternate assessment is administered in a one-on-one setting in an online, paper, or hybrid of online and of paper mode. The IEP team determines which mode is best for the student. Once the student responds to a test item, the response of either A, B, or NR (No Response) for Access Point A items or A, B, C, or NR for Access Point B and C items is recorded on the data collection form or directly in the online test, and the student moves to the next item in the test. Items are not presented again once a response is indicated, and the student must answer every item within a task before moving to the next task. The recorded responses are scored as any other online test by comparing with the correct answer key. If the first six items in the test are marked as NR, the test administrator may decide to engage the early stopping rule (ESR) and stop testing.

1.3.7. 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

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.

[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.

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.

1.3.8. 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 Statutes 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.

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:

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 passages/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 designed 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 and subscores in reporting. Item types, cognitive levels of understanding to be tested, the range in the number of items, and other elements of test design or construction are established in the test specifications based on the need to assess the content 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 the test will be constructed. The test specifications also provide a basis for interpreting test results by providing subscore categories and descriptions of them.

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

MDE held meetings with Minnesota educators to define general test specifications for each grade for the MCA-III and MCA-IV. 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. Depending on the content area being assessed, 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, reporting categories (subscores) and other elements of test construction or design. 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 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) and Alternate MCA (Alt MCA)

A systematic and iterative process was used to create the MTAS-III and Alt MCA test specifications. For MTAS, 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. For Alt MCA, this identification of the benchmarks was part of the test specifications committee's initial steps. 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.
- 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 and Alt MCA, including multiple substrands, cognitive levels, and benchmarks. Teacher groups took the recommended benchmarks and 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.

For MTAS, 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.

For Alt MCA, the next step was to write access points for each extended benchmark. The access point breaks down the extended benchmark into specific skills the item is targeting to assess. Access points begin with "The student will (verb)…." and are similar to an objective written for an Individualized Education Program (IEP) statement for a student. Each extended benchmark has three access points that increase in cognitive complexity from Access Point A (which is aligned to the core idea of the extended benchmark at the knowledge level) to Access Point C (which is most fully aligned to the extended benchmark at the comprehension or application skill level).

MDE recognizes that the students who take the alternate assessment 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 or access points for three grade bands in mathematics and reading and each grade-level assessment in science.

For MTAS, 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.

For Alt MCA, following the writing of the access points, committee members described item specifications for writers, visual models, and appropriate contexts for each extended benchmark. Following a public review period, the committee convened to address suggested edits or clarifications in both extended benchmark language and access points. At the final meeting, they also looked at vertical alignment across grades and drafted samples for different grades intended to be used by educators.

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 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.

2.2. Item Development

This section describes the item writing process used during the development of test items (including stimuli such as passages and phenomena) and tasks, in the case of the MTAS-III and Alt MCA. 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 and MTAS/Alt MCA, items and tasks are written and internally reviewed by the testing service provider before submission to MDE. The Science MCA items and stimuli such as phenomena are written by Minnesota educators before review by the testing service provider and submission to MDE. MDE and Pearson collaborate on item bank analyses and the creation of item development plans to guide annual asset development. Progress against those plans is monitored throughout the development process and adjustments are made as needed in consultation with MDE.

[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. Content Limits and Item Specifications

For MCA-III and MTAS, 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.

For MCA-IV and Alt MCA, while test specifications designate how the test will be built each year, specifications and limits by benchmark are described for item writers in the Item Development Guide. These limits and specifications were written by the test specifications committee and used to keep consistent development of items to the benchmarks over time. Adjustments are made based on educator review committee and MDE feedback in the first several years of development of revised standards, but over time they become stable. The Item Development Guide is considered a secure material as part of the test development process.

2.2.1.1. Minnesota Comprehensive Assessments

Item specifications are provided for each assessed benchmark for the Mathematics and Reading MCA-III assessments in the test specifications and for the Science MCA-IV and Alternate MCA assessments in the Item Development Guide. These 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. Alternate Assessments

Criteria outlined by the National Alternate Assessment Center served as a guide in the development of the MTAS-III and Alt MCA to help ensure that items were based on the Minnesota Academic Standards. All the content of the alternate assessments is academic and derived directly from the standards in mathematics, reading, and science. The content limits of the MTAS-III and the access points of the Alt MCA 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 distinguishing between being able to match a definition to a word compared to determining the meaning of a given word. In science, this might be addressed by assessing both a science practice and core idea within an item (e.g., comparing models of the stages of different animals throughout their lifecycles to identify the stages in common).

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 content area expertise and 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. Alternate Assessments

Alternate assessment 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. Minnesota's testing service provider conducts item writer training for the Alt MCA that includes an overview of the requirements for alternate assessments based on alternate achievement standards, characteristics of students with significant cognitive disabilities, descriptions of tasks, principles of universal design, and test specifications. 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 MCA-III and MCA-IV. Training includes an overview of the test development cycle and specific training in the creation of high-quality, culturally affirming MC, TE, and CR items. Experienced service provider staff members lead the trainings and provide specific and evaluative feedback to participants.

2.2.3.2. Alternate Assessments

Minnesota's testing service provider conducts item writer training for the Alt MCA 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 tasks and reading passages. 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;
- 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 such as passages and phenomena) or tasks (for MTAS-III and Alt MCA) 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 and CR 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. Alternate Assessments

Before an Alt MCA task 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 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 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 and CR items:
    a. Are rubrics aligned to the items being asked?
    b. Is the type of item appropriate?

2.3.3.2. Alternate Assessments

Item review committees are composed of special education and content educators in a given content area along with multilingual learner educators at each meeting. 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 MCAs and are therefore familiar with the Minnesota Academic Standards. The collaboration between special education and content area educators ensures that the MTAS-III and Alt MCA assess 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. These committees evaluate 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. Approximately 15 committee meetings are convened involving Minnesota community members who represent students statewide.

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 Mathematics and Reading MCA-III adaptive assessments, 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 English learner (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.

[4] MH statistical indices require 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. 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.

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 7: 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 MCAs and Alt MCAs. Only the Science MTAS included the embedded field test items in spring 2022, so DIF analyses were conducted for those items. In 2023 and 2024, DIF analyses were conducted for MCA-IV and Alt MCA embedded field test items that appeared on the MCA Science and MTAS Science and Reading assessments, respectively. In 2025, DIF analyses were also conducted for the MCA-IV and Alt MCA embedded field test items that appeared on the MCA Reading and MTAS Reading assessments and for the Alt MCA embedded field test items that appeared on the MTAS Mathematics assessments.

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 forms (i.e., Science MCA, MTAS, and Alt MCA) 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 $2x2xj$ 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 (a_{MH}), as well as the lower ($a_{MH,lower}$) and upper ($a_{MH,upper}$) limits of the 95% confidence interval, are recorded.

To improve the ease of interpretation, the test statistic (a_{MH}), as well as the ($a_{MH,lower}$) and upper ($a_{MH,upper}$) 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):

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

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

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

(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)	$\|\boldsymbol{MH\ D - DIF}\|$ is not significantly different from zero, or is less than one.
B-DIF (slight to moderate)	$\|\boldsymbol{MH\ D - DIF}\|$ is significantly different than 0.0 (but not 1.0) and is at least 1.0. OR $\|\boldsymbol{MH\ D - DIF}\|$is significantly different than 1.0 but is less than 1.5.
C-DIF (moderate to large)	$\|\boldsymbol{MH\ D - DIF}\|$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, the Mathematics and Reading MTAS, and the Science Alt MCA 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 and MCA-IV

Data review meetings have been held annually with a few exceptions in certain years. The MCA-III and MCA-IV 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. Alternate Assessments

The Alt MCA 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 Alt MCA data review panels also consider disaggregated information about performance of students most likely to participate in the Alt MCA. 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 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 7: 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 grade 5, grade 8, and high school Science MCA-IV assessments; and all alternate 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 MCA-III for mathematics and reading and the MCA-IV for science to fulfill their requirement for each content area. Students with Individualized Education Programs (IEPs) who meet the alternate assessment eligibility criteria, as defined in the *Eligibility Requirements and Decision-Making Tool 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 for mathematics and reading and the Alt MCA for science 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. Mathematics MCA-III

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 calculators may be used when a calculator is allowed, while grade 11 students may use the online calculators 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.

3.2.2. Reading MCA-III

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.

3.2.3. Science MCA-IV

The grade 5, grade 8, and high school Science MCA-IV assessments are administered online, with paper test materials available for eligible students. Separate online test forms are also available—such as assistive technology and script forms—for eligible students. The online and paper assessments are each divided into two sections, allowing districts to administer the test over multiple days, if they choose. In the Science MCA, items are associated with a phenomenon, and each phenomenon 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 calculators 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 calculators.

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 the previous section of the test. Each district sets their testing schedule within the state-designated testing window.

3.2.4. Mathematics MTAS-III

Any district employee who has completed the applicable MTAS Test Administrator trainings 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.

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/Alt MCA Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to students.

3.2.5. Reading MTAS-III

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, Test Administrators 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. The following 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/Alt MCA Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to students.

3.2.6. Science Alt MCA

The Science Alt MCA is administered in a one-on-one setting in an online, paper, or hybrid (any combination of online and paper) mode to meet student needs. In the online mode, the Test Administrator guides the student's interactions and supports the student's use of the online test form, or the student interacts directly with the online assessment with supervision from the Test Administrator. In the paper mode, the student and Test Administrator use only paper materials for testing (such as the Task Administration Manual, Phenomena Book, and Presentation Pages). The Test Administrator or district staff then enters the student responses online in TestNav for scoring. In a hybrid administration, both the online test form and paper materials are used to support the student during test administration.

Once the student responds to a test item, the response of either A, B, or NR (No Response) for Access Point A items or A, B, C, or NR for Access Point B and C items is recorded on the Data Collection Form or directly in the online test, and the student moves to the next item in the test. Items are not presented again once a response is indicated, and the student must answer every item within a task before moving to the next task. The recorded responses are scored as any other online test by comparing with the correct answer key. If the first six items in the test are marked as NR, the Test Administrator may decide to engage the early stopping rule (ESR) and stop testing.

The Science Alt MCA includes object lists that provide guidance on the use of objects or manipulatives for students who need this type of support. The separate Phenomena Book includes materials used to present Alt MCA tasks to students. Although the Science Alt MCA 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/Alt MCA Task Administration Manual* to provide standardization of the content and to maintain the representation of the construct to students.

3.2.7. 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 to the applicable testing service provider, 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 Statewide Assessments*.

3.3.1. Minnesota Comprehensive Assessments

The Mathematics and Reading MCA-III and Science MCA-IV 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 for scoring, but districts are required to return secure test materials used in conjunction with online tests (such as scripts) and 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 and reading, and 12- and 18-point fonts for 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. Alternate Assessments

Secure test materials for the Mathematics and Reading MTAS-III and the Science Alt MCA include the Task Administration Manuals, Presentation Pages, Response Option Cards, and Passages, Scenarios, and/or Phenomena Books for applicable grades. Following administration, all used and unused Task Administration Manuals, Presentation Pages, visual descriptions, braille materials, and Passages, Scenarios, and/or Phenomena Books must be returned to Minnesota's testing service provider. All Response Option Cards must be securely disposed of by 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 for ELLs 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

The available universal 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 Minnesota Assessment Hub website (Minnesota Assessment Hub > Resources & Training > Policies and Procedures).

Universal supports and accommodations provided during statewide assessments should be familiar to a student through use in classroom instruction, classroom assessments, and district assessments. In general, a new support or accommodation should not be first introduced to a student at test administration time because it may negatively affect their performance on the test. The decision to provide a universal support or accommodation for the first time during testing should be made carefully by a team that is very familiar with the student's needs; however, practice with the support or accommodation is encouraged whenever possible. Student resources are available to familiarize students with the test they will be taking, including universal supports and accommodations.

The increasing number of universal supports available for all students, especially online tools and accessibility supports, provides more students access to a wider range of supports that they use in classroom instruction. However, an unintended consequence for students with an IEP or 504 plan is that often only accommodations are documented in the IEP or 504 plan. If a student has an IEP or 504 plan, it is recommended that it document all needed universal supports to ensure they are provided during testing.

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. Note: While universal supports are available to all students, some require a specific code to be applied in advance in order for students to access them. Universal supports may be provided by the school, provided by MDE, or embedded in the online test, and are divided into the following categories:

- Tools include classroom materials that students use in instructional, daily classroom, 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 multilingual learners) based on needs identified by an educator (or team of educators) familiar with the characteristics and needs of the student(s).

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. Districts are not required to document for MDE which universal supports or accommodations are provided on the MTAS/Alt MCA; however, the assistive technology online forms for Alt MCA require codes to be indicated. Accommodating student needs is integral to alternate assessments, and the Test Administrator may provide needed supports (for example, objects, tactile graphics, counters, or any type of calculator) or adapt materials (for example, enlarging materials or providing a braille version) as long as the support is not specifically prohibited in the task script.

3.5. 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 whether the use of a support or accommodation by students with disabilities and/or English learners (ELs) is effective and whether its use alters the focal construct of the assessment (Abedi and Ewers, 2013). Their research drew on a wide range of studies, meta-analyses, and expert consensus, including Jaman Abedi's research on accommodations, technical reports on accommodations for students with disabilities and English learners from the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), state policy syntheses, and best practices and guidelines. The research process included five main steps:

- Establishing criteria: Supports and accommodations were evaluated on effectiveness, validity, impact on diverse student backgrounds, relevance, and feasibility.
- Evidence gathering: Randomized controlled trials (RCTs) are preferred for assessing accommodations, but most studies use existing data without random assignment, which may introduce bias.
- Classification system: A notation system categorized supports/accommodations by evidence of effectiveness and validity, after expert review and consensus. Categories include "Use," "Use/Low Evidence," "Not Use," and "Unsure."
- Decision rubric: Recommendations depended on impact to the focal construct and associated risks. Supports not affecting validity were prioritized for use; higher risk was assigned where validity might be compromised.
- Ongoing review: The system is adaptable and incorporates new research findings as they emerge.

MDE reviewed its allowed supports (for English language proficiency and standards-based accountability assessments) and accommodations (focused on standards-based accountability assessments) against Abedi and Ewers' findings, and summaries are provided in the following tables. 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 Universal Supports—Accessibility Supports

Support/Accommodation	Research and Recommendations
Amplification Devices (Provided by School) Student uses an amplification device, including personal hearing aids, FM systems, or other amplification systems. This support may be provided to a student who normally uses these devices in instruction and daily classroom settings to amplify sounds.	No research has specifically considered the role of amplification as a support during standardized testing. However, according to Bertachini et al. (2015); Heeney (2007); Mealings (2022); Smart et al. (2018); and Wilson et al. (2021): • The use of sound amplification devices in the classroom is generally supported for students with disabilities. Research supports the effectiveness of the support and recommends its use in the classroom. The risk of the support giving students an unfair advantage is likely low.
Answer Masking (Embedded in MCA and Alt MCA Tests) The answer masking accessibility feature lets a student choose which answer choices will show on the screen. It is available only for multiple-choice and multiple-response questions.	No research has specifically considered the role of answer masking as a support. However, according to Hokken et al. (2023) and Patros et al. (2016): • Students with attention deficit hyperactivity disorder (ADHD) may have difficulty with visual selective attention or choice impulsiveness, leading

Support/Accommodation	Research and Recommendations
This feature may be most beneficial for a student with attention difficulties, print disabilities, and low vision who use a similar strategy or feature during instruction to visually reduce content.	to difficulty selecting the correct response from a list of options. According to Bertoni et al. (2019); Huurneman et al. (2014); Ogata et al. (2019); Roach & Hogben (2007); and Stein (2014): • Students with dyslexia or low vision may experience a visual crowding effect, leading to difficulties processing multiple responses. Though none of this research specifically discusses answer masking, it may be inferred that students with disabilities who struggle with visual crowding or selective attention would benefit from answer masking. The risk of the support giving students an unfair advantage is likely low.

Support/Accommodation	Research and Recommendations
Calculator (Handheld), Mathematics Manipulatives, or Abacus for Science MCA, Grades 3–8 and 11 Mathematics MCA, Mathematics MTAS, and Science Alt MCA (Provided by School) Student uses a handheld calculator for testing, or a student uses one or more of the following materials in place of or alongside a calculator: • Mathematics manipulatives such as unifix cubes or base-10 blocks • An abacus calculation tool Standard, scientific, or graphing calculators with a talking feature may also be used in an individual setting if the need is documented in the IEP or 504 plan and the calculator policies in Chapter 8 and Appendix B of the *Procedures Manual* are followed. A handheld calculator may be provided to a student who normally uses a handheld calculator in instruction. A student who may benefit from using mathematics manipulatives is typically identified as a student with visual disabilities or math-related disabilities, such as dyscalculia, and is a student with Learning Disabilities, Autism, Visual Impairments, or Intellectual Disabilities. MDE does not produce a comprehensive list of allowable manipulatives. For an abacus, a student with low vision or with documented processing impairments may benefit from the use of an abacus in place of a calculator. While specific Texas Instruments and Desmos (beginning in 2024-25) calculators are the types of calculators available in TestNav for online testing, any brand of calculator may be used when handheld calculators are allowed.	Limited research describes the use of mathematics manipulatives during standardized testing. However, according to Bone (2020); Bone et al. (2021); Bouck & Park (2018); Bouck et al. (2018); and Elliott et al. (2009): • Experts support the effectiveness of mathematics manipulatives for students with disabilities. Research indicated that manipulatives help students grasp math concepts by making abstract ideas tangible, reducing cognitive load, and improving accuracy and independence. They also support externalizing thinking and encourage persistence with complex tasks, especially when combined with other accommodations/supports. • Despite uncertainties, research supports the use of the accommodation. The risk of the accommodation giving students an unfair advantage is likely low. According to Bouck and Bouck (2008); Fuchs et al. (2000); Lutsic & Zhang (2023); Jiang et al. (2023); Jiang & Cayton-Hodges (2023); Russell (2006); Russell (2014); and Shaftel et al. (2006): • Calculators, which often are automatically included for mathematics tests, are widely used by all students. Most research indicates that this support provides a significant benefit to students. • Research supports the effectiveness of the support and recommends its use. There is no risk the support gives students an unfair advantage if provided in accordance with test administration policies.

Support/Accommodation	Research and Recommendations
Co:Writer Universal/Read&Write for Science MCA (Embedded in Test) These TestNav extensions allow students to use speech-to-text and/or word prediction to complete constructed-response items for science. This support may benefit a student who uses speech-to-text writing supports in their classroom instruction. Students with processing-related needs, print disabilities, and writing disabilities may find this support helpful. Multilingual learners may also benefit from this support. Ensure the student is familiar with this support prior to testing.	No research has specifically considered the role of speech-to-text as a support during standardized testing. However, according to Almgren Bäck et al. (2024); Alrawashdeh et al. (2024); Bouck et al. (2011); Cullen et al. (2008); Dawson et al. (2019); Evmenova et al. (2010); Garrett et al. (2011); Lee (2011); Liu (2019); Levine et al. (2023); MacArthur & Cavalier (2004); Matre & Cameron (2024); McCollum et al. (2014); Noakes et al. (2019); and Svensson (2021): • Research generally supports the use of speech-to-text and word prediction among students with disabilities to increase speed and accuracy, though some research has noted increased error rates. Students without disabilities demonstrate, on average, either limited or inconsistent benefits. Regarding multilingual and English learners, research by Lewis (2024) and Arcon et al. (2017): • Showed positive effects of speech-to-text compared to handwriting. The risk of the support giving students an unfair advantage is likely low.
Color Contrast (Embedded in MCA and Alt MCA Tests) The color contrast tool lets a student change the color of the screen and words in TestNav. The available color choices are: • Black text on White (Default) • Black text on Cream • Black text on Light Blue • Black text on Rose • White text on Black • Yellow text on Blue • Gray text on Green This feature may be most beneficial for a student who uses various contrasts between the text and background color on other devices during instruction for any reason including attention difficulties, print disabilities, and low vision.	No research has specifically considered the role of color contrast as a support during standardized testing. However, according to Janáky et al. (2014); Ilhan et al. (2019; Richardson et al. (2014), and Rogers et al. (2022): • Research supports the use of color contrast to benefit students with disabilities, attention difficulties, low vision, and color-blindness. The risk of the support giving students an unfair advantage is likely low.

Support/Accommodation	Research and Recommendations
Color Overlays (Provided by School) A transparent colored sheet that is placed over a page of text or is affixed to a device screen. This support may provide assistance to a student who normally uses colored overlays to reduce visual stress when reading.	No research has specifically considered the role of color overlays as a support during standardized testing. However, according to Denton & Meindl (2016); Griffiths et al. (2016); Henderson et al. (2013); Jakovljević et al. (2021); Razuk et al. (2018); and Sjoblom et al. (2016): • Findings regarding the effectiveness of color overlays are mixed, with some studies finding benefits, while others find no effect or detrimental effects. The risk of the support giving students an unfair advantage is likely low.
English Glossary for Science MCA (Embedded in Test) The English glossary feature provides concise definitions or explanations of terms in English. The words selected for glossing represent academic vocabulary frequently found in grade-appropriate science texts; however, the lists are not exhaustive and students may still encounter unfamiliar terms in the test. Words that have been glossed are underlined in the online test. When selected, a pop-up window with the definition appears with an audio option. Beginning with science in spring 2025, the English glossary is a replacement for pop-up translations. This feature is intended to increase accessibility for a wider population of multilingual learners and, potentially, support all learners. English glossaries may provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or dual-language programs. Additionally, English glossaries may be beneficial to any student who typically receives classroom support with academic vocabulary development.	English glossaries have been studied as a support for English learners. According to Abedi/CRESST Policy Brief (2001); Abedi et al. (2020); Cohen et al. (2017); Kieffer et al. (2009)/(2012); Li & Suen (2012); Liu et al. (2020); Myers & Kopriva (2015); Pennock-Roman & Rivera (2011); Roohr & Sireci (2017); Wolf et al. (2009); Wolf et al. (2012); Yang (2020): • Findings regarding the effectiveness of English glossaries as a test support are mixed, with some studies finding marginal benefits, while others find none. • Some researchers have suggested the need for additional training on using a glossary effectively, as many students avoid using them during the test. The risk of the support giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Line Reader Mask (Embedded in MCA and Alt MCA Tests) The line reader mask accessibility feature blocks out part of the screen using a gray box with a small window. The size of both the gray box and window can be adjusted. This feature helps a student focus on one part of a question, reading passage, or science phenomenon at a time by blocking out part of the screen. A student who uses a similar strategy or feature during instruction for any reason including visual processing needs may benefit from this feature.	No research has specifically considered the role of line reader masks as a support. However, according to Hokken et al. (2023) and Patros et al. (2016): • Students with ADHD may have difficulty with visual selective attention or choice impulsiveness, leading to difficulty selecting the correct response from a list of options. According to Bertoni et al. (2019); Huurneman et al. (2014); Ogata et al. (2019); Roach & Hogben (2007); and Stein (2014): • Students with dyslexia or low vision may experience a visual crowding effect, leading to difficulties processing multiple responses. It may be inferred that students with disabilities who struggle with visual crowding or selective attention would benefit from line reader masks. The risk of the support giving students an unfair advantage is likely low.
Low-Vision Aids (Provided by School) A device or tool designed to improve the visual performance of a person with low vision. Examples of low-vision aids are magnifying glasses, electronic magnifiers, cardboard cutouts, and a larger monitor screen size and different resolution for online assessments. A student who is accustomed to viewing enlarged text or graphics or navigation buttons may need low-vision aids to view content on paper or online assessments. This support also may meet the needs of a student with low vision and other print disabilities. The use of this support should be considered along with other supports, such as accessibility features embedded in the online test (for example, magnifiers) or an accommodation (such as large print paper test books).	Limited research has specifically considered the role of low-vision aids during standardized testing. However, according to Granquist et al. (2018); Latham (2018); Lusk (2012); Mulloy et al. (2014); and Yu et al. (2020): • The use of low-vision aids is generally beneficial in improving reading ability. The risk of the support giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Magnifier (Embedded in MCA Test) The magnifier accessibility feature lets a student make part of the screen larger by dragging a magnified box over the area they want to make larger. This feature may be most beneficial for a student who uses a similar feature during instruction to magnify a specific part of the screen.	See above discussion of research on Low-Vision Aids.
Online Version of Mathematics Tables (Embedded in MCA Test) Students can access online versions of multiplication or hundreds tables in place of or alongside a calculator on all grades of the Mathematics MCA. For grades 3-8, tables are only allowed in sections where a calculator is also allowed. A student with visual processing or spatial perception needs or a documented and persistent calculation disability (for example, dyscalculia) may find this support beneficial. Ensure the student is familiar with these formats and/or uses them during instruction.	No research has specifically considered the role of mathematics tables as a support. However, according to Spurlock (2020): • Using mathematical tables during class was associated with improved performance on later standardized testing. Using mathematics tables during testing may serve to reduce cognitive load. According to Gilmor et al. (2015): • Reducing cognitive load during a mathematics test may also improve performance. The risk of the support giving students an unfair advantage is likely low

Support/Accommodation	Research and Recommendations
Paper Version of Mathematics Tables for Grades 3-8 Mathematics MCA (Provided by MDE) Students use multiplication or hundreds tables provided by MDE in place of or alongside a calculator while taking the paper Mathematics MCA. Only paper versions of the multiplication and hundreds tables posted on the [Minnesota Assessment Hub]() website may be provided during testing (Minnesota Assessment Hub > Resources & Training > Student Tools and Supports). Note: The paper version of the mathematics tables is only available for students in grades 3-8 who are already testing using a paper test book, which is considered an accommodation for students with an IEP or 504 plan. A student with visual processing or spatial perception needs or a documented and persistent calculation disability (for example, dyscalculia) may find this accommodation beneficial. Ensure the student is familiar with these formats and/or uses them during instruction.	See above discussion of research on Online Version of Mathematics Tables.

Support/Accommodation	Research and Recommendations
Paper Version of Mathematics Tables for Grade 11 Mathematics MCA (Provided by MDE) Students use a paper version of the multiplication or hundreds table provided by MDE in place of or alongside a calculator while taking the grade 11 Mathematics MCA in a paper test book or online. Only paper versions of the multiplication and hundreds tables posted on the Minnesota Assessment Hub website may be provided during testing (Minnesota Assessment Hub > Resources & Training > Student Tools and Supports). A student with visual processing or spatial perception needs or a documented and persistent calculation disability (for example, dyscalculia) may find this support beneficial. Ensure the student is familiar with these formats and/or uses them during instruction.	See above discussion of research on Online Version of Mathematics Tables.
Pop-Up Translations in Hmong, Spanish, and Somali for Mathematics MCA (Embedded in Test) The pop-up translation accessibility feature provides translations for academic terms and direction words. Words that have been translated are underlined in the online test. When selected, a pop-up window with the translation appears. No audio is provided for the translations. These translations may provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or dual-language programs.	According to Abedi, et al. (2020); Cohen et al. (2020); Kopriva et al. (2007); Li & Suen (2012); Pennock-Roman & Rivera (2011); and Yang (2020): • Findings regarding the effectiveness of bilingual glossaries (including pop-up bilingual translations) as a test support are mixed, with some studies finding marginal benefits, while others find none. The risk of the support giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Standard Text-to-Speech for Mathematics MCA and Science MCA/Alt MCA (Embedded in Test) The standard text-to-speech accessibility feature provides computer-generated audio, and a student can select the parts of the item they want to listen to. For science, the standard text-to-speech has been updated to remove directional text and provides a comprehensive read-aloud experience for students. Standard text-to-speech may be beneficial for a student with reading-related, vision, and visual processing needs. A student who uses this feature during daily instruction may benefit from this support. Standard text-to-speech may also be beneficial for multilingual learners whose listening proficiency may be higher than their reading proficiency.	According to Acosta et al. (2008); Alrawashdeh et al., (2024); Almgren Bäck et al. (2024); Barton (2001); 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); Svensson et al. (2019); Wei (2024);Wolf et al. (2009); and Wood et al. (2018): • 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 students with disabilities benefit from this support. Text-to-speech can help students reduce cognitive load, engage with challenging or age-relevant texts that they might not be able to read on their own because of difficulties with decoding, enhance comprehension by combining written text with audio, enabling students to better absorb and remember information, and maintain concentration and stamina when reading lengthy or challenging texts. The risk of the support giving students an unfair advantage is low.
Styluses for Devices (Provided by School) A stylus may be used by the student to navigate through the online assessment, use tools, and indicate responses. A stylus may provide assistance to a student with fine motor needs or print disabilities who normally uses it in daily instruction.	No research considered the use of styluses as a support.
Templates to Reduce Visual Print Field (Provided by School) A material (such as a cardboard cutout) that blocks out part of the screen or test book. This support may help a student focus on one part of a question or test content at a time by blocking	No research has specifically considered the role of templates to reduce visual print field as a support. However, according to Hokken et al. (2023) and Patros et al. (2016): • Students with ADHD may have difficulty with visual selective attention or choice impulsiveness, leading

Support/Accommodation	Research and Recommendations
out part of the screen or test book. These materials may provide assistance to a student who normally uses them in daily instruction.	to difficulty selecting the correct response from a list of options. According to Bertoni et al. (2019); Huurneman et al. (2014); Ogata et al. (2019); Roach & Hogben (2007); and Stein (2014): • Students with dyslexia or low vision may experience a visual crowding effect, leading to difficulties processing multiple responses. It may be inferred that students with disabilities who struggle with visual crowding or selective attention would benefit from this support. The risk of the support giving students an unfair advantage is likely low.
Translated Test Administrator Script for ACCESS and WIDA Alternate ACCESS (Written, Oral) (Provided by School) The test directions in the *Test Administrator Scripts* for ACCESS may be translated into a language other than English in an oral or written format. Only the test directions may be translated; no translations of test items or response options are allowed. Translations must be made by an interpreter (in-person or recording), not via an app or online platform (such as Google Translate). WIDA provides translations of the test directions in 16 languages. Districts may translate the scripted directions into additional languages as needed. Note: Translated directions are not allowed for Kindergarten ACCESS and WIDA Alternate ACCESS as the directions are embedded within the test, and Test Administrators may never use a language other than English while administering the test. This support may benefit students who typically need directions provided in another language.	According to Abedi (2020); Acosta et al. (2008); Barton (2001); Bolt and Thurlow (2004); Brown (2007); Burch (2002); Buzick & Stone (2014); Calhoon et al. (2000); Castellon-Wellington (2000); Christensen et al. (2011); Cormier et al. (2010); DeBacker et al. (2024); Dolan et al. (2005); Elbaum (2007); Fuchs et al. (2000); Helwig et al. (2002); Johnson, Kimball, Brown, and Anderson (2001); Kopriva et al. (2007); Li (2014); Pennock-Roman and Rivera (2011); Pennock-Roman and Rivera (2012); Rios et al. (2020); Sato et al. (2007); Tindal et al. (1998); and Wolf et al. (2009): • Collective research provides varied conclusions as to the effectiveness of this support. Although results vary across grades, subjects, disability type, and level of proficiency in a subject or skill, the overall consensus confirms students with disabilities benefit from this support, though benefits may be smaller for math and science tests than reading tests. Findings suggest that the effectiveness of translated directions is heavily moderated by the student’s first language literacy and reading support in non-reading subjects yields limited gains; for instance, it may hinder rather than help math problem solving and symbol or equation processing. • The risk of the support giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Voice Feedback Devices (Provided by School) Voice feedback devices and whisper phones allow a student to vocalize as they read and work problems. This support also includes a student reading the test out loud in an individual setting. This support may benefit a student who normally vocalizes what they read in daily instruction.	No research has specifically considered the role of voice feedback devices as a support during standardized testing. However, according to Sanjurjo (2009): • Some preliminary evidence with elementary school students suggests that whisper phones may be effective at improving reading fluency. According to Prior et al. (2011); Roberts et al. (2024); Robinson et al. (2019); Schimmel & Ness (2017); and Smyrnakis et al. (2021): • Findings are mixed regarding the benefits of reading aloud, as some studies find that oral reading benefits comprehension and memory, while other studies have no effect. Reading aloud may have particular benefits for students with learning disabilities or students who are just learning to read. The risk of the support giving students an unfair advantage is likely low.
Word-to-Word Translation and Glossary Using Word Lists for Science MCA (Provided by MDE) Translated and glossed word lists may allow students to access the meaning of unfamiliar words in test items. This support may help students better comprehend the test questions and answer choices by reducing confusion or misunderstanding resulting from ambiguous or low frequency academic vocabulary. Glossed word lists provide concise definitions or explanations of terms in English. Academic terms and direction words are translated in the following languages: Amharic, Arabic, Chinese, Hmong, Karen, Oromo, Russian, Somali, Spanish, and Vietnamese. The lists must be printed versions of the current-year lists posted on the Minnesota Assessment Hub website (Minnesota Assessment Hub > Resources & Training > Student Tools and Supports).	According to Abedi, et al. (2020); Cohen et al. (2020); Kopriva et al. (2007); Li & Suen (2012); Pennock-Roman & Rivera (2011); and Yang (2020): • Students access translated or glossed word lists differently based on test format. Online, they use the testing platform and may hear audio if available. On paper tests, students receive English or bilingual glossaries as separate reference documents. • Findings regarding the effectiveness of translated words as a test support are mixed, with some studies finding marginal benefits, while others find none. The studies found that word lists are most effective when well-designed, paired with other supports, and familiar to students. A common observation was that students frequently lacked understanding about the proper use of word lists. Therefore, their benefits can be limited if students don’t know how to use them. Research indicates that word lists offer modest gains when used alongside additional supports.

Support/Accommodation	Research and Recommendations
Translated and glossed word lists may provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or dual-language programs. Translated word lists provide direct one-to-one translations of terms in a student's target language(s); as a result, they will be most beneficial for multilingual learners who are literate in the target language(s) and have encountered these translated terms during instruction. Glossed word lists may be beneficial to any student who typically receives classroom support with academic vocabulary development.	The risk of this support giving students an unfair advantage is low.
Word-to-Word Translations Using Dictionaries or Programs for Mathematics MTAS and Science Alt MCA (Provided by School) Mathematics and science terms in the task (bold text in the Task Administration Manual and/or information in the Presentation Pages, Scenarios Book, Phenomena Book, or online test) or response options may be entered into a translation program or looked up in a dictionary. A commercial word-to-word dictionary contains mathematical and scientific terms in English and in the primary language of a given learner. No definitions or pictures may be provided—only direct translations of the terms—and a commercial dictionary must be used. When looking at online translation applications or programs to provide this support, ensure that no definitions or pictures are provided. Because applications and programs can change frequently, MDE does not provide a list; however, note that Google Translate is **not** allowed because it can provide definitions. These materials or tools may provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or	See above discussion of research on Word-to-Word Translation and Glossary Using Word Lists for Science.

Support/Accommodation	Research and Recommendations
dual-language programs, who use them during instruction.	
Word-to-Word Translation Using Word Lists for Mathematics MCA (Provided by MDE) Translated word lists may allow students to access the meaning of unfamiliar words in the test items. Academic terms and direction words are translated in the following languages: Amharic, Arabic, Chinese, Hmong, Karen, Khmer, Laotian, Oromo, Russian, Somali, Spanish, and Vietnamese. The lists must be printed versions of the current-year lists posted on the [Minnesota Assessment Hub](#) website (Minnesota Assessment Hub > Resources & Training > Student Tools and Supports). These lists are intended only for the purpose of providing this support during testing and are not intended as the basis for instruction. Translated word lists may provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or dual-language programs. Translated word lists provide direct one-to-one translations of terms in a student's target language(s); as a result, they will be most beneficial for multilingual learners who are literate in the target language(s) and have encountered these translated terms during instruction.	See above discussion of research on Word-to-Word Translation and Glossary Using Word Lists for Science.

Support/Accommodation	Research and Recommendations
Zoom Feature (Embedded in MCA and Alt MCA Tests) A student can use the Zoom accessibility feature to make the entire screen larger by using keyboard commands or pinching on touchscreen devices. TestNav supports zoom levels of up to 500%. This feature may be most beneficial for a student with vision impairments who uses a similar strategy or feature during instruction. The user interface is designed to respond to zoom by making the assessment content larger, and also collapsing menu systems (much like when the mobile version of a website is accessed). This means that a student who is instructionally using zoom within a browser may have a similar experience within TestNav.	See above discussion of research on Low-Vision Aids.

Table 3.2. Research Base for Universal Supports—Administrative Considerations

Support/Accommodation	Research and Recommendations
Different Time of Day (Provided by School) The time of day an assessment is administered is changed. A student who needs to test at a different time of day than originally scheduled and/or than is planned for other students may do so for any reason. This may include adjusting test administration if a student does not have necessary aids (such as glasses, hearing aids) when they are to begin testing and/or has another extenuating circumstance occur (such as medication not taken, traumatic event).	According to Hahn et al. (2012) and Randler et al. (2016): • Although students may perform better on some cognitive tasks when allowed to do so during their preferred time of day, no research has indicated that time of day has any effects on test performance. Given that this support may also be provided as a result of extenuating circumstances (such as traumatic events), research on the effects of stress on test-taking was also considered. According to Heissel et al. (2021) and Robson et al. (2023): • Increased levels of stress can negatively impact test performance. Thus, allowing students to take a test at a different time of day due to a traumatic event may improve their test experience. The risk of this support giving students an unfair advantage is low.

<table>
<tr><th>Support/Accommodation</th><th>Research and Recommendations</th></tr>
<tr><td>Environment (Provided by School)
The assessment is administered in a special setting (for example, certain lights, acoustics, seating arrangements).

A student who needs a sensory friendly environment to calm or alert their sensory system or to decrease distractions may benefit from this support. A student who requires a special setting or feature may have it provided for any reason as long as test security requirements are followed.</td><td>No research has specifically considered the role of sensory environment changes as a support during standardized testing. However, according to Mallory & Keehn (2021) and Unwin et al. (2021):
A sensory friendly environment may result in increased attention among students with autism spectrum disorder (ASD).
The risk of the support giving students an unfair advantage is likely low when used in accordance with test administration policies.</td></tr>
<tr><td>Extended/Flexible Testing Time (Provided by School)
Statewide assessments are not timed. Flexible or extended testing time means that a student requires additional time to complete an assessment that should be considered when scheduling testing.

For MCA and MTAS/Alt MCA, taking a single test over multiple days or sessions is allowable as a general support for all students. Testing may be split over multiple days with one or more sections completed on a given day.

For ACCESS and WIDA Alternate ACCESS, only extended time within a day for a single domain is considered a universal support—extended time for a single domain test across multiple days is considered an accommodation.

This support may be most helpful for a student who needs extra time to process written text, to write, or to use other accommodations (including braille tests, American Sign Language (ASL) interpretation, scribe, or script) or equipment. Extended time and/or more frequent breaks should be planned for if this is a support or practice typically provided during instruction or other assessments.</td><td>According to Bernard & Witmer (2025); Brown et al. (2011); Buehler (2001); Cahalan Laitusis et al. (2007); Cohen et al. (2005); Crawford and Tindal (2004); DiCerbo et al. (2001); Fletcher (2009); Holmes & Silvestri (2019); Lewandowski et al. (2007); Lindstrom & Gregg (2007); Thurlow and Bolt (2001); and Walz et al. (2000):
Research is divided on whether extending testing time is effective. Some studies revealed that students with disabilities in lower grades and students with ADHD or low-level reading abilities benefited. In other studies, students with disabilities benefited little or did not benefit at all and general education students benefited. Some studies noted that students did not always use extended time when it was made available to them. Experts recommend the support, which is used in most states, be used thoughtfully and carefully, only when absolutely needed. The benefits can include a decrease in cognitive demands, the chance for individuals to show what they know without being disadvantaged by slower response times, and greater engagement and perseverance during testing.
The risk of the support giving students an unfair advantage is low.</td></tr>
</table>

Support/Accommodation	Research and Recommendations
Individual or Small-Group Setting (Provided by School) A student is tested individually or in a small-group setting. MDE does not define "small group" as a specific number of students; this is determined by the district. A student may be tested in a different setting than planned for the majority of test administrations as needed to meet specific needs, such as to reduce distractions. An individual or small-group setting can be used for any student that would benefit from this smaller testing environment. A small group setting may be required for the administration of some supports and accommodations.	According to Lewandowski et al. (2015); Lewandowski et al. (2021); Lovett et al. (2019); and Weis & Beauchemin (2020): • Students with disabilities rate private testing rooms as less distracting and attempted more items, but no performance gains occurred. The risk of this support giving students an unfair advantage is low.
Written, Repeated, or Clarified Testing Directions (Provided by School) For MCA, the Test Monitor may write out scripted directions from the applicable [Testing Directions](#) (from the *Script to Read to Students*, *Answering Questions and Providing Assistance*, and *Transcript of Student Directions* (for online testing) sections) or the *Test Administrator Scripts* for ACCESS. After the scripted directions are read, Test Monitors may also repeat, explain, or clarify the scripted directions that may be confusing to a specific student or group of students. This support has been updated to include guidance about explaining and clarifying scripted directions; simplified/clarified directions is no longer a separate support. This support may benefit a student who typically needs directions provided verbally or in written form more than one time to engage in assessments. It may also provide support to multilingual learners, including English learners (ELs), former ELs, or students in immersion or dual-language programs. Scripted directions may be repeated or clarified as needed, following the	According to Kuti (2011) and One Feather (2010): • Unclear directions may limit the effectiveness of the test, but clarified directions does not improve students' test performance. The risk of this support giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
guidelines below. For written directions, Test Monitors should consider what is needed by a student; providing too much written information (which may also be repeated verbally, as needed) may be distracting or overwhelming.	

Table 3.3. Research Base for Universal Supports—Tools

Support/Accommodation	Research and Recommendations
Answer Eliminator (Embedded in MCA Tests) Students can use the answer eliminator tool to cross out an answer choice for multiple-choice and multiple-response questions.	No research has specifically considered the role of answer eliminator as a support. However, according to Hokken et al. (2023) and Patros et al. (2016): • Students with ADHD may have difficulty with visual selective attention or choice impulsiveness, leading to difficulty selecting the correct response from a list of options. According to Bertoni et al. (2019); Huurneman et al. (2014); Ogata et al. (2019); Roach & Hogben (2007); and Stein (2014): • Students with dyslexia or low vision may experience a visual crowding effect, leading to difficulties processing multiple responses. It may be inferred that students with disabilities who struggle with visual crowding or selective attention would benefit from using the answer eliminator. The risk of the support giving students an unfair advantage is likely low.

<table>
<tr><th>Support/Accommodation</th><th>Research and Recommendations</th></tr>
<tr><td>Calculators (Embedded in Test)
Students can use the calculator tools to help them answer questions. The calculators are only available on mathematics and science questions where students are allowed to use calculators. The following calculators are available:
Online grades 3-8 Mathematics MCA:
Grades 3-5: Basic (TI-108™ and Desmos Four-Function)
Grades 6-7: Scientific (TI-30XS™ and Desmos Scientific)
Grade 8: Graphing (TI-84 Plus CE™ and Desmos Graphing)
Online grade 11 Mathematics MCA: Graphing (TI-84 Plus CE™ and Desmos Graphing)
Online Science MCA: Basic (TI-108™ and Desmos Four-Function)</td><td>According to Bouck and Bouck (2008); Fuchs et al. (2000); Lutsic & Zhang (2023); Jiang et al. (2023); Jiang & Cayton-Hodges (2023); Russell (2006); Russell (2014); and Shaftel et al. (2006):
Calculators, which often are automatically included for math tests, are widely used by all students. Most research indicates that this support provides a significant benefit to students.
There is no risk the support gives students an unfair advantage when used in accordance with test administration policies.</td></tr>
<tr><td>Classroom Materials (Provided by School)
Classroom materials include materials provided by the school or the student that are typically used in instruction, such as highlighters, specific writing utensils, pencil grips, or place markers. These materials may be used as needed with paper test materials or related paper test materials for online assessments (such as scratch paper). Test administration resources (such as object lists for Alt MCA and testing directions for MCA) specify materials allowed and prohibited by test.

These materials may provide assistance to a student who normally uses these types of supports in daily instruction.</td><td>No research has specifically considered the role of classroom materials as a support. However, according to Ashbaugh (2018); Leonard et al. (2021); Yue et al. (2015); and Zante et al. (2024):
Using a highlighter may improve reading comprehension and later memory of material, though some studies have found no effect on reading performance.
The risk of the support giving students an unfair advantage is likely low.</td></tr>
</table>

Support/Accommodation	Research and Recommendations
Formula Sheets (Embedded in Test) Students can access online, grade-level formula sheets in grades 5-8 and 11 Mathematics MCA.	No research has specifically considered the role of formula sheets as a support during standardized testing. However, according Chang & Shieh (2013); Laderach (2017); and Paquin et al. (2020): • Among college students, self-created formula sheets may improve test performance, and teacher provided sheets are perceived positively. Since research on this support is limited, further rationale may be derived from cognitive load theory. According to Gilmor et al. (2015): • Reducing cognitive load improves performance on math tests. The risk of the support giving students an unfair advantage is likely low.
Highlighter (Embedded in Test) Students can use the highlighter tool to highlight words. They can choose the color they wish to use and can also listen to highlighted text.	No research has specifically considered the role of highlighters as a support. However, according to Ashbaugh (2018); Leonard et al. (2021); Yue et al. (2015); and Zante et al. (2024): • Using a highlighter may improve reading comprehension and later memory of material, though some studies have found no effect on reading performance. The risk of the support giving students an unfair advantage is likely low.
Music or Noise Buffers (Provided by School) Music or noise buffers may include individual study carrels, headsets, earplugs, and individual portable buffers set on the student's desk, or an audio player that generates white noise or instrumental music. The noise buffer or instrumental music can be accessed through headphones or in an individual setting. Music or noise buffers may help reduce noise distractions for the student taking the test. Music and noise buffers provided by the school may be used, but only by students who have used them in instructional or other assessment settings.	No research has specifically considered the role of music or noise buffers as a standardized testing support. However, according to Dalton & Behm (2008); de la Mora Velasco & Hirumi (2020); Hallam et al. (2002); Lehmann & Seufert (2017); and Thompson et al. (2012): • Findings are mixed regarding the effects of background music on academic activities. Some studies indicate that fast, aggressive, or loud music may have more negative effects compared to other types of music. According to Kulawiak (2021) and Smith & Riccomini (2013): • Effects of noise-reducing headphones are limited, but they may have benefits for some students with disabilities. The risk of the support giving students an unfair advantage is likely low.

Support/Accommodation	Research and Recommendations
Ruler (Embedded in Test) Students can use the ruler tool on specific Science and grade 3 Mathematics MCA questions to measure length.	No research has specifically considered the use of rulers as a standardized testing support. However, according to Gómezescobar et al. (2020); Gómezescobar (2025); and Kwon et al. (2024): • Students in elementary school often struggle to use a ruler effectively. Thus, ensuring students are familiar with available tools is important for its use. The risk of the support giving students an unfair advantage is likely low.
Scratchpad (Embedded in Test) Students can use the scratchpad tool to write or draw notes during the test.	No research has specifically considered the use of a scratchpad or scratch paper as a standardized testing support. However, according to Pengelley et al. (2023) and Prisacari & Danielson (2017): • Using scratch paper is more common on paper tests than computerized tests, possibly due to the cognitive load of switching between screens and paper. Thus, a scratchpad may provide a benefit beyond scratch paper due to lessened cognitive load. The risk of the support giving students an unfair advantage is likely low.
Scratch Paper (Provided by School) Schools may provide or allow students to provide scratch paper for paper and online assessments. Specific materials may be used as scratch paper, including blank paper, individual whiteboards, commercially produced graph paper, or student testing tickets. This support is widely supportive for all students, especially for the ACCESS Writing test.	See above discussion of research on Scratchpad.

Support/Accommodation	Research and Recommendations
Sensory Tools (Provided by School) Sensory tools include materials such as fidgets or weighted pads that a student uses, or has access to, during testing. This support may provide assistance to a student who uses sensory tools to calm or alert their sensory system in their daily instruction or other assessment settings.	No research has specifically considered the use of sensory tools as standardized testing supports. However, according to Aspiranti & Hulac (2022); Graziano et al. (2020); Hulac et al. (2020); Kriescher et al. (2023); Rohrberger (2011); and Soares & Storm (2020): • Research suggests that students with ADHD may benefit from using sensory tools. There is some evidence that using fidget spinners may increase on-task behavior for students with ADHD while other evidence found that they may distract general education students leading to poorer attention and lower performance. The risk of the accommodation giving students an unfair advantage is likely low.
Straightedge (Embedded in Test) Students can use the straightedge tool in grades 8 and 11 Mathematics MCA to create a line of best fit. The straightedge tool is now available on all questions.	No research considered the use of a straightedge as a support.
White Noise (Embedded in Test) Students can use the white noise tool to listen to background noise as a way to reduce ambient sound and help them focus on the test. There are three white noise sounds to choose from: waves, white noise, and wind.	No research has specifically considered the use of white noise as a standardized testing support. However, according to Cook et al. (2014); Fox et al. (2015); Helps et al. (2014); Nigg et al. (2024); Rosalez et al. (2020); and Witty (2021): • White noise typically had no benefit on student performance. Some limited evidence with small sample sizes showed possible benefits for students with ADHD. The risk of the accommodation giving students an unfair advantage is likely low.

Table 3.4. Research Base for Accommodations

Support/Accommodation	Research and Recommendations
Approved Accommodation IEP or 504 plan teams determine the need for an accommodation not specifically addressed as an accommodation or support for testing. Approval by MDE is required before the accommodation may be used.	Because these accommodations are individually requested and approved, research is not available.

Support/Accommodation	Research and Recommendations
The use of an approved accommodation is most beneficial for a student who uses the support during instruction.	
Accommodated Text-to-Speech for Mathematics MCA (Embedded in Test) Accommodated text-to-speech provides computer-generated audio for charts and graphs in addition to the audio that is provided for standard text-to-speech; the student can select the parts of the item they want to listen to. With the improvements to standard text-to-speech for Science MCA, accommodated text-to-speech is no longer provided for science. Each individual student's needs must be taken into account when determining which version of text-to-speech will be the most beneficial for them. This accommodation may be beneficial, for example, for a student who is blind and does not yet have adequate braille skills. If not used regularly during instruction and/or the student does not know the content that the items are assessing, this accommodation is likely to be confusing and may impede the performance on assessments. Because accommodated text-to-speech also provides audio for all the labels, graphs, and charts, it may be beneficial only for a limited number of students—for example, those who may require language or visual processing supports. Cognitive overload may occur if a student is presented with too much visual and aural information simultaneously, resulting in the student being unable to process the information. The default for most students should be the standard text-to-speech, which is available as a universal support on the Mathematics MCA.	According to Acosta et al. (2008); Alrawashdeh et al., (2024); Almgren Bäck et al. (2024); Barton (2001); 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); Svensson et al. (2019); Wei (2024);Wolf et al. (2009); and Wood et al. (2018): • 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 students with disabilities benefit from the standard text-to-speech accommodation. The risk of the accommodation giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Assistive Technology (Provided by School or Embedded in Test) Technology that is used to maintain, increase, or improve the functional capabilities of a student with disabilities. This accommodation is used specifically to document assistive technology devices that are used for testing. 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 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. Separate TestNav forms for MCA and Alt MCA Science tests – one for screen readers and one for non-screen readers – allow assistive technology devices to interact directly with the online test. Assistive technology is most beneficial for a student who uses these supports during instruction.	According to Cormier et al. (2010); 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. 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
Paper Test Book (Braille – MCA) (Provided by MDE) IEP or 504 plan teams may select contracted or uncontracted braille test book for students who are blind or partially sighted and are competent braille readers. Braille materials are provided in Unified English Braille format. Both contracted and uncontracted versions are available; contracted braille uses the same letters, punctuation, and numbers as uncontracted, but adds abbreviations, contractions and other short forms of words. A separate phenomena book is provided automatically with the Science MCA paper test book to make it easier for students to access phenomena information alongside the items included in the test book. Students must answer all items in the test book. **Paper Test Materials (Braille – Alt MCA) (Provided by MDE)** IEP teams may select uncontracted braille test materials for students who are blind or partially sighted and are competent braille readers. Braille versions of the Science Alt MCA Presentation Pages, Phenomena Book, and Response Option Cards are provided. Braille materials are available in UEB Uncontracted Math/Science (Science Alt MCA) and UEB Uncontracted (Reading Alt MCA) formats. For Reading and Mathematics MTAS, a braille version of the test materials may be created at the district.	According to Bennett, Rock, and Kaplan (1987); Bennett, Rock, and Novatkoski (1989); Bolt and Thurlow (2004); Coleman (1990); Stone et al. (2010); 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. In Minnesota, where assessments are untimed, timing is less of a concern. The risk of the accommodation giving students an unfair advantage is low.
Paper Test Book (Large Print) (Provided by MDE) A large print test book has an increased font size. Large print test books are available in 18-point (Reading, Mathematics, and Science MCA) and 24-point fonts (Reading and Mathematics MCA only).	According to Beattie et al. (1983); Bennett, Rock, and Jirele (1987); Brown (2007); Burk (1998); Grise et al. (1982); Perez (1980); Stone et al. (2010); 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

Support/Accommodation	Research and Recommendations
For the Science MCA, the 24-point large print test book will not be available for the MCA-IV series. Instead, the 18-point large print test book format will be improved to maximize the layout of the content on an 8.5” x 11” page size, making the materials more manageable for students. A separate phenomena book will also be provided automatically with the Science MCA paper test book to make it easier for students to access phenomena information alongside the items included in the test book. Students must answer all items in the test book. This accommodation is beneficial for a student with low vision who needs an increased font size to see the test items and typically accesses instruction through similar materials, as determined by the IEP or 504 plan team. The use of this accommodation may be considered along with other supports, such as low-vision aids (for example, magnifying glasses).	with low vision and specific learning disabilities significantly benefit from this accommodation. There is no risk the accommodation gives students an unfair advantage.
Paper Test Book (Regular Print) (Provided by MDE) A 12-point regular print test book for Reading, Mathematics, and Science MCAs may be used by any student with an IEP or 504 plan whose disability prevents them from accessing the content on a computer or device. A separate phenomena book will be provided automatically with the Science MCA paper test book to make it easier for students to access phenomena information alongside the items included in the test book. Students must answer all items in the test book. Use of this accommodation must be determined by the IEP or 504 plan team rather than student or administrator preference. In some cases, a student with an injury that prevents them from testing online (for example,	According to Anakwe (2008); Candrlic et al. (2014); Dhitareka et al. (2018); Öz & Özturan (2018); and Prisacari & Danielson (2017): • No differences were observed in student performance, reliability, or validity based on administration type (paper vs. online). According to Pengelley et al. (2023) • Students performed better on a paper versus an online test of similar difficulty. Students performed significantly better on the difficult items on the paper test, while students did slightly better on the easy items for the online test. However, none of these studies specifically considered the use of a regular print paper test book as a test accommodation. More research is needed in this area. The risk of the accommodation giving students an unfair advantage is unclear.

Support/Accommodation	Research and Recommendations
concussion restrictions) may be allowed to use a paper test book if supported by medical documentation, even if they do not have an IEP or 504 plan.	

Support/Accommodation	Research and Recommendations
Allow Student to Read a Reading Test Aloud, Recording Their Voice for Later Playback (Provided by School) A student may record their voice while reading the reading test aloud and replay the recording while taking the test. A student who is a struggling reader, especially one who has difficulty decoding information when reading, may benefit from this support on the Reading MCA, as other read-aloud supports are not allowed.	Goldstone et al. (2021) and Finizio (2008) describe this variation of the student reads aloud to self accommodation, however, there is currently no available data regarding the effectiveness of this accommodation. Further research is needed in this area. The risk of the accommodation giving students an unfair advantage is unclear.

Support/Accommodation	Research and Recommendations
Scribe (Provided by School) Students with disabilities 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. This accommodation may be applicable in instances when visual or motor difficulties, including injuries, prevent the student from indicating their own responses. • Scribes for online tests will navigate the test, access text-to-speech and other tools (as requested by the student), and enter all test item responses provided by the student directly into the online test. • Scribes for paper tests will indicate the responses provided by the student into the paper test book. ○ An audio recording may be used by the student to record and edit answers for paper test materials if the student is unable to mark their answers or uses similar strategies in instruction. Audio recordings cannot be used with the online MCA because each item must have an answer entered in order to move forward in the test. A student with significant motor or processing difficulties may benefit from this accommodation. The student's IEP or 504 plan should document the need for a scribe, except in injury situations. The student should be competent in the use of a scribe as determined by the student's IEP or 504 plan team.	According to Thurlow and Bolt (2001): • Experts recommend that students with disabilities, including students who use ASL, submit answers via computer, whenever possible, rather than relay answers to a scribe. Concerns include the possibility of scribes unintentionally altering student responses, a mismatch between what the student uses during classroom instruction and standardized testing, and uncertainty about how scribes are trained and interpret student work. According to Arcon et al. (2017); Cox et al. (2006); Fuchs et al. (2000); Koretz and Barton (2003, 2004); Koretz and Hamilton (2000); Lai & Berkely (2012); MacArthur and Graham (1987); MacArthur and Cavalier (2004); and Tippets and Michaels (1997): • Students with disabilities research is limited, especially regarding the impact a disability has on test taking. A body of research suggests, however, that students with disabilities 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.

Script/Human Reader (Mathematics and Science) (Provided by MDE) The Test Monitor reads from the script while the student responds directly into the online test (science only) or paper test book (mathematics and science). The script is for use by the Test Monitor only. For Mathematics MCA, because the online mathematics test is adaptive and the test items presented vary by student, the script/human reader accommodation must be administered with a paper test book (regular print, large print, or braille test book). The script requires that Test Monitors read all the text for labels, graphs, and charts, which may be beneficial only for a limited number of students—for example, those who may require language or visual processing supports. Cognitive overload may occur if a student is presented with too much visual and aural information simultaneously, resulting in the student being unable to process the information. For Science MCA, the script/human reader accommodation can be used in conjunction with the online test because the test is not adaptive; it can also be administered with a paper test book (regular print, large print, or braille test book) or with the assistive technology online forms. The scripted language has been updated to remove directional text and provides a comprehensive read-aloud experience for students. This accommodation may be beneficial for a student with reading-related disabilities, or by a student who is blind and does not yet have adequate braille skills. If not used regularly during instruction and/or the student does not know the content that the items are assessing, this accommodation is likely to be confusing and may impede the performance on assessments.	According to Acosta et al. (2008); Barton (2001); Bolt and Thurlow (2004); Brown (2007); Burch (2002); Buzick & Stone (2014); 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); Li (2014); 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 this accommodation helps students with disabilities. It eases both reading and cognitive demands, particularly for items containing extensive written text, although its advantages may be smaller for math and science assessments. The risk of the accommodation giving students an unfair advantage is low.

Support/Accommodation	Research and Recommendations
Script/Human Reader (Signed Interpretation of Mathematics and Science Scripts) (Script provided by MDE for use by district or school interpreter) Signed interpretation (ASL or signed English) of the Mathematics MCA and Science MCA scripts may be provided for students who are deaf or hard of hearing. Interpreters may provide the content directly from text-to-speech in the student's online test, if this is typically how interpretation is provided in instruction, but approval by MDE is required.	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 Cawthon et al. (2011): • Students do not gain an unfair advantage from a signed interpretation of the test, as variance in performance was best explained by subject matter knowledge over standard vs. ASL-accommodated conditions. According to Cawthon et al. (2011), 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 hard of hearing 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.

3.5.1. 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 services or identified as disabled under Section 504 of the Rehabilitation Act of 1973 and (b) ELs.

3.5.2. Data Audit

The data collected is intended to provide MDE with information about districts' use of accommodations on statewide 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.

3.6. Modes of Assessment

The standards-based accountability assessments and English language proficiency 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 accommodated fixed form.

3.6.1. 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.

3.6.1.1. 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 10: 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.

3.6.1.2. 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. Refer to Chapter 3: Test Administration for more information about the WPM and CRM algorithms. 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.

3.6.2. Weighted Penalty Model

The Minnesota CAT assessments are implemented through a CAT system which consists of algorithms such as those for item selection, content balancing, and item exposure control. The WPM is a component of the CAT algorithms used for content balancing for the Minnesota CAT assessments. 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:

$$Prop_j = (nadm_j + Prevalence_j \times nremaining)/testlength \tag{3.1}$$

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:

$$X_j = (Prop_j - Mid_j). \tag{3.2}$$

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 $Upper_j$:

$$P_{ij} = (\frac{1}{kD_j}X_j^2 + \frac{D_j}{k}) \times Z_{ij}, if\ Prop_j < Lower_j \tag{3.3}$$

or

$$P_{ij} = (\frac{1}{kA_j}X_j^2 + \frac{A_j}{k}) \times Z_{ij}, if\ Prop_j \geq Upper_j \tag{3.4}$$

or

$$P_{ij} = X_j \times Z_{ij}, if\ Upper_j > Prop_j \geq Lower_{j.} \tag{3.5}$$

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:

$$F_i''' = \sum_{j=1}^{J} P_{ij} \times w_j, \tag{3.6}$$

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:

$$F_i' = \frac{F_i''' - \min(F_i''')}{max(F_i''') - \min(F_i''')}, \tag{3.7}$$

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:

$$SI_i(\hat{\theta}) = \frac{I_i(\hat{\theta})}{I_{\max(\hat{\theta})}}, \tag{3.8}$$

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:

$$F_i'' = -SI_i(\hat{\theta})^2, \tag{3.9}$$

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

$$F_i = w' \times F_i' + w'' F_i'', \tag{3.10}$$

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.

3.6.3. Conditional Randomesque Method

The CRM is a component of the CAT algorithms for the Minnesota CAT assessments. 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.

3.6.4. 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 8: 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 accommodated assessments, described below, have been placed on the same metric as that of the operational adaptive items. Therefore, scores from the accommodated and adaptive forms of the assessment are on the same metric and can be directly compared.

3.6.5. Online Fixed-Form Assessments

Science MCA-IV in grade 5, grade 8, and high school 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 online fixed 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.

3.6.6. Fixed-Form Assessments for Alternate Assessments and Accommodations

Fixed forms are available for the alternate assessments, MTAS/Alt MCA, and for the online and paper accommodated forms for the MCAs. For student responses to be scored for paper test forms, student responses in paper test books must be entered online into a data entry form by a Test Administrator or district staff. The Mathematics and Reading MTAS-III assessments are administered in a one-on-one setting using a paper test form. The Science Alt MCA is also administered in a one-on-one setting but can be administered using online, paper, or hybrid (any combination of online and paper) modes to meet student needs. In the online mode, the Test Administrator guides the student's interactions and supports the student's use of the online test form, or the student interacts directly with the online assessment with supervision from the Test Administrator. In the paper mode, the student and Test Administrator use only paper materials for testing (i.e., the Task Administration Manual, Phenomena Book, and Presentation Pages). The Test Administrator or district staff then enters the student responses online in TestNav for scoring. In the hybrid mode, both the online test form and paper materials are used to support the student during the test administration.

The MCA-III and MCA-IV assessments have paper accommodated forms that can be administered to eligible students (i.e., students who are unable to take the test on a computer or device). The MCAs also have several online accommodated forms available for eligible students that contain the same content as the paper accommodated forms. Like the online fixed-form assessments, the items on the accommodated fixed forms are fixed prior to administration. One primary difference between the online fixed-form and paper accommodated forms is that the paper accommodated forms are first given in a one-on-one or small group setting and then manually entered into the computer by the Test Administrator or district staff whereas responses to the online assessments, which include the accommodated forms, are provided directly by the student. A second difference is that the paper accommodated forms do not contain field test items, whereas both the online adaptive and online fixed-form assessments do.

Chapter 4: Scoring

Table 4.1 outlines the scoring methods for the spring 2025 administration of the MCAs and the MTAS/Alt MCA, including how the scale scores were calculated as described in Chapter 7. Minnesota's assessments use a combination of machine, human, and automated scoring by the Intelligent Essay Assessor (IEA) automated scoring engine and Math Reasoning Engine (MRE). All dichotomous items (i.e., MC items and TE items) for Reading MCA-III, Science MCA-IV, and MTAS/Alt MCA were machine scored. Most dichotomous items for Mathematics MCA-III were also machine scored, but a small number of math-only CR items were scored using MRE. In addition to dichotomous items, Science MCA-IV also included polytomous items (i.e., CR items) that required scoring by the IEA engine and human scoring.

Table 4.1. Scoring Methods

Test	Scoring Method(s)	Scale Score Calculation
Mathematics MCA-III	Machine-scored dichotomous items Small number of items scored using MRE	IRT pattern scoring
Reading MCA-III	Machine-scored dichotomous items	IRT pattern scoring
Science MCA-IV	Machine-scored dichotomous items CR items scored using a hybrid of automated and human scoring	Number-correct scoring with RSSS tables
Mathematics MTAS-III	Machine-scored polytomous items	Number-correct scoring with RSSS tables
Reading MTAS-III	Machine-scored polytomous items	Number-correct scoring with RSSS tables
Science Alt MCA	Machine-scored dichotomous items	Number-correct scoring with RSSS tables

4.1. Machine Scoring

Machine scoring refers to the scoring of student responses to the dichotomous MC and TE items by the online testing system based on pre-defined scoring rules and answer keys. TE items require scoring models that go beyond traditional answer keys, defined during item development and encoded in the item's XML using rule-based logic. For example, common interaction types include choice interaction (students select one or more options), text entry (students type a brief response), hot spot or text selection (students click or highlight designated regions of an image or text), or match interaction (students match items from two sets into correct pairs). Scoring rules for each interaction type are reviewed during item development and quality control processes.

To ensure the accuracy and reliability of machine scoring, Minnesota's testing service provider conducts a key check and adjudication to verify that all scoring rules and keys are implemented correctly for each item prior to final scoring and reporting. A key check is the process of verifying the correct answer key for MC items, while adjudication is the process of verifying that student responses for TE items are scored appropriately and the functionalities of the items perform correctly.

For each TE item, the frequency distribution of responses scored correctly is created, along with the frequency distribution of responses scored as incorrect. Content specialists analyze for correctness every response string not checked in a previous administration in the frequency reports and indicate whether the response should be scored as correct. The adjudication process typically occurs in two rounds during the middle and end of the testing window for operational items and once at the end of the testing window for field test items. If discrepancies are identified during the key check or adjudication process, MDE content staff and Minnesota's testing service provider's content specialists review the flagged item(s) to ensure that the item is correctly printed (for paper accommodated forms) or displayed (for online forms) and to certify the key is the correct answer.

4.2. Scoring of Constructed-Response Items

Each Science MCA-IV grade-level test form has one or two operational 3-point CR items that required scoring based on holistic rubrics using a combination of human raters and Minnesota's testing service provider's IEA automated scoring engine. IEA is the first score for the operational CR items, with human scoring incorporated in two ways: (a) to conduct a second score on 10% of responses for reliability evaluation and (b) to score routed responses that the engine flagged as low-confidence and require human input, in real time. For the field test CR items, human-scored responses are used to train and calibrate IEA. Once trained using human responses, IEA can be used to score the operational responses to the item in future administrations.

4.2.1. Automated Scoring

Automated scoring refers to the use of machine learning to score students' constructed responses that provides efficiencies over human scoring such as the ability to score millions of responses in a day, thus reducing score turnaround and reporting times. With more than two decades of use, the IEA automated scoring engine has been shown to be as accurate, or more so, than human scorers. It was used to score the Science MCA-IV operational CR items in spring 2025. The IEA automated scoring engine was trained on all CR items based on previous field test responses that were human scored.

4.2.1.1. Intelligent Essay Assessor (IEA) Automated Scoring Engine

The IEA automated scoring technology is used to score written student responses for a wide range of prompt types, including short answers, constructed responses, and essays. Like human scorers, IEA can consider both writing quality and content accuracy, evaluating development, organization, style, and mechanics alongside the key concepts and ideas required by each prompt. Because it can manage multiple writing styles, ranging from argumentative and informational to narrative and persuasive, IEA is an ideal choice for large-scale assessments. Although IEA can consider multiple types of items, it is trained on a specific rubric determined by MDE. In this case, it is trained on a holistic rubric with score points of 0–3.

The IEA team trains IEA using advanced machine learning and natural language processing technologies to score student responses. Separate models are developed for each CR item based on hundreds to thousands of human-scored student responses, including the materials used to train human scorers, enabling IEA and human scorers to score alike. IEA is trained to associate different aspects of student responses with the scores assigned by human scorers. A machine learning-based approach is used to determine the optimal set of features, and the weights for each of those features, to best model the scores for each item. From these comparisons, a prompt- and trait-specific scoring model is derived to predict the scores human scorers would assign to any new responses that have not been previously scored. The result is a precise, standardized scoring method that can be adapted to almost any prompt and rubric arrangement.

4.2.1.2. Engine Training and Model Deployment

Training occurs whenever a new CR item is introduced using hundreds to thousands of student responses that have been scored by professional human scorers. The training datasets encompass a broad spectrum of score points, writing styles, and student ability levels, ensuring robust model performance across various populations. Each model is built from a field test dataset containing the text of student responses and the associated human-assigned scores and uses machine learning algorithms that identify the patterns and features most predictive of those human-assigned scores. These features are derived from linguistic, structural, and semantic aspects of writing, including development, organization, mechanics, and idea relevance. The trained model is then applied to an independent test dataset containing student responses that were not used to train the engine. The model's scores are compared to human scores assigned to that validation dataset to determine whether the candidate model should be used operationally.

During training, IEA is evaluated against the following set of established criteria based on research literature by Williamson et al. (2012) to determine whether it is ready for operational deployment. These measures are computed between pairs of human scores and between IEA and humans to evaluate how performance is the same or different. The human-human results provide a context for interpreting the IEA-human statistics. The goal is for IEA to perform as well or better than the humans.

- Quadratic-weighted kappa (QWK) between IEA-human should be within 0.1 of human-human.
- Absolute standardized mean difference (|SMD|) between IEA-human should be less than or equal to 0.15.

The final model is selected based on its performance on a test dataset that is unseen during engine training, based on the evaluation metrics stated above. If the model meets established criteria, it is deployed to score new responses. Human scorers continue to provide second scores for 10% of responses and scores for low-confidence cases during IEA automated scoring.

4.2.1.3. Continuous Flow

A Continuous Flow scoring solution for the FT and OP items optimizes the integration of human and automated scoring to provide reliable, accurate scoring by brokering communication between IEA and the human scoring platform. This facilitates training and evaluation in real time by allowing responses to move seamlessly between human scorers and the engine. This dynamic system enables the engine to learn from human input, while any complex or uncertain responses are routed to human scorers. As human-scored responses accumulate, IEA continuously updates and refines its predictive models. Once a model satisfies the established evaluation criteria, it is deployed as the primary scoring method for that CR item.

4.2.1.4. Human Oversight

When the engine cannot confidently score a response, that response is marked with a low-confidence flag and routed to a human scorer. This process ensures that only high-confidence automated scores are reported and any uncertain results are reviewed by expert human scorers. Low-confidence scores may result from the following:

- Statistical outliers in feature measures
- Absence of a valid numerical score
- Split-score models where certain score levels are not included in pretraining
- Program-defined constraints on automated scoring
- Cross-trait inconsistencies (e.g., one trait scored high, the other low)
- Responses exceeding the upper bound of the scoring scale

4.2.2. Human Scoring

Human raters score 100% of the Science MCA-IV field test constructed responses used to train IEA, 10% of the operational CR items for the second reliability score, and responses routed for human review during automated scoring for the operational items. These processes are conducted under a rigorous scoring protocol developed with input from MDE. All human-scored responses takes place within Minnesota's testing service provider's Online Scoring and Reporting (OSCAR) system, with daily monitoring and quality assurance conducted by trained supervisory staff. The following roles contribute to the scoring process:

- Scorers apply scores to student responses within OSCAR according to established rubrics and training materials.
- Scoring supervisors monitor scorer performance using statistical reports and backreading procedures.
- Scoring directors oversee the scoring quality, ensure fidelity to training materials, and monitor resolution scoring issues.
- Content specialists ensure consistency and review scoring outputs for accuracy.
- Project managers document procedures, monitor risks, and coordinate the day-to-day scoring operations.

Human scoring involves rangefinding, development of training and qualifying sets, scorer recruitment, scorer training, live scoring, and scorer monitoring. As described in the following sections, the process begins with a rangefinding committee to establish the anchor scores (i.e., anchor sets) based on the scoring rubric that represents exemplar responses to ensure consistency and fairness across scorers. Training sets are then built to show scorers how to apply the rubric consistently across a wide range of student responses, ensuring fair and reliable scoring. Once scorers are recruited based on standard vendor criteria, they are trained using the training sets developed by Minnesota's testing service provider and are required to qualify on qualification sets. Scorers need to pass the qualification criteria before being promoted to live scoring. Once scoring is underway, Minnesota's testing service provider monitors the accuracy and consistency of the scoring following a robust scorer monitoring protocol.

4.2.2.1. OSCAR Scoring System

Constructed responses are scored by humans in the OSCAR scoring system that securely assigns student responses to scorers. Each scorer uses a unique login and completes online training modules on both the scoring system and holistic rubric application prior to live scoring. Scorers view responses, assign scores based on the training and rubrics, and submit the scores through the system. OSCAR then randomly assigns the next student response. No student-identifying information is visible during scoring; all responses are associated with anonymous document IDs. Blank responses are automatically scored as blank to streamline scoring and reduce scorer burden. OSCAR generates detailed reports for internal monitoring and MDE review, including scorer statistics, agreement rates, and frequency distributions.

4.2.2.2. Rangefinding

Rangefinding is the process of preparing anchor papers to guide scorer training and ensure consistent application of the scoring criteria across all scorers. When a new CR item is field tested, Minnesota's testing service provider conducts a rangefinding meeting with Minnesota educators to establish exemplar student responses for each rubric score point, known as anchor sets, that serve as anchor papers for training, qualification, and monitoring during live scoring. Prior to the meeting, Minnesota's testing service provider compiles representative samples of student responses to the field test items as good examples across score point ranges. During the meeting, the committee reviews and scores sample responses using the rubric, discusses interpretations of each score point, and identifies high and low examples within score levels to better understand the ranges and types of responses. MDE approves the resulting rangefinding responses from the committee. These anchor sets inform training sets and qualification materials for future operational use of the CR items to ensure scoring consistency and validity.

4.2.2.3. Scorer Recruitment

Highly qualified scorers are essential to achieving a high degree of consistency in scoring students' responses. All scorers hold at least a bachelor's degree and have completed a formal application process. Prior experience in science education or scoring is preferred. Scorers are required to sign confidentiality agreements and adhere to strict security protocols. Scoring supervisors are selected from the pool of scorers based on demonstrated scoring expertise, communication skills, and leadership ability. For the spring 2025 administration, approximately 30–40 scorers were recruited six weeks before scoring began.

4.2.2.4. Training and Qualification

All scoring personnel complete training, led by Minnesota's testing service provider's content experts, to understand the rubric and score responses accordingly. Training is conducted within the OSCAR scoring system through its video conferencing tool. It begins at the end of April with scoring supervisors who receive in-depth instruction on scoring rubrics, anchor responses, and scorer monitoring protocols to prepare them to support scorers and resolve issues. Scorers then receive comprehensive training that includes an introduction to the assessment and rubrics, a review of scoring criteria and rangefinding anchor sets, guided practice with annotated examples, and independent practice using training sets. Some responses clearly represent the score point, while others represent borderline responses to demonstrate how to define the line between score points.

To qualify for live scoring, scorers complete two qualification sets of 10 responses and are required to meet an exact agreement of at least 80% on at least one of them, as shown in Table 4.4. Scorers who pass qualification begin live scoring one item at a time, using scoring guides to review item-specific expectations before being assigned additional items from the same grade.

Table 4.2. Scorer Qualification Criteria

Assessment	Score Point Range	Minimum Exact Agreement
Science 5	0–3	80%
Science 8	0–3	80%
Science HS	0–3	80%

4.2.2.5. Scorer Monitoring

A multi-layered scorer monitoring process is implemented, as described below. Supervisory staff review reports daily to detect scoring anomalies, support remediation, and ensure fidelity to scoring standards. If a scorer demonstrates consistent disagreement with anchor responses or drifts from expected scoring patterns, retraining is conducted. Retraining includes review of rubric language, anchor sets, and practice sets. Calibration sets are also developed as needed to realign scoring consistency across scorers or rooms. Calibration sets are mini training sets that usually consist of 2–5 responses chosen from live responses that scoring leadership have identified or previous training papers. They are usually created for stabilizing inter-rater reliability, to move scores to achieve an acceptable frequency distribution or to deal with unusual scoring trends or drift.

- Backreading: Supervisors review scores assigned by scorers to detect drift or inconsistency, providing additional coaching or instruction as needed.
- Validity Responses and Reports: Responses scored by the scoring director in advance, serving as a benchmark called "true scores," are inserted into live scoring to evaluate scorer accuracy, and validity reports that show how often a scorer agrees with the "true scores" are monitored throughout scoring. Scorers falling below validity thresholds are retrained or placed on probation.
- Frequency Distribution Reports: Daily reports monitor condition code distributions (e.g., blank, off-topic, not English, insufficient or illegible) and identify outlier scorers who assign score points at rates inconsistent with the group. If a scorer over- or under-uses condition codes compared to the room average, supervisory staff intervene to provide guidance and retraining.

- Second Scoring and Inter-Rater Reliability Reports: These reports track exact, adjacent, and non-adjacent agreement between scorers.

The pre-scored validity responses are randomly assigned to scorers through OSCAR to ensure accurate scoring by individual scorers and avoid scorer drift from the anchor sets. The process for gathering validity evidence is as follows. Scores that are not consistent with project expectations are excluded from final scoring statistics.

- Validity responses are selected, either from rangefinding or concrete score point responses approved by a scoring director.
- The validity set is randomly administered to scorers and supervisors through the OSCAR system.
- Each scorer is required to attain the same criteria as shown in Table 4.4. Any scorer who falls below these criteria are retrained and potentially placed on probation. If a scorer falls below 60%–70% on the validity responses, they could be released from the project.
- A validity report helps identify struggling scorers or alert supervisory staff to group drift. A struggling scorer is below the validity requirement and/or significantly below the group average, while scorer drift refers to scoring responses inconsistently. Scoring directors and supervisors remediate any struggling scorers, and the director considers retraining or calibration of an item if needed.

Lastly, a minimum of 10% of student responses are second scored, with scorers having no indication whether a response has been scored previously, as described in Section 4.2.3. Inter-rater reliability reports are produced based on these second scores.

4.2.3. Inter-Rater Reliability

To evaluate scoring consistency, 10% of the operational constructed responses scored by IEA are randomly selected for second scoring by an independent human rater. During human scoring of the field test responses, inter-rater reliability is also conducted among the human raters. These responses are indistinguishable from first-score responses in the OSCAR system, ensuring that second scorers are blind to prior scores. Inter-rater agreement is evaluated based on the following criteria. If the first and second scores are nonadjacent (i.e., discrepant scores), the scores are escalated for resolution scoring where a third, and if necessary, fourth scorer review the response and assign a score to resolve scorer disagreements, with final adjudication made by a scoring director or content specialist.

- Exact Agreement: Identical scores assigned by both raters, with an expectation of at least 70% of responses having exact agreement
- Adjacent Agreement: Scores within 1 point of each other
- Non-Adjacent Disagreement: Scores differing by 2 or more points

Inter-rater reliability results are used to guide scorer remediation, retraining, or removal from the project. Scorers consistently performing below agreement thresholds are subject to additional monitoring and corrective action. Table 4.3 presents average exact agreement rates from the spring 2025 human scoring of the field test responses. Minnesota's testing service provider provides MDE with secure, ongoing access to scorer performance data throughout the scoring window. Inter-rater agreement is also calculated for the operational online CR items scored by IEA 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 “Item Statistics Reports.”.

4.3. Math Reasoning Engine (MRE)

Minnesota’s testing service provider’s MRE evaluates responses for the equation editor (EE) item type based on rubric criteria specific to the expected item responses and does not require human scoring to train the engine. The engine can interpret the math a student enters into the equation editor response section of the item, so it is unnecessary to list every possible correct response, unlike more traditional machine scoring methods. The MRE scoring criteria are defined in terms of the mathematical characteristics of a response the engine uses to evaluate the extent to which a student response satisfies a scoring rubric (e.g., equivalence, form, precision, and constraints).

Before a mathematics item is scored operationally, it goes through a rubric refinement and validation process. After field testing, online student responses are aggregated into a set of individual unique responses that are backread by human scorers. When the human scorers disagree with the score assigned by the scoring engine, they flag the response for further review. The flagged responses are then reviewed by Minnesota’s testing service provider’s content experts to determine if scoring rules should be adjusted to better fit the rubric. This validation process is held within the same timeframe as the MN adjudication process. The Content team does review the PDF of Top 30 responses for each MRE-scored item during the MN adjudication process to ensure that the item is scoring as expected.

If scoring rule adjustments are made, MRE can rescore the entire item based on the new scoring rules configuration. During rescores, every changed response score is identified and reviewed by content experts to evaluate the impact of the rule change on all responses to ensure that the change had the intended effect and nothing more. The data provided to committee members during data reviews are based on the final scoring rules reflecting how the responses are scored operationally.

Chapter 5: 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/or 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. This chapter provides an overview of the types of scores reported—raw scores, scale scores, and performance levels—along with brief descriptions of each, guidelines for their proper use, and cautions against misuse.

As with any large-scale assessment, the statewide assessments provide a point-in-time snapshot of information regarding student performance. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences about student performance. 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 and WIDA Alternate ACCESS, individual student scores are used to inform continuing eligibility and exit decisions for students receiving English learner services.

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 MCAs and MTAS/Alt MCA are available on the Individual Student Reports (ISRs) Resources page (Minnesota Assessment Hub > Reporting > Individual Student Reports (ISRs) Resources) and the Additional Reporting Resources page (Minnesota Assessment Hub > Reporting Resources > Additional Reporting Resources). WIDA provides scoring and reporting resources on the ACCESS Scores and Reports page (WIDA > Assess > WIDA ACCESS > ACCESS Scores and Reports) and the WIDA Alternate ACCESS page (WIDA > Assess > WIDA Alternate ACCESS). 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).

5.1. Description of Scores

5.1.1. 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, MCA-IV, MTAS-III, and Alt MCA score codes include the following:

- Valid Score (VS)
- Not Attempted (NA)
- Not Complete (NC)
- Valid Score – Early Stopping Rule (VS-ES) (Alt MCA only)
- Wrong Grade (WG)

For the MCAs, 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. New for the 2024-25 school year: The early stopping rule can be used for a very small percentage of students taking the Science Alt MCA with the most significant cognitive disabilities. These students have no consistent, observable system of expressive communication to communicate responses to participate in assessments. The early stopping rule can be applied after the student has no response to the first 6 items so these students do not have to take the entire test. These students receive a VS-ES 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. If a student was administered the incorrect grade cluster domain for their grade in school, Test WES will apply the WG (wrong grade) test code to that domain. 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.

5.1.2. 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 performance levels (MCA, MTAS/Alt MCA) or proficiency levels (ACCESS, WIDA Alternate ACCESS).

5.1.2.1. Raw Score

Raw scores are used for the Science MCA-IV and all MTAS/Alt MCA 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, substrand, or content area (also called subscore[6]) 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 subscore. They cannot be compared across tests or administrations because test items and their levels of difficulty may vary across different tests and 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 scale score and subscore performance levels 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 and subscore performance levels. Therefore, raw number-correct scores are not utilized for the Mathematics and Reading MCA-III.

5.1.2.2. Scale Score

For the MCAs and MTAS/Alt MCA, 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 a performance level, which delineates whether a student met the performance standards in a manner that is fair across forms and administrations. The performance levels for MCA-III and MTAS-III are *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, and *Exceeds the Standards*. The performance levels for MCA-IV and Alt MCA are *Beginning*, *Intermediate*, *Meets*, and *Advanced*. 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 and MCA-IV subject and grade range from G01 to G99, where G is the grade tested. For each MCA assessment, scale score G50 is the cut score for *Meets the Standards/Meets*, and G40 is the cut score for *Partially Meets the Standards/Intermediate*. The MCA cut score for *Exceeds the Standards/Advanced* 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-IV are transformations of raw number-correct scores. More than one raw score point may be

[6] The term of subscore can refer to strand performance or content areas.

assigned the same scale score, except at cut scores for each performance 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. The scale scores for Alt MCA range from 100 to 300. For each MTAS-III and Alt MCA assessment, a scale score of 200 is the cut score for *Meets the Standards/Meets*, and 190 is the cut score for *Partially Meets the Standards/Intermediate*. The MTAS-III and Alt MCA cut score for *Exceeds the Standards/Advanced* can vary by grade and subject. As with the MCAs, MTAS-III and Alt MCA scale scores from different grades and subjects are not directly comparable.

The meaning of scale scores is tied to the grade-specific content and performance levels associated with a given set of Minnesota Academic Standards. Thus, MCA-II scores are not directly comparable to MCA-III scores and the MCA-III scores are not directly comparable to the MCA-IV scores because those scores reflect different content and performance standards. Similarly for MTAS-III and Alt MCA, the scores from assessments based on different academic standards are not directly comparable. Refer to Chapter 7: 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.

5.1.2.3. Performance/Proficiency Levels

To help families and schools interpret scale scores, performance levels are reported. performance levels provide insight into how a student's test performance relates to grade-level standards. For the MCAs and MTAS/Alt MCA, the student's scale or raw score determines each performance level, also sometimes referred to as proficiency level. The range for a performance level is set during the standard setting process. The first year that a new series of assessments developed to align with the updated academic standards is implemented, panels of Minnesota educators set the cut scores for performance levels and created the performance level descriptors (PLDs), 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 performance on the MCAs and MTAS/Alt MCA. For each test, certain performance levels are designated as demonstrating proficiency of grade-level standards. Refer to Chapter 6 Standard Setting for details on standard setting processes.

Table 5.1 presents a summary of the performance levels for the Minnesota statewide assessments. For individual students, the performance 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 performance 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/Beginning* performance level, for example, or who move down in performance 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 MCAs and MTAS/Alt MCA 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 5.1. Performance Levels for the Minnesota Statewide Assessments

Test	Performance/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
MCA-IV, Alt MCA	*Beginning*	No
	Intermediate	No
	Meets	Yes
	Advanced	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*	No
	Level 2: Emerging	Yes**
	Level 3: Developing	Yes**
	Level 4: Expanding	Yes
	Level 5: Bridging	Yes

*Proficiency on the ACCESS requires an overall composite ≥ 4.5 and at least three of the four domains ≥ 3.5. If only one domain is < 3.5 but the overall composite is ≥ 4.5, additional evidence must be considered for an English Learner (EL) services exit decision. For more information, see the MDE EL Education page (MDE > Districts, Schools and Educators).
**Proficiency on the WIDA Alternate ACCESS requires an overall composite ≥ 2. If an overall composite is 2 or 3, additional evidence must be considered for an English Learner (EL) services exit decision. For more information,

see the Minnesota Department of Education EL Education page (MDE > Districts, Schools and Educators > English Learner Education).

5.2. 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 the MCAs and MTAS/Alt MCA is PearsonAccess Next, which provides dynamic data that can be used to gauge students' performance 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.

5.2.1. 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, these reports provide scale scores and performance level designations associated with the student's performance. Subscore performance is also included for each student. Table 5.2 presents the available student-level reports. Sample student reports can be found on the Individual Student Reports (ISRs) Resources page (Minnesota Assessment Hub > Reporting > Individual Student Reports (ISRs) Resources) for MCA and MTAS/Alt MCA and the ACCESS Scores and Reports page (WIDA > Assess > WIDA ACCESS > ACCESS Scores and Reports) for ACCESS or WIDA Alternate ACCESS page (WIDA > Assess > WIDA Alternate ACCESS) for WIDA Alternate ACCESS.

Table 5.2. Student-Level Test Reports

File or Report	Format	Applies to MCA-III and MTAS-III?	Applies to MCA-IV and Alt MCA?	Applies to ACCESS and WIDA Alternate ACCESS?
On-Demand Reports	Online (PDF and excel file)	Yes (within 60 minutes after testing or data entry is completed)	No[7]	No

[7] On-Demand Reports are not available in the first operational administration due to standard setting but will be available in future years.

File or Report	Format	Applies to MCA-III and MTAS-III?	Applies to MCA-IV and Alt MCA?	Applies to ACCESS and WIDA Alternate ACCESS?
Early Student-Level Results	Online (excel file)	No	No	Yes (beginning of Posttest Editing in late May)
Individual Student Report (ISR)[8]	Paper and online PDF for ISR	Yes	Yes	Yes
Students Result Labels (Optional)	Paper and online PDF	Yes, paper by request	Yes, paper by request	Not available
District Student Results (DSR) and School Student Results (SSR)	Excel file	Yes	Yes	Yes
Student Assessment History Report	Online and Excel file	Yes	Yes	Yes

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

- On-Demand Reports for Reading and Mathematics MCA-III/MTAS-III
- ISR for Reading/Mathematics MCA-III/MTAS-III and Science MCA-IV/Alt MCA-IV
- Student Results Label (optional) for Reading/Mathematics MCA-III/MTAS-III and Science MCA-IV/Alt MCA-IV

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

5.2.1.1. On-Demand Reports/Preliminary Results

For the 2024-25 school year, On-Demand Reports were only available for Reading and Mathematics MCA-III and MTAS-III. On-Demand Reports are not available in the first operational administration of an assessment due to standard setting. On-Demand Reports provide a student's preliminary scale score and performance level in

[8] 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.

PearsonAccess Next within 60 minutes after testing or data entry is completed. Information on subscores 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/Alt MCA Test Administrator 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 and WIDA Alternate 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 once the Posttest Editing window opens. Early results are not updated during Posttest Editing, and final student-level results are available in the District & School Student Results (DSR & SSR) in Assessment Secure Reports.

5.2.1.2. 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/Alt MCA begin with overall results, including a description and visuals representing student performance level on the grade-level standards. For the 2024-25 school year, two-page ISRs separated by subject were created for Reading and Mathematics MCA-III and MTAS-III, while Science MCA-IV and Alt MCA-IV ISRs underwent a complete redesign. Sample ISRs are posted under Reporting on the Individual Student Reports (ISRs) Resources page (Minnesota Assessment Hub > Reporting > Individual Student Reports (ISRs) Resources).

A personalized video ISR resource with individual student results for Reading and Mathematics MCA-III and MTAS-III is available for each student with at least one valid score. For the 2024-25 school year, due to timing of standard setting and release of results, students who completed the Science MCA-IV or Alt MCA received a generic video detailing the standard setting process. 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 for Reading and Mathematics MCA-III and MTAS-III 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 are available in 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 is presented in a bar graph along with a description of the assigned performance level next to the graphic. For Reading and Mathematics MCA-III, school, district, and/or state average scale scores are presented on the same graphic for comparison. For Science MCA-IV, the percentage of students demonstrating proficiency in the grade-level standards at the school, district, and state is provided. The MCA ISRs provide all three averages, whereas the MTAS/Alt MCA ISRs provide the state average only 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/Alt MCA, the number of students is frequently quite small for school or district populations.

Next, the MCA ISRs include descriptions of each subscore and the subscore performance level for each reported as *Below Expectations*, *At or Near Expectations*, and *Above Expectations*. The MTAS/Alt MCA 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 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 assessment results for students. Reading and Mathematics MCA-III and MTAS-III ISRs were provided to districts in late August, while Science MCA-IV and Alt MCA-IV ISRs were provided on Oct. 1 following standard setting.

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.[9] 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. 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.

5.2.1.3. 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 MCAs and MTAS/Alt MCA include the test name, test date, student information, scale scores, and performance level for each subject tested for a single test. The individual

[9] Districts with the largest enrollments of students identified as EL can have reports shipped to schools.

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 the *Procedures Manual*). All districts can access electronic PDFs of Student Results Labels under Published Reports in PearsonAccess Next.

5.2.1.4. 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 history for students who are currently enrolled in the district, based on student enrollment data submitted to MDE.

5.2.1.5. 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 MCAs, MTAS/Alt MCA, ACCESS, and WIDA Alternate ACCESS, including demographic information, student performance information, and test scores.

5.2.2. 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 or MCA-IV 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

Nonsecure summary-level reports for the MCAs provided on the Minnesota Assessment Hub include the following:

- Subscore reports for MCA-III and MCA-IV

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/Alt MCA
- Test Results Summary for ACCESS, WIDA Alternate ACCESS, MCA, and MTAS/Alt MCA

Nonsecure summary reports are also available from MDE's 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/Alt MCA
- Assessment Files for ACCESS, WIDA Alternate ACCESS, MCA, and MTAS/Alt MCA

5.2.2.1. 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 for ACCESS and WIDA Alternate ACCESS.

5.2.2.2. Subscore Reports

The Subscore Report (Minnesota Assessment Hub > Reporting > Subscore Report) provides the public access to school-, district-, and state-level MCA subscore data, also known as strand performance levels/content area performance details, for a selected test, year, grade, and subject. The subscore performance levels are determined by comparing the school or district performance to the state expectation at the *Meets* performance level. The subscore performance levels are reported as *Below Expectations (B)*, *At or Near Expectations (N)*, and *Above Expectations (A)*. For each grade and subject, this report includes the percentage of students at B, N and A levels for each subscore calculated by aggregating the B. N, and A 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 performance level.

5.2.2.3. Alternate Assessment Participation

The Alternate Assessment Participation report (MDE website > Data Center > Secure Reports) provides the district's MTAS/Alt MCA 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.

5.2.2.4. 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 accountability assessments.

5.2.2.5. 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. Assessment participation summary data is available through the Minnesota Report Card.

5.2.2.6. Assessment Files

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

5.3. Appropriate Uses of Assessment Results

The Minnesota statewide assessments are summative assessments that provide data about student performance from one point in time and are designed primarily to determine school and district accountability related to equitable implementation of the academic standards (meaning 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 (the WIDA English Language Development Standards or the Minnesota Academic Standards).

The assessments provide a snapshot of a student’s overall performance, 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 student’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 performance/proficiency levels over time between student groups
- Percent of students by performance/proficiency level within a subject or grade level
- Change in students moving between performance/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.

5.3.1. Individual Students

Individual student results show a broad overview of student learning of grade-level standards and these results are intended to be interpreted alongside more fine-grained information about the individual student from their classroom teacher. Federal and state legislation requires that individual student results be reported for all statewide assessments; therefore it is 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. When making educational decisions about individual students, a comprehensive approach should be taken, and other classroom-level assessment and instructional data should be used in conjunction with statewide assessment data.

5.3.2. 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-IV results cannot be directly compared to previous MCA-III administrations because the MCA-IV assesses different academic standards. The same holds true for the MTAS-III and Alt MCA assessments. For example, the grades 3-8 Mathematics MTAS-III began assessing new standards in 2011 and thus cannot be directly compared to the grades 3-8 Mathematics MTAS-III from prior years. 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 5.3. Comparing MCA and MTAS/Alt MCA Assessment Results from Year to Year

Subjects and Assessments	Series Number	Grades	Academic Standards Last Revised	First Year Assessed on Revised Standards	Years Scores are Comparable
Mathematics MCA and MTAS	III	3-8	2007	2011	2011 to 2025
Mathematics MCA and MTAS	III	11	2007	2014	2014 to 2025
Reading MCA and MTAS	III	3-8, 10	2010	2013	2013 to 2025
Science MCA and MTAS	III	5, 8, HS	2009	2012	2012 to 2024
Science MCA and Alt MCA	IV	5, 8, HS	2019	2025	2025 and beyond

- When individual student graduation stakes associated with high school MCA tests changed in 2013-14, students no longer needed to achieve proficiency as a way 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 receiving special education services 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 5.4 summarizes this information.

Table 5.4. Comparing ACCESS, Alternate ACCESS, and WIDA Alternate ACCESS Assessment Results from Year to Year

English Language Proficiency Test	Subjects	Years Scores Are Comparable
ACCESS	All Subjects (Domains)	2021-2025
Alternate ACCESS	All Subjects (Domains)	2021-2023
WIDA Alternate ACCESS	All Subjects (Domains)	2024-2025

The percentage of students in each performance 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 performance 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* performance 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 performance level comparisons across years, particularly because testing program changes in content alignment are accompanied by changes in the definition of performance 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 subscore 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 subscore, 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 subscores 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 performance 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.

5.4. 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.

5.4.1. 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-IV and the accommodated forms of the MCAs and MTAS/Alt MCA), 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 8: 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 year, measurement error is factored in when determining the subscore performance levels. Refer to Chapter 7: Scaling for details on the subscore performance 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 10: 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.

5.4.2. 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-IV 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. Therefore, it's important to be cautious when interpreting scores from students at the high or low ends of the scale. For a more accurate understanding of their performance, it's best to consider these scores alongside data from district, local, and classroom assessments.

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.

5.4.3. Interpreting Score Means and Variability in Performance and Percent Proficient

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. For a more precise interpretation of student performance, it's preferable to look at the percentage of students who are proficient instead of relying on the average scale score.

5.4.4. Using Subscore-Level Information

Subscore-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 subscore level, the error of measurement increases because the number of possible items is small. To provide comprehensive diagnostic data for each subscore, 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 subscore-level performance should be made cautiously. Significant effort is made to approximate the overall difficulty of the subscores from year to year during the test construction process, but fluctuations in difficulty can occur across administrations. Observing trends in subscore-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 subscore information. For example, observing trends in the proportions of students at the three different subscore performance 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.

5.4.5. 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 assessments 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 6: Standard Setting

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). Performance standards are set to determine the level of performance students must demonstrate to be classified into the defined performance levels. During standard setting, these levels are defined by cut scores that mark the minimum level of performance a student must meet to be classified into a given level.[10]

The MCA-III and MTAS-III have four performance levels: *Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, and *Exceeds the Standards*. The MCA-IV and Alt MCA have four performance levels: *Beginning*, *Intermediate*, *Meets*, and *Advanced*. 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*). (This section summarizes the standard setting process and results for the MCA and MTAS/Alternate MCA only.)

Table 6.1 presents 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 the table. This chapter summarizes the process for establishing the performance 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 6.1. Standard Setting Meetings

Test	Grades	Date	Location	Method	Content 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
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
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-IV	5, 8, high school	July 8–10, 2025	Eagan, Minn.	Bookmark	2019

[10] Achievement levels (MCA-III and MTAS) are the equivalent of performance levels (MCA-IV and Alt MCA) or proficiency levels (ACCESS and WIDA Alternate ACCESS) depending on the assessment. Similarly, achievement level descriptors (ALDs) for MCA-III and MTAS are equivalent to performance level descriptors (PLDs) for MCA-IV and Alt MCA. When summarizing the process for defining the levels across the different series of assessments in this technical manual, “performance levels” and “PLDs” are used for simplicity.

Test	Grades	Date	Location	Method	Content Standards
Science Alt MCA	5, 8, high school	July 14–15, 2025	Eagan, Minn.	Modified ID Matching	2019

6.1. Process Components

Minnesota's testing service provider, MDE, and MDE's TAC worked together to design the standard setting activities for each assessment, with the testing service provider facilitating the standard setting under the supervision of MDE. The Bookmark standard setting procedure (Lewis et al., 1996) was chosen for the MCA-III and MCA-IV assessments, the ID Matching method (Ferrara et al., 2002) was used for most MTAS-III assessments, and a slightly modified version of the ID Matching method was used for Alt MCA. Although similar to the widely implemented Bookmark method, the ID Matching procedure asks panelists to indicate which performance level is best matched by the knowledge and skill requirements necessary to respond successfully to each test item based on the PLDs. 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.

This section provides an overview of the general components included in a Minnesota standard setting and the steps followed in the process. Details for a specific standard setting meeting 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)).

6.1.1. Selecting a Method

There are a variety of standard setting methods, all of which require the judgment of content and education experts, who are commonly referred to as panelists during standard setting meetings. The key differences among the various standard setting methods can be understood through contrasting examples. The most cited example 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 example 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 examples , 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, which are items that provide examinees with a set of predefined answer options such as multiple-choice or true-false formats (Cizek, 2001; Hambleton and Plake, 1997). In contrast, the policy-capturing method, which is a standard setting approach that models the decision rules judges use when evaluating performances defined by several observable task features, 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 committee 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.

Before selecting the standard setting method, Pearson and MDE conducted a study in February 2024 comparing the Bookmark method and a modified version of the Item Descriptor (ID) Matching method to determine which one is most suitable for the upcoming standard settings (Moyer & Reid, 2024). Based on these results and subsequent consultation with the Technical Advisory Committee (TAC) in April 2024, MDE selected the Bookmark standard setting method for the MCA-IV assessments and Modified ID Matching for Alt MCA assessments.

6.1.2. 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 committee 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 standard settings, MDE convenes educator committees to recommend performance standards, or cut scores, for each assessment. Each committee is organized by grade or grade band that has a lower and upper grade for which panelists set standards. Each committee also has its own facilitator and is physically separate from the other committees. MDE invites approximately 10–30 panelists from across Minnesota for each committee to set cut scores for each assessment. Details of the credentials and demographics of the panelists for each meeting 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)).

6.1.3. Table Leaders

During the standard setting, panelists are divided into groups called *tables*. Each table has one table leader selected by MDE who is expected to promote and keep track of the table-level discussion and represent their committee's point of view during committee-level discussions, vertical articulation, and/or community and educational partners (CEP) meetings. Table leaders are trained on their roles and responsibilities on Day 1 of the standard setting.

6.1.4. Ordered Item Sets

Central to both the Bookmark and ID Matching procedures is the production of an ordered item set (OIS) that typically presents the items from easiest to most difficult in terms of their theta difficulty (an exception is the OIS in the modified ID Matching method that presents the items in administration order during Round 1 so that items related to the same stimulus can be reviewed together). While the OIS is often produced from the operational items from the first operational administration of an assessment, 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. As such, the operational items can be supplemented by items from various sources (e.g., operational items from a different form or field test items) in the OIS. While recommending cut scores on the entire item bank is ideal in some respects, including too many items may cause cognitive overload when reviewing the OIS. Therefore, the OIS is designed to be of reasonable length with a uniform distribution of item difficulty across the theta scale, ensuring comprehensive coverage of the ability continuum.

6.1.5. Feedback Data

Some methodologies provide student performance data to panelists for decision-making, while other types of feedback include consequential data (i.e., impact data), 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. Table 6.2 summarizes the feedback data presented during each standard setting meeting.

Table 6.2. Summary of Feedback Data by Round

Assessment	Method	Round	Data Presented
MCA-III	Bookmark	Round 2	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 only)
MTAS-III	ID Matching Modified Angoff	Round 2	R1 panelist feedback data
		Round 3	R2 panelist feedback data Series-II MTAS historical impact data Series-III MTAS-III operational impact data

Assessment	Method	Round	Data Presented
MCA-IV	Bookmark	Round 1	Individual cut scores Panelist cut score agreement
		Round 2	Individual cut scores Panelist cut score agreement Impact data
		Round 3	Panelist cut score agreement Impact data
Alt MCA	ID Matching	Round 1	Individual item judgments Panelist item judgment agreement
		Round 2	Individual item judgments Panelist item judgment agreement Panelist cut score agreement Impact data
		Round 3	Panelist cut score agreement Impact data

6.1.6. Standard Setting Process

MDE and Minnesota's testing service provider began each standard setting meeting by briefing the committees on the purpose of the meeting and use of the outcomes, advising panelists that (a) the principal outcome was a set of cut score recommendations, (b) the educator committees were one of many components in the complete policymaking process of standard setting, and (c) their cut score recommendations are reviewed for overall consistency and continuity by Minnesota's Education Commissioner who finalizes the cut scores. 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 committee to begin the process of setting performance standards.

The MCA-III and MCA-IV standard setting meetings were conducted in three rounds. Panelists used the PLDs to generate threshold descriptors, or "just barely" descriptions, as a group. The "just barely" descriptions characterize the knowledge and skills of students at the threshold or borderline of each level (i.e., the cut score). To create the "just barely" descriptions, panelists simplified PLD statements to describe the minimum skills at each performance level. Working in groups, they drafted "just barely" descriptions for each level, which were then reviewed and refined as a whole group for clarity and coherence across the levels. The finalized descriptions were printed for reference during judgment rounds. After creating the threshold descriptors and completing the standard setting training, experience assessment activities, and practice activities, the committee began the process of setting performance standards. During the judgment process, panelists use the "just barely" descriptions to evaluate each item in a set of test items ordered by item difficulty and estimate whether such a student would answer the item correctly with a response probability of at least 0.67 (RP67), or a 67% likelihood of a correct response. The 67% threshold reflects a balance between demonstrating understanding while allowing for occasional errors, making it intuitive and realistic for panelists (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006).

After the Round 1 cuts were made, psychometricians and statistical analysts evaluated the results and produced feedback data for each table and the committee as a whole. After completing discussions on the Round 1 feedback, panelists again worked through the OIS, placing their cut scores for *Meets /Meets*, followed by *Partially Meets/Intermediate* and *Exceeds/Advanced*. After Round 2, in addition to the feedback data presented in the previous round, impact data presenting the percentage of students at or above the level for each possible cut score was provided for discussion. After completing discussions on the Round 2 feedback, panelists again worked through the OIS and provided cut score recommendations beginning again with the *Meets /Meets* level.

The MTAS-III and Alt MCA standard settings were also conducted in a series of three rounds but used the ID Matching or modified ID Matching method (except for grades 3–8 Mathematics MTAS-III that used Modified Angoff in the first two rounds and Reasoned Judgment in Round 3). During the standard setting process, panelists identified the threshold region (also known as the region of uncertainty) between performance levels by indicating the first item in the OIS that clearly matched the *Meets /Meets* level and the last page that clearly matched the *Partially Meets/Intermediate* level. After identifying the threshold region, panelists examined each item in the threshold region to determine the first item that more closely matches the PLDs for *Meets/Meets* than the PLDs for *Partially Meets/Intermediate*. 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 all levels.

After completion of the experience assessment, practice, and training activities, panelists were provided with the OIS and other materials associated with their committee. For the Alt MCA standard setting process, the OIS was developed around a response probability of 0.50, or RP50. The use of RP50 is particularly appropriate for the alternate assessment student population that has a wide range of abilities. RP50 reflects a more balanced and interpretable point on the difficulty scale, reducing the risk of misclassifying students by avoiding unrealistic expectations of learning and instead focusing on demonstrated understanding that aligns more closely with the design intent and accessibility goals of the Science Alt MCA.

Once they indicated their readiness to make their independent judgments, panelists began their Round 1 activities. During Rounds 2 and 3, panelists discussed their recommendations in small groups at their tables. Panelists were provided with table-level and committee-level feedback on their previous round recommendations and discussed their recommendations at their table and at the committee level, with the goal of identifying major sources of variation among panelists. Impact data was also provided during Rounds 2 and 3. Impact data provides panelists insight into the percentage of students who would fall into each performance level based on the current recommended cut scores. Understanding, rather than consensus, was the goal of the discussion. After the discussion, panelists again identified their cut scores. Panelists were reminded that this was an individual activity. The Round 3 cut scores were considered the committees' final recommendations that were reviewed at a vertical articulation committee and/or a CEP meeting.

6.2. Mathematics MCA-III and MTAS-III, Grade 11

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 panelists from across Minnesota to set cut scores for each assessment. The standard setting activities and results are recorded in the standard setting report (MDE, 2014). Operational items common to the 2014 online and paper accommodated test form administration modes served as the base for the MCA-III OIS. The OIS was augmented with 21 additional operational items selected from other operational forms 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 OIS. This led to an OIS that included 77 items for the Mathematics MCA-III. The MTAS-III OIS included both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS-III item bank.

6.2.1. Recommended Cut Scores

Table 6.3 presents the panelist-recommended cut scores for the grade 11 Mathematics MCA-III and MTAS-III after Round 3. Cut scores are shown on the theta metric. The final MTAS-III cut scores were identified by selecting the observed theta score nearest to the theta value associated with the panelists' recommended page number in the OIS. The nearest observed theta in the operational test form raw-score-to-theta table is the final recommended cut. Table 6.4 presents the impact data and the percentage of students in each performance category based on the cut scores after Round 3.

Table 6.3. Panelist-Recommended Cut Scores on the Theta Metric for Mathematics Grade 11

Assessment	*Partially Meets*	*Meets*	*Exceeds*
Mathematics MCA-III, Grade 11	-0.5371	0.1034	0.9989
Mathematics MTAS-III, Grade 11	1.0260	1.6731	2.8329

Table 6.4. Impact Data Associated with Panelist-Recommended Cut Scores for Mathematics Grade 11

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

6.2.2. Commissioner Approval

After the standard setting meeting, Minnesota' Education Commissioner 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.

6.3. Mathematics MCA-III and MTAS-III, Grades 3–8

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). Committees were organized by grade band, with approximately 12–15 panelists per committee. 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 the standard setting report (MDE, 2011). The MCA-III OIS contained 60 operational items from the 2011 MCA-III assessments that spanned the range of content, item types, and difficulty represented on a typical test. The MTAS-III task book contained all the operational tasks from the 2011 MTAS-III, ordered in the same sequence as they appeared on the test.

6.3.1. Recommended Cut Scores

Table 6.5 presents the panelist-recommended cut scores based on the panelists' Round 3 bookmark placements. Cut scores are shown on the theta metric for MCA-III and the raw score metric for MTAS-III. Table 6.6 presents the associated impact data.

Table 6.5. Panelist-Recommended Cut Scores (Round 3) on the Theta Metric for MCA-III and on the Raw Score Metric for MTAS-III for Mathematics Grades 3–8

Assessment	*Partially Meets*	*Meets*	*Exceeds*
Mathematics MCA-III, Grade 3	-1.21	-0.51	0.61
Mathematics MCA-III, Grade 4	-1.05	-0.43	0.42
Mathematics MCA-III, Grade 5	-0.86	-0.03	1.04
Mathematics MCA-III, Grade 6	-0.72	0.06	0.95
Mathematics MCA-III, Grade 7	-1.19	0.08	0.95
Mathematics MCA-III, Grade 8	-0.82	-0.03	0.84
Mathematics MTAS-III, Grade 3	13	17	24
Mathematics MTAS-III, Grade 4	14	17	24
Mathematics MTAS-III, Grade 5	12	19	25
Mathematics MTAS-III, Grade 6	11	17	24
Mathematics MTAS-III, Grade 7	12	18	21
Mathematics MTAS-III, Grade 8	12	16	21

Table 6.6. Impact Data Associated with Panelist-Recommended Cut Scores for Mathematics Grades 3–8

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MCA-III, Grade 3	14	17	41	28
Mathematics MCA-III, Grade 4	17	17	32	34
Mathematics MCA-III, Grade 5	21	27	36	15
Mathematics MCA-III, Grade 6	25	27	30	17
Mathematics MCA-III, Grade 7	14	38	30	18
Mathematics MCA-III, Grade 8	22	26	31	21
Mathematics MTAS-III, Grade 3	15	13	38	34
Mathematics MTAS-III, Grade 4	14	8	52	26

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MTAS-III, Grade 5	12	31	45	12
Mathematics MTAS-III, Grade 6	15	24	51	11
Mathematics MTAS-III, Grade 7	15	30	28	27
Mathematics MTAS-III, Grade 8	18	12	37	33

6.3.2. Vertical Articulation

Vertical articulation is an evaluation of the standard setting results across grades from the different grade-level committees. Panelists review the impact data based on the recommended Round 3 cut scores and discuss the reasonableness of the cut scores across grades. They have the opportunity to adjust the cut scores to better match expectations for student performance across the grade levels. Vertical articulation helps ensure results are meaningful and demonstrate a logical progression of student performance across the grade levels. Articulation panelists are stakeholders in the results of the assessment system from a broad range of perspectives. Members of an articulation committee 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 committee 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 and 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 performance level in a grade were provided for review and discussion.
9. The panelists discussed 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 6.7 presents the cut scores from the vertical articulation, and Table 6.8 presents the associated impact data. Cut scores are shown on the theta metric for MCA-III and the raw score metric for MTAS-III.

Table 6.7. Vertical Articulation Committee's Smoothed Cut Scores on the Theta Metric for MCA-III and on the Raw Score Metric for MTAS-III for Mathematics Grades 3–8

Assessment	*Partially Meets*	*Meets*	*Exceeds*
Mathematics MCA-III, Grade 3	-1.22	-0.52	0.60
Mathematics MCA-III, Grade 4	-1.06	-0.44	0.57
Mathematics MCA-III, Grade 5	-0.88	-0.04	1.01
Mathematics MCA-III, Grade 6	-0.75	0.03	0.96
Mathematics MCA-III, Grade 7	-0.91	0.03	0.94
Mathematics MCA-III, Grade 8	-0.83	-0.03	0.83
Mathematics MTAS-III, Grade 3	13	17	24
Mathematics MTAS-III, Grade 4	14	18	24
Mathematics MTAS-III, Grade 5	12	19	25
Mathematics MTAS-III, Grade 6	11	17	23
Mathematics MTAS-III, Grade 7	12	18	21
Mathematics MTAS-III, Grade 8	12	17	21

Table 6.8. Impact Data Associated with Articulation Committee's Smoothed Cut Scores for Mathematics Grades 3–8

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MCA-III, Grade 3	14	17	41	28
Mathematics MCA-III, Grade 4	17	17	37	29
Mathematics MCA-III, Grade 5	21	27	36	16
Mathematics MCA-III, Grade 6	24	27	32	17
Mathematics MCA-III, Grade 7	20	30	32	18

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MCA-III, Grade 8	22	26	31	21
Mathematics MTAS-III, Grade 3	15	13	38	34
Mathematics MTAS-III, Grade 4	14	13	47	26
Mathematics MTAS-III, Grade 5	12	31	45	12
Mathematics MTAS-III, Grade 6	15	24	45	17
Mathematics MTAS-III, Grade 7	15	30	28	27
Mathematics MTAS-III, Grade 8	18	18	32	33

6.3.3. Commissioner Approval

After the standard setting meeting, Minnesota's Education Commissioner reviewed the recommended cut scores for overall consistency and continuity. Table 6.9 presents the final cut scores approved by the commissioner for the 2011 MCA-III and MTAS-III administrations on the theta metric, and Table 5.10 presents the impact data associated with the final cut scores from 2006 for the MCA-III and from 2011 for the MTAS-III.

Table 6.9. Commissioner-Approved Cut Scores on the Theta Metric for Mathematics Grades 3–8

Assessment	*Partially Meets*	*Meets*	*Exceeds*
Mathematics MCA-III, Grade 3	-1.22	-0.52	0.60
Mathematics MCA-III, Grade 4	-1.06	-0.44	0.57
Mathematics MCA-III, Grade 5	-0.88	-0.04	1.01
Mathematics MCA-III, Grade 6	-0.75	0.03	0.96
Mathematics MCA-III, Grade 7	-0.91	0.03	0.94
Mathematics MCA-III, Grade 8	-0.83	-0.03	0.83
Mathematics MTAS-III, Grade 3	0.22	0.92	2.31
Mathematics MTAS-III, Grade 4	0.56	1.27	2.61
Mathematics MTAS-III, Grade 5	0.17	1.54	3.13
Mathematics MTAS-III, Grade 6	0.19	1.60	2.74
Mathematics MTAS-III, Grade 7	0.51	1.62	2.11
Mathematics MTAS-III, Grade 8	0.42	1.42	2.10

Table 6.10. Impact Data Associated with Commissioner-Approved Cut Scores for Mathematics Grades 3–8

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MCA-III, Grade 3	14	17	41	28
Mathematics MCA-III, Grade 4	17	17	37	29
Mathematics MCA-III, Grade 5	21	27	36	16
Mathematics MCA-III, Grade 6	24	27	32	17

Assessment	*Does Not Meet* (%)	*Partially Meets* (%)	*Meets* (%)	*Exceeds* (%)
Mathematics MCA-III, Grade 7	20	30	32	18
Mathematics MCA-III, Grade 8	22	26	31	21
Mathematics MTAS-III, Grade 3	15	13	38	34
Mathematics MTAS-III, Grade 4	14	13	47	26
Mathematics MTAS-III, Grade 5	12	31	45	12
Mathematics MTAS-III, Grade 6	15	24	45	17
Mathematics MTAS-III, Grade 7	15	30	28	27
Mathematics MTAS-III, Grade 8	18	18	32	33

6.4. 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). Committees were organized by grade band (3–4, 5–6, 7–8, and 10), with approximately 10 panelists per committee. The standard setting activities and results are recorded in the standard setting report (MDE, 2013).

Operational items from one of the 2013 test administration online fixed forms served as the base for the MCA-III OIS. The grades 3–8 OIS was augmented with two additional operational passages and corresponding items selected from other operational forms. MDE selected additional passages for inclusion in the OIS 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 OIS. This led to an OIS that included 59–70 items across grades 3–8. At grade 10, cut scores were recommended based on the single paper form so the core of the OIS comprised the operational items contained in that form. The grade 10 paper OIS was augmented using one field test passage and associated items from that form, which led to a total of 57 items. The grades 5–8 OIS was created based on one of the two forms and additional field test items administered on the 2013 tests, leading to an OIS with 47–53 items across the grades.

The MTAS-III OIS included both operational and field test items to more fully represent the range of academic achievement encompassed within the MTAS-III item bank.

6.4.1. Recommended Cut Scores

Table 6.11 presents the panelist-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, final cut scores were identified by selecting the nearest observable theta to the theta value associated with the panelists' recommended page number in the OIS. The nearest observable theta in the operational test form raw-score-to-theta table is the final recommended cut. Table 6.12 presents the impact data, or the percentage of students in each performance level based on the cut scores after final moderation for the Reading MCA-III and MTAS-III.

Table 6.11. Panelist-Recommended Cut Scores on the Theta Metric (Final Moderation) for Reading

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

Table 6.12. Impact Data Associated with Panelist-Recommended Cut Scores (Final Moderation) for Reading

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

6.4.2. Vertical Articulation and Moderation

Following Round 3 for the initial grades, a vertical moderation session was conducted to allow table leaders to evaluate the recommended cut scores in the context of a system of standards across grade levels. During the vertical moderation process, table leaders reviewed the impact data based on the recommended Round 3 cut scores and discussed the reasonableness of the cut scores across the initial grades. Following the evaluation of recommended cut scores across the initial grade levels (grades 3, 5, 7, and 10), table leaders from each committee 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), a final moderation session was held to allow table leaders to evaluate the entire system of performance standards and make any final revisions. This vertical articulation process allowed table leaders the opportunity to evaluate standard-setting results across all grades, discuss their appropriateness, and adjust them to align with student performance expectations at each grade. This process ensures results are meaningful and show a logical progression across grade levels.

6.4.3. Commissioner Approval

After the standard setting meeting, Minnesota's Education Commissioner reviewed the recommended cut scores for the 2013 MCA-III administration for overall consistency and continuity and approved all the panelist-recommended cut scores.

6.5. Science MCA-IV and Alt MCA

Standard setting for the grade 5, grade 8, and high school Science MCA-IV took place on July 8–10, 2025, in Eagan, Minn., using the Bookmark standard setting procedure (Lewis et al., 1996). Standard setting for the Science Alt MCA took place on July 14–15, 2025, in Eagan, Minn., using a modified version of the ID Matching procedure (Ferrara et al., 2008). Committees were organized by grade (5, 8, and high school), with 13–18 panelists per committee. The standard setting activities and results are recorded in the standard setting report (Pearson, 2025). For each committee, Minnesota's testing service provider developed an OIS with operational items and used field test items to fill in gaps in the difficulty distribution, resulting in OISs containing 55–65 items for MCA-IV and 31–33 items for Alt MCA.

6.5.1. Recommended Cut Scores

Table 6.13 presents the panelist-recommended cut scores for the Science MCA-IV and Alt MCA, as taken from Round 3. Table 6.14 presents the associated impact data with the cut scores. Cut scores are shown on the theta metric.

Table 6.13. Panelist-Recommended Cut Scores on the Theta Metric (Round 3) for Science

Assessment	*Intermediate*	*Meets*	*Advanced*
Science MCA-IV, Grade 5	-0.0882	0.8149	1.8159
Science MCA-IV, Grade 8	0.2564	1.1191	1.8919
Science MCA-IV, High School	-0.1161	0.4710	1.4558
Science Alt MCA, Grade 5	-0.423343	0.476414	1.073627
Science Alt MCA, Grade 8	-0.915870	-0.123162	0.862686

Assessment	*Intermediate*	*Meets*	*Advanced*
Science Alt MCA, High School	-0.498639	0.364329	1.017367

Table 6.14. Impact Data Associated with Panelist-Recommended Cut Scores for Science

Assessment	*Beginning* (%)	*Intermediate* (%)	*Meets* (%)	*Advanced* (%)
Science MCA-IV, Grade 5	44.2	33.5	19.0	3.4
Science MCA-IV, Grade 8	56.1	30.4	11.1	2.3
Science MCA-IV, High School	44.3	20.4	29.0	6.2
Science Alt MCA, Grade 5	13.5	34.2	23.7	28.8
Science Alt MCA, Grade 8	5.1	18.1	39.9	36.8
Science Alt MCA, High School	8.9	25.2	29.3	36.6

6.5.2. Community and Educational Partners Committee

The purpose of the CEP meeting was to share the impact data based on the Round 3 recommended cut scores with Minnesota stakeholders, discuss its implications, and gather recommendations for public communication. The combined Science MCA-IV and Alt MCA CEP meeting took place on July 16, 2025, with table leaders from both standard setting committees and members of the Minnesota education and business communities. The lead facilitator explained the standard setting process and presented an overview of the assessments, while the table leaders were responsible for explaining the reasoning behind the recommended cut scores. Participants offered useful perspectives on how the results may be perceived and what information should accompany them to provide the context necessary for accurate interpretation.

6.5.3. Commissioner Approval

After the CEP meeting, Minnesota's Education Commissioner reviewed the recommended cut scores for overall consistency and continuity and approved the panelist-recommended cut scores on July 18, 2025.

Chapter 7: Scaling

The MCAs and MTAS/Alt MCA 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 6: Standard Setting, cut scores 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. Although test forms across years may follow the same blueprint outlined in the test specification and measure the same content standards, 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, no test forms are 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 8: Equating and Linking).

7.1. 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, such as a scale-score metric, so that test forms from different years can be equated.

Mathematics and Reading MCA-III do not use the raw score but instead use a model-based score (pattern scoring) to calculate the theta score as the initial score. However, tests like the Mathematics and Reading 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 a performance level that represents the degree to which students meet the Minnesota grade-level academic standards. For accountability assessments such as the MCAs and MTAS/Alt MCA, the final scaling results are a designation of performance levels (*Does Not Meet the Standards*, *Partially Meets the Standards*, *Meets the Standards*, or *Exceeds the Standards* for MCA-III and MTAS-III and *Beginning*, *Intermediate*, *Meets*, and *Advanced* for MCA-IV and Alt MCA).

7.2. Measurement Models

Item response theory (IRT) is used to derive the scale scores for all 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 also facilitates the creation of equivalent forms across years—forms that maintain comparable content representation and test difficulty. .

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 multiple-choice (MC), fill-in-the-blank (FIB), and technology-enhanced (TE) items. Models designed for use with items that allow multiple scores are called *polytomous models*. Both dichotomous and polytomous IRT 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 multi-point performance tasks. 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 generalized partial-credit (GPC) IRT model (Muraki, 1992) for the Science MCA-IV 3-point CR items.

7.2.1. Rasch Models

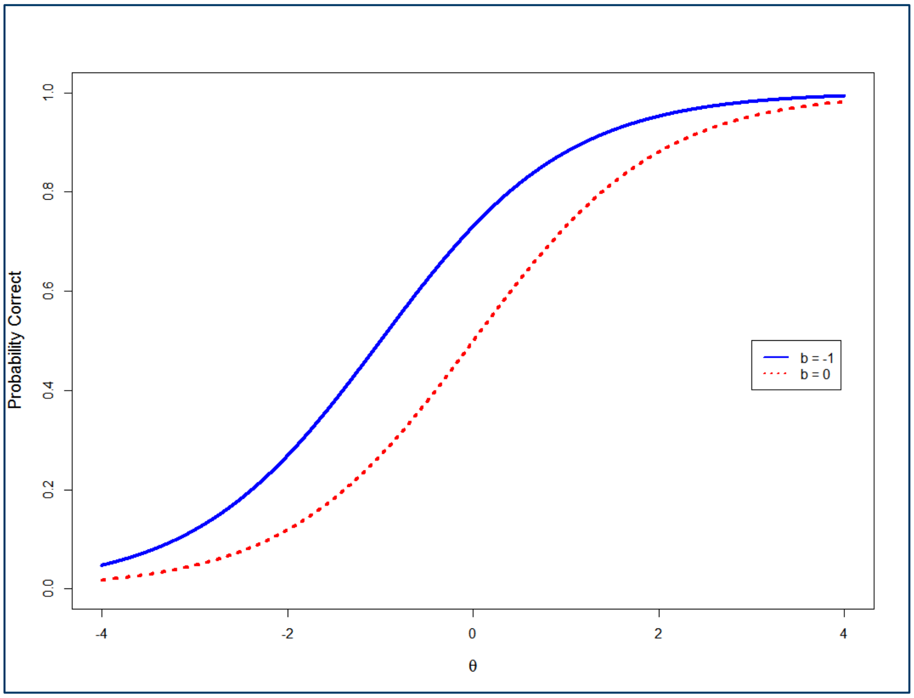
The dichotomous Rasch model can be written as the following mathematical equation, where the probability (P_{ij}) of a correct response for person *i* taking item *j* is given by:

$$P_{ij} = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)} = \frac{1}{1+\exp[-(\theta_i - b_j)]} \tag{7.1}$$

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. The b-parameter in item response theory (IRT) is typically defined as the location on the ability scale at which the probability of selecting a correct answer is 50%. Figure 7.1 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*-parameter value of -1.0, and the dotted curve represents an item with a *b*-parameter of 0.0. A *b*-parameter 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*-parameter 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*-parameter 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*-parameter value for an item have a low probability of getting the answer correct.

Figure 7.1. 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:

$$P_{ijx} = \frac{exp\sum_{k=0}^{x}(\theta_i - b_{jk})}{\sum_{v=0}^{m_j-1} exp\sum_{k=0}^{v}(\theta_i - b_{jk})} \tag{7.2}$$

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

$$\boldsymbol{\sum_{k=0}^{0}(\theta_i - b_{jk}) = 0.} \tag{7.3}$$

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 7.2 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 7.2. Rasch Partial Credit Model Category Response Functions for Example Polytomous Item With b1 = -1.5, b2 = -0.3, b3 = 0.5, and b4 = 2

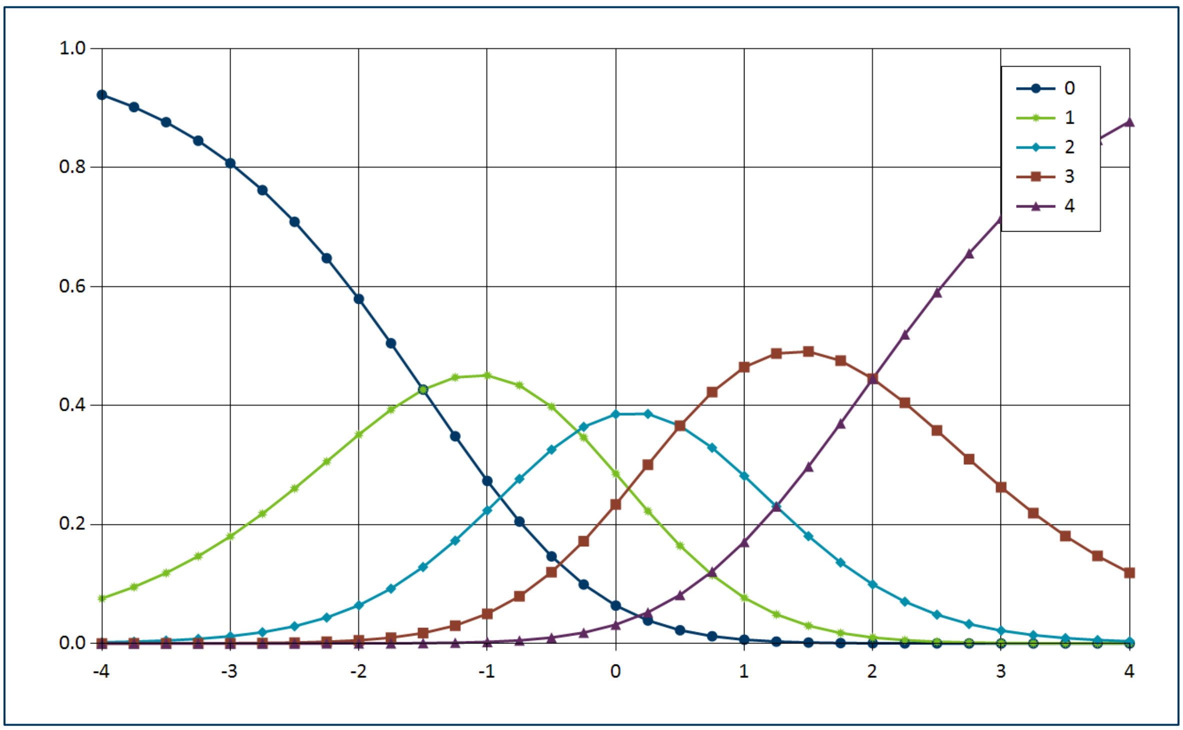
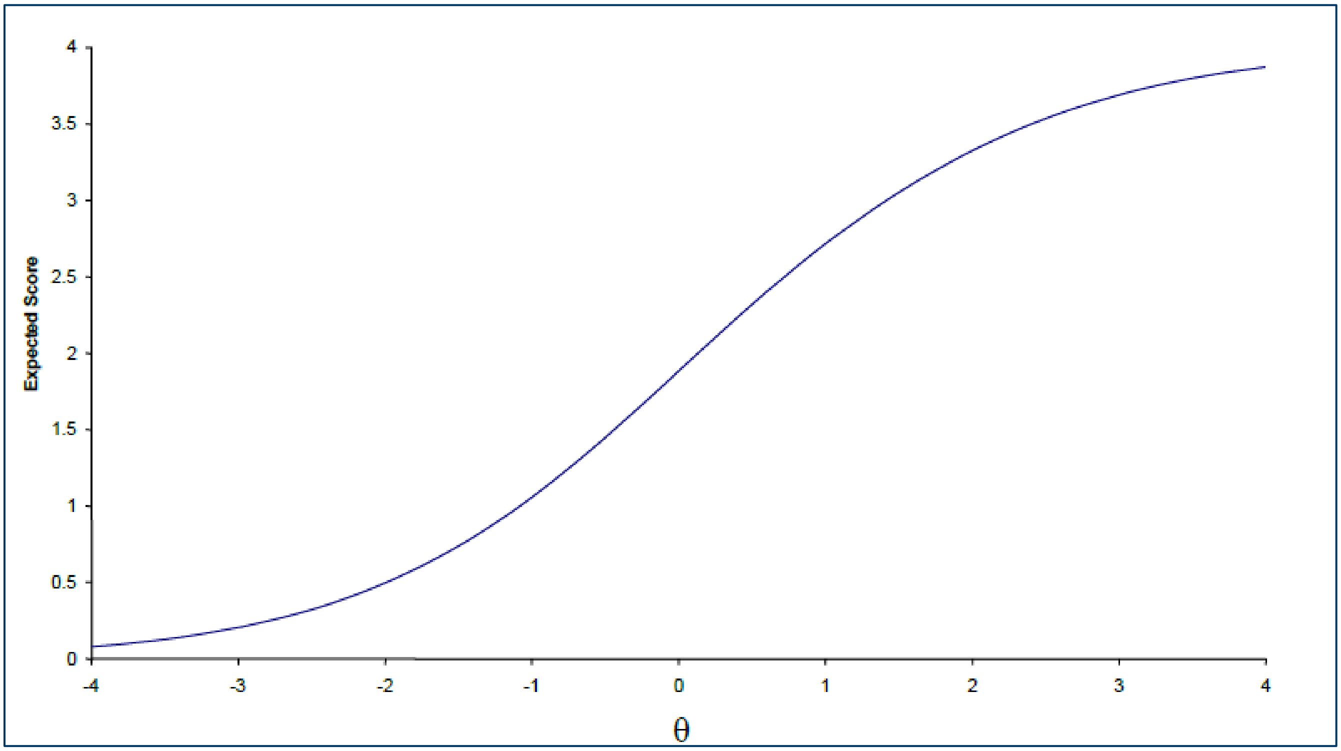


Figure 7.3 presents the average score for every ability value for the sample item given in Figure 7.2. The figure shows that students with ability θ = 0 should, on average, receive a score of "2" on the item, whereas students with ability at about 1 should average about 2.5 points on the item.

Figure 7.3. 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, 2023). The program estimates item difficulty for MC items and category boundary parameters for polytomously scored items such as multi-point performance tasks or CR items. The dichotomously scored operational items in Alt MCA and field test items in both the MTAS and Alt MCA use the Rasch model. The polytomously scored operational items in MTAS use the RPC model.

7.2.2. 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, while GPC is used for the Science MCA-IV 3-point CR 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 with higher discrimination parameter better 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 dichotomously that are not in multiple choice format, such as FIB items, the 2PL model may be more 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*:

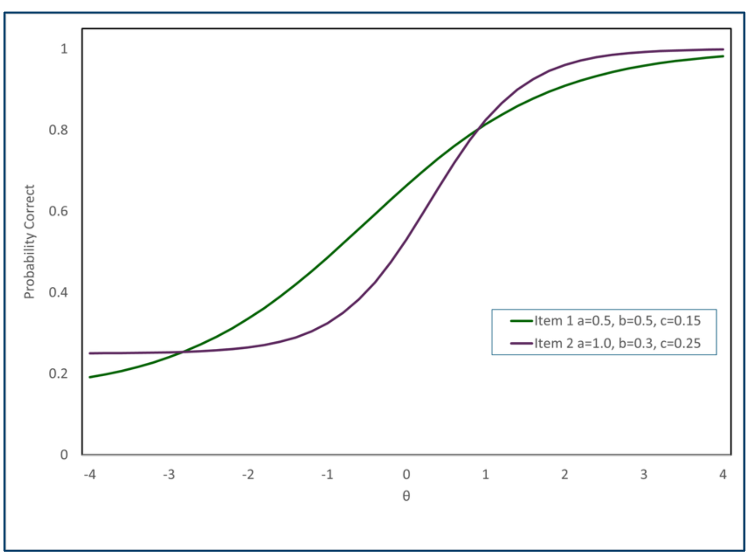
$$P_{ij} = c_j + \frac{1-c_j}{1+\exp\left(-1.7a_j\left(\theta_i - b_j\right)\right)} \tag{7.4}$$

where a_j, b_j, c_j are the item's slope (discrimination), location (difficulty), and lower asymptote parameters (guessing), 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 7.4 can be reduced to the standard Rasch equation (7.1) by setting *c* = 0, *a* = 1, and removing the 1.7 scaling constant.

Figure 7.4 presents examples of 3PL model item-response functions. Several differences from the Figure 7.1 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 7.4 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 7.4. 3PL Item Response Functions for Two Sample Dichotomous Items



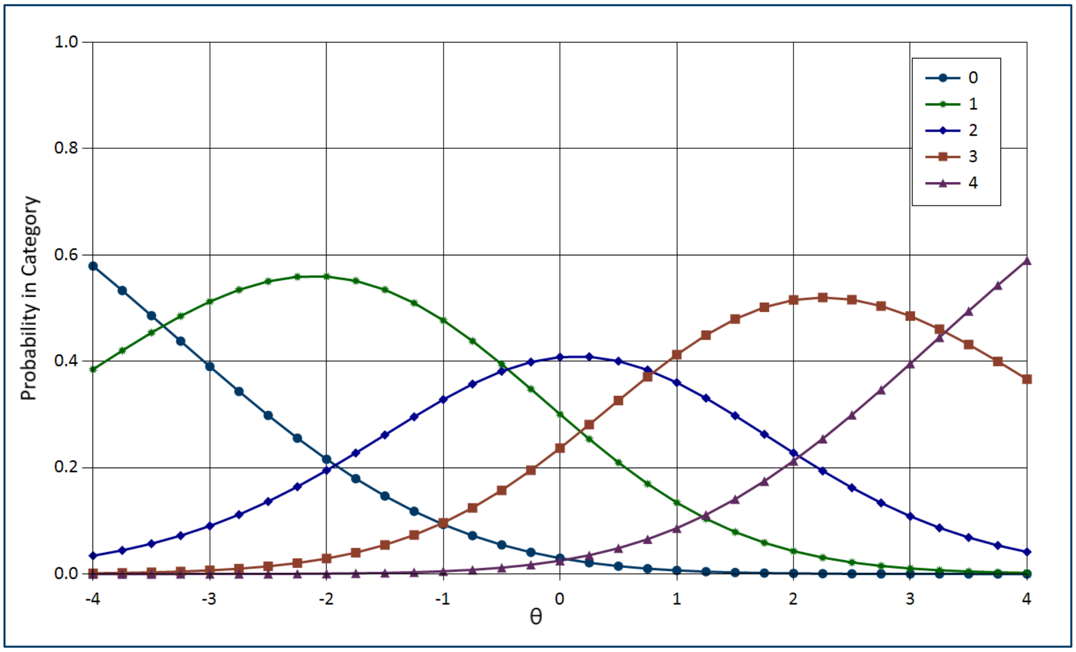
The polytomous IRT model described in this section is the GPC model. Instead of having a single probability correct, as in the 3PL model, the GPC model has a separate probability for each possible response category. The GPC model is mathematically defined as the probability of person *i* scoring in response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

$$P_{ijk} = \frac{\exp\left[\sum_{v=1}^{k} 1.7a_j(\theta_i - b_j + d_{jv})\right]}{\sum_{c=1}^{m} \exp\left[\sum_{v=1}^{c} 1.7a_j(\theta_i - b_j + d_{jv})\right]} \tag{7.5}$$

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 7.5 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 7.5, for example, d_2 (= 3.7) is the distance on the θ scale that the crossing point for the "zero" and "1" curves is from the location parameter (*b* = 0.3); the *b*-parameter for this item is 3.7 units greater than the value of θ at the crossing point. As another example, *b* is one-half of a unit less than the value of θ at the crossing point for the response functions for scores of "2" and "3" (because 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 7.5. 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

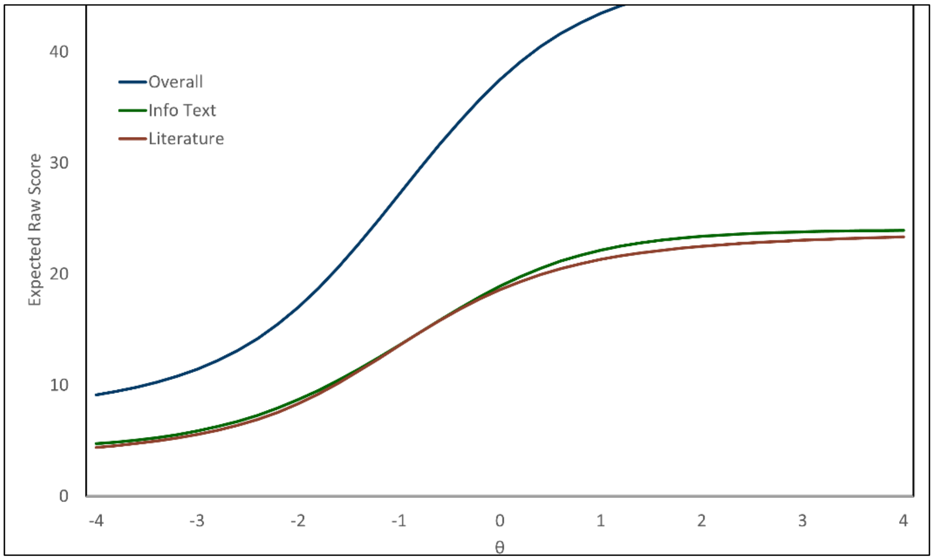


Calibration of MCA items for the 2PL/3PL/GPC models is achieved using the computer program IRTPRO 6.0.6 (Cai et al., 2022). IRTPRO estimates parameters simultaneously for dichotomous and polytomous items via a statistical procedure known as marginal maximum likelihood. Simultaneous calibration of these 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.

7.2.3. 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 7.6 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 7.6.

Figure 7.6. 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 the MTAS/Alt MCA tests because they are administered to relatively few students. The Rasch model generally performs better than more complex models when sample sizes are small.

7.3. 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 (raw score to scale score table) or a measurement model-based score (pattern scoring). These two initial scores are described below.

7.3.1. Number-Correct Scoring

Scale scores for fixed forms, such as the Science MCA-IV and MTAS/Alt MCA, 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/task. 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 2025 administration is the base year for Science MCA-IV and Alt MCA. In subsequent administrations, the ability score metric is equated back to the spring 2025 base administration.

In the case of assessments based on the Rasch measurement model (MTAS/Alt MCA), the number-right and model-based scoring approaches are mathematically equivalent. In other words, if two students have the same number of correct scores, they will always receive the exact same model-based ability estimate. 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.

7.3.2. 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 use 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.

7.3.3. 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 Mathematics and Reading MCA-III 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.

7.3.3.1. 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 Mathematics and Reading MCA-III are assigned maximum likelihood scores that are based on the items the student answers correctly and the difficulty of those items. The assessments include multiple item types such 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:

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR} \tag{7.6}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N} \left[c_i + \frac{1-c_i}{1+exp[-Da_i(\theta-b_i)]}\right]^{x_i} \left[1 - c_i + \frac{1-c_i}{1+exp[-Da_i(\theta-b_i)]}\right]^{1-x_i} \tag{7.7}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N} \frac{exp \sum_{k=0}^{x_i}(\theta_i - b_{jk})}{\sum_{v=0}^{m_j-1} exp \sum_{k=0}^{v}(\theta_i - b_{jk})} \tag{7.8}$$

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.

7.3.3.2. Raw-to-Theta Transformation

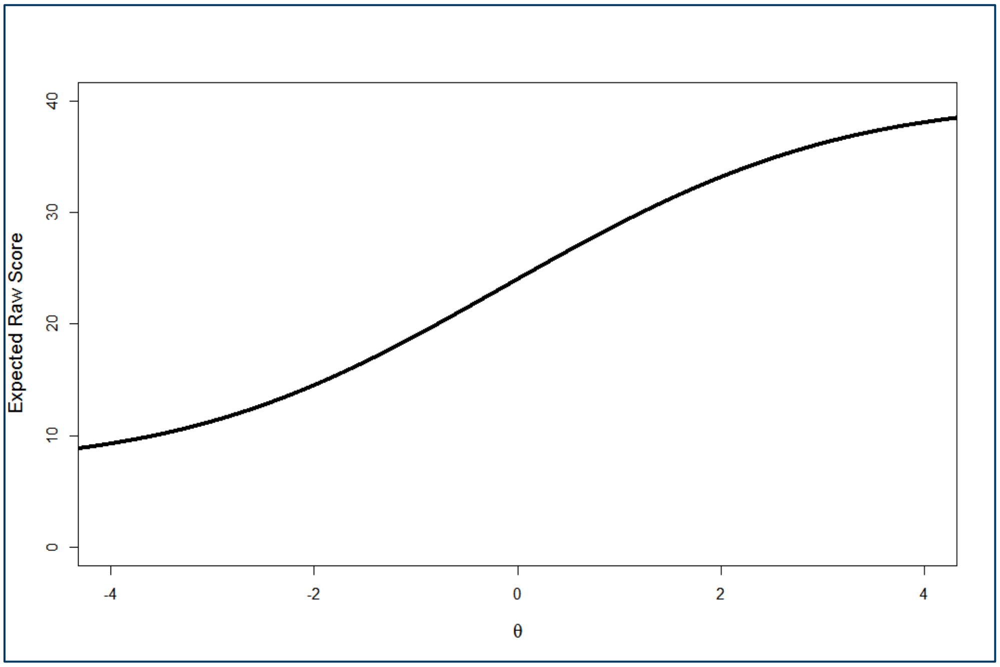
The raw-to-theta transformation can be described as a reverse table lookup on the test characteristic function, defined as follows:

$$\mathrm{TCF}(\theta) = \sum_{j=1}^{N} \sum_{K=0}^{m-1} kP_{ik}(\theta) \tag{7.9}$$

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 7.7 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, 2023) or POLYEQUATE (Kolen, 2004) are used to find the transformations.

Figure 7.7. Example Test Characteristic Function for 40-Item Test



7.4. MCA Scaling

To simplify comparison of student scores across years, the equated student ability estimates are transformed mathematically to a more convenient metric. For MCA-III and MCA-IV, 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/Meets* is set to g50, where g is the grade prefix. The cut score to achieve *Partially Meets/Intermediate* is set to g40. At grade 3, for example, students scoring below 340 are designated *Does Not Meet/Beginning*, students with scores from 340 to 349 are designated *Partially Meets/Intermediate*, and a score of 350 to the next cut score is necessary to achieve *Meets/Meets*. The *Exceeds/Advanced* performance level score is not set to the same value across grades, but it generally ranges from g59 to g74.

7.4.1. Transformation

The general transformation formula used to obtain scale scores for the MCAs is the following:

$$Scale\ Score = (\theta - \theta_{Std2}) * Spread + Center + Grade * 100 \tag{7.10}$$

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. The transformation formula uses cut scores on the θ scale. For Mathematics and Reading MCA-III and Science MCA-IV, Minnesota's Education Commissioner approved cut scores that were already on the θ scale.

One goal for the scale transformation was to make the performance level scale score cuts as consistent as possible across grades. Using a linear transformation-like equation (7.10) allows two of the three scale cut scores to be fixed. As stated above, the cut score for *Meets/Meets* 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/Beginning* and *Partially Meets/Intermediate* 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:

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})} \tag{7.11}$$

where θ_{Std1} is the theta ability cut score between *Does Not Meet/Beginning* and *Partially Meets/Intermediate*, and θ_{Std2} is the theta ability cut score between *Partially Meets/Intermediate* and *Meets/Meets*. 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/Meets* and *Exceeds/Advanced* 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.

In order to construct raw-score-to-scale-score conversion for Science MCA-IV, some additional scoring rules are necessary. For Science MCA-IV, 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-IV. The second-highest, number-correct scores less than all correct should obtain a scale score that is one scale score lower 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 scale score 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 dichotomous 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-IV 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.

Furthermore, 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.

7.5. MTAS/Alt MCA Scaling

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

$$Scale\ Score = (\theta - \theta_{Std2}) * Spread + Center \tag{7.12}$$

where θ is the post-equated ability estimate, $\theta_{Std2}$is the ability cut score between *Partially Meets/Intermediate* and *Meets/Meets*, *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 MTAS/Alt MCA use the same transformation equation.

Chapter 6: Standard Setting describes the process of setting the performance standards for MTAS-III and Alt MCA, a procedure culminating in the Education Commissioner approving the cut scores. The ability cut scores corresponding to the Education Commissioner's approved raw score cuts were used to set the MTAS-III and Alt MCA scales.

As with the MCAs, the aim was to make the performance level scale score cuts as consistent as possible across grades. Using a linear transformation-like equation (7.12) allows two of the three scale cut scores to be fixed. For all grades and subjects of MTAS/Alt MCA, the cut score for *Meets/Meets* was set to 200 by setting *Center* = 200. The cut score between *Does Not Meet/Beginning* and *Partially Meets/Intermediate* was set to be equal to 190. 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/Intermediate* and the *Meets/Meets* scale score cuts. The *Spread* constant for each grade and subject combination was selected to force the first scale cut score to be equal to 190. The formula used to find the *Spread* is as follows:

$$Spread = \frac{10}{(\theta_{Std2} - \theta_{Std1})}, \tag{7.13}$$

where θ_{Std1} is the theta ability cut score between *Does Not Meet/Beginning* and *Partially Meets/Intermediate*, and θ_{Std2} is the theta ability cut score between *Partially Meets/Intermediate* and *Meets/Meets*. 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/Meets* and *Exceeds/Advanced* was allowed to vary across grades and subjects.

For Alt MCA, the LOSS is set to 100 and the HOSS is set to 300. Because Alt MCA uses raw-score-to-scale-score conversion, some additional scoring rules are necessary. Restrictions are placed on the transformation for very high and very low scores. For example, the scale score associated with the raw score of zero is always set to the LOSS. A raw score of all correct is always assigned to the HOSS, regardless of the result of the transformation equation. The scale score for any lower-than-perfect raw score is set to 299 if it is equal to or higher than HOSS.

7.6. Subscores

The primary goal of each assessment is to provide an indicator of student progress in each subject area. Subject area performance is reported as the total scale score and performance 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 subscore reporting.

7.6.1. Subscore Reporting

The MTAS/Alt MCA assessments report subscores as raw score (i.e., number correct) points. Subscores are reported as raw score points for MTAS/Alt MCA because the limited number of raw score points for each subscore does not provide enough information to convert the raw scores into more interpretable scale scores. Since each subscore contains limited points, transforming raw scores into scale scores can result in subscale estimates that lack reliability or consistency.

As with subject area scores, subscores 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 subscore information that can be more meaningfully interpreted, the MCAs report subscore level performance on a common scale that reflects relative achievement across the student population. In 2011 and 2012, Mathematics MCA-III subscore 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 subscores. For Science MCA-IV, MLE scoring is used to estimate subscore theta values.

For the MCAs, two indicators of subscore performance were previously provided: a subscore scale score and a subscore performance level classification. Starting with the 2025 administration, only the performance level classification will serve as the indicator of subscore performance; subscore scale scores will no longer be reported.

7.6.2. Subscore Performance Levels

Beginning in 2016, MDE has reported subscore-level PLDs for the MCAs in the ISRs. Subscore 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 a performance level. Both the student subscore 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" level. Both $\hat{\theta}$ and $CSEM(\hat{\theta})$ are described in greater detail in Chapter 10: Reliability. The computation of the lower and upper limit is made to the fourth decimal place as follows:

$$Lower\ Limit = \hat{\theta} - CSEM(\hat{\theta}) \tag{7.14}$$

$$UpperLimit = \hat{\theta} + CSEM(\hat{\theta}). \tag{7.15}$$

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

$$Below\ Expectations\ (B): Upper\ Limit < Score\ Target \tag{7.16}$$

$$Above\ Expectations\ (A): Lower\ Limit > Score\ Target \tag{7.17}$$

$$At\ or\ Near\ Expectations\ (N): Lower\ Limit \leq Score\ Target \leq Upper\ Limit. \tag{7.18}$$

Table 7.1. Score Targets of Subscore Performance 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

Grade	Strand	Score Targets
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 7.2. Score Targets of Subscore Performance 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 7.3. Score Targets of Subscore Performance Levels for Science MCA-IV

Grade	Strand/Substrand	Score Targets
5	ES	0.8149
5	LS	0.8149
5	PS	0.8149
8	ES	1.1191
8	LS	1.1191
8	PS	1.1191
HS	LS1	0.4710
HS	LS2	0.4710
HS	LS3	0.4710
HS	LS4	0.4710

As can be seen from the above computations of the lower and upper limits, two students can differ in their performance 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 performance levels.

Caution is always required when interpreting subscale performance. Because some subscale performance is based on a few items, the error of measurement increases. That is, individual student subscale estimates used to generate the subscale performance levels 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 performance for a student. Using subscore performance 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.

7.7. 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 602, 2023-2024 Administration.* The scaling procedures for WIDA Alternate ACCESS are available online in the following report on the WIDA website (wida.wisc.edu/resources): *Annual Technical Report for WIDA Alternate ACCESS English Language Proficiency Test Series 602, 2023-2024 Administration*.

7.8. Scale Score Interpretations and Limitations

Because the on-grade scale scores associated with the MCA 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 (7.11) and (7.13), 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, performance 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*), but scale score gain comparisons between individual students are not appropriate.

For assessments that use raw-to-scale score conversions (i.e., MTAS/Alt MCA and Science MCA-IV), 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 are either delivered through fixed forms or administered adaptively, attempts to interpret scale scores in raw score terms are generally inappropriate.

The main purpose of the scale score is to show how close a student is from each performance level, without relying on raw scores that can vary from one test form to another.. Scale scores do not change in their representation of proficiency across years, whereas raw scores do not maintain their proficiency level meaning across years.

Finally, it must be emphasized that there are substantial differences in test content and scoring metrics across the MCA series of assessments (i.e., MCA-IV, MCA-III and MCA-II). These differences should discourage attempts to draw inferences based on score comparisons between students now taking the MCA-III or MCA-IV 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 Science MCA-IV score from 2025 to the grade 5 Science MCA-III score average from previous years. However, limited and focused linking procedures or prediction analyses may still serve useful purposes.

7.9. 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 of frequency distributions and summary statistics for scale scores by grade and subject under the sections entitled "Frequency Distribution Reports."

Chapter 8: 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 MCAs and MTAS/Alt MCA. 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 602, 2023-2024 Administration*. The equating and linking procedures for WIDA Alternate ACCESS are available online in the following report on the WIDA website (wida.wisc.edu/resources): *Annual Technical Report for WIDA Alternate ACCESS English Language Proficiency Test Series 602, 2023-2024 Administration*.

8.1. 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 within and 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 within and across test administrations.

Historically, a “two-stage with pre- and post-equating” design was used to maintain comparable difficulty across MCA administrations and with the large-scale paper form administrations of 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 MCAs and MTAS/Alt MCA. 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.

Although MDE has transitioned to a pre-equating design, a post-equating design was implemented for the Science MCA-IV and Alt MCA assessments in 2025. Spring 2025 is the first operational administration of the new science assessments. For their initial operational year, the base scale for each grade-level science assessment was reset, or updated, using a post-equating design. In a post-equating design, all items on a test form are placed on the base scale after an operational administration, and the banked item parameters are used for scoring. The pre-equating design for MCA-III and MTAS and the post-equating design for Science MCA-IV and Alt MCA are fully described in the sections that follow.

8.2. MCA 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 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, 2013 for reading, 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., performance level and scale score) that can be validly compared to previous years.

For the new Science MCA-IV and Alt MCA tests, the intent of post-equating is to produce a test with a new base scale based on the most recent student performance data. This approach enables the base scale to be developed with consideration of the student performance data from the current operational test administration. This takes into account modifications in test content, changes in methods of administration, and student motivation. The post-equating process calibrates all new operational items and places them on a common scale. Once the base scale is set using post-equating for the first year of new series of assessment, pre-equating is then used to calibrate all new field test items to the base scale and to produce test forms for future administrations. In this way, each subsequent science assessment is constructed from a pool of items for which parameters have been equated to the base scale thus allowing scores on a new test form to be comparable to scores on the previous test form. The base scales for Science MCA-IV and Alt MCA were established in 2025 using post-equating.

8.2.1. Operational Test Construction and Review

8.2.1.1. Fixed-Form Assessments

Test construction for the new Science MCA-IV fixed-form assessments began by selecting the operational items for the initial operational administration. To support test construction, MCA-IV field test items embedded in the MCA-III test were equated to the MCA-III base scale prior to the first year of MCA-IV. Placing the new field test items on this temporary scale enabled evaluation of the new science tests' psychometric properties using the available field test data. Using the field test items available in the item pool, psychometricians and content specialists from Minnesota's testing service provider and MDE constructed new operational forms by selecting items that meet the content specifications of the subject tested and targeted psychometric properties. Psychometric properties targeted included test difficulty, precision, and reliability. The test 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. Because the new test forms for the initial operational year were on the temporary MCA-III scale, the MCA-III cut scores and psychometric curves from past years were used as reference targets during test construction. Minnesota's testing service provider and MDE staff were not required to exactly match these reference targets. Reasonable flexibility was allowed during the evaluation process, acknowledging that a new assessment was being developed in alignment with new science content standards. Therefore, achieving complete alignment with the reference targets between the new test and its earlier version was not established as the main goal.

Psychometricians and content staff then reviewed the newly constructed test forms to ensure that specifications and difficulty levels had been maintained. Although every item on the test had 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 reexamined 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—was also evaluated, and items were further examined for their statistical quality, range of difficulties, and spread of information. Staff members also reviewed forms to ensure a wide variety of content and situations were 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 were designed to verify that keyed answer choices were correct and that the order of answer choices on the test form varied appropriately.

If any of these procedures uncovered an unsatisfactory item, it was replaced with a replacement from the item bank and the review process began again. This process for reviewing each newly constructed test form helped ensure each test was of the highest possible quality.

8.2.1.2. 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. Instead of form-to-form equating used in fixed-form tests, adaptive tests rely on IRT item calibration and linking to place all items and scores on a common ability scale. This maintains score comparability across administrations even though test difficulty varies by examinee.

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.

Mathematics and Reading 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.

8.2.2. Field Test Items

When a newly developed item has survived committee reviews (passage review for reading, phenomena 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 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-IV administration, there may be 17 different forms containing the same operational test items; however, each form would also contain several unique field test phenomena 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 the 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, majority field test items are placed in similar positions in the middle on each test form. For the MCA paper accommodated forms, there is one operational form that does not contain field test items.

8.2.3. 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 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 assessments 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. Additionally, 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 and other online assessments, students must respond to at least 90% of the items on the test to be classified as "attempted."

In some test administrations, such as the 2025 Science MCA-IV administration, sampling is used to obtain a representative collection of student responses for field test scoring. Sampling criteria are used to select students for each field test CR item so that each item for a specific grade level has the same demographic characteristics of the state for the specific grade level. The sampling criteria include student group frequency distributions and percentages for gender, ethnicity, students with disabilities, and English language learners. Once testing is complete, a statewide data file is provided that includes demographic subgroup information for all students in each science grade level. After exclusion rules are applied to the data file, demographic subgroup frequency distributions and percentages are then generated for each field test CR item at each grade level and compared to the sampling criteria to evaluate how close the sample demographic subgroup frequences and percentages for each item align with the sampling criteria. The sampling process is conducted iteratively to achieve the best possible alignment with the established sampling criteria.

8.2.4. Data 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 calibrations. 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 two times during the testing window for operational items (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(s) are the correct answers.

8.2.5. Operational Item Equating Procedures

The Mathematics and Reading MCA-III operational tests are pre-equated, meaning that the existing item bank parameters are used for scoring and reporting results. However, post-equating was conducted in 2025 for the Science MCA-IV operational items to obtain new item bank parameters to reset, or update, the base scale for each grade-level science assessment. For Science MCA-IV, the commercial software IRTPRO was used for operational item calibrations. For science, FIB items are calibrated as two-parameter items, 3-point CR items are calibrated using the GPC model, and all other items are calibrated as three-parameter items. These models are described in Chapter 7: Scaling.

8.2.6. Operational Calibration

All operational items on each grade-level science form were calibrated using a post-equating design to place them on an operational base scale. The post-equating was conducted using spring 2025 student performance data with exclusion rules applied. To establish the new base scale for each science assessment, the student response data for all operational items across the new test forms were freely calibrated simultaneously (i.e., a concurrent calibration). In this way, all student response data associated with an item, regardless of the form on which the item was administered, is used in the estimation of the item parameters. Because all items are calibrated together, it ensures they are placed on the same underlying scale. The operational item parameters which are now on the base scale are then used to pre-equate the field test items to the same common scale and to create raw score-to-scale-score lookup tables for the online tests and paper accommodated tests for reporting purposes.

8.2.7. Field Test Item Equating Procedures

For all MCA assessments, the commercial software IRTPRO is used for field test 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, FIB items are calibrated as two-parameter items, 3-point CR items are calibrated using the GPC model, and all other items are calibrated as three-parameter items. These models are described in Chapter 7: Scaling.

8.2.8. 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 for field test calibration, a drift analysis is conducted to flag operational 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 emphasis 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 for the items that have more than 1000 responses 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 and the Reading MCA-III grades 3–8 and 10 online assessments, operational 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 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. Drift analyses were not conducted for the Science MCA-IV assessments in spring 2025; however, these analyses will continue with future operational administrations.

8.2.9. Field Test Calibration

Field test calibration is the procedure used to obtain item parameter estimates for newly administered field-test items by placing them onto the same measurement scale as the established operational item bank. Through this process, the field-test item parameters become comparable to the existing operational parameters and are suitable for use in future test construction.

For the computer-adaptive Mathematics and Reading MCA-III assessments, 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 the fixed-form online Science MCA assessments, 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 fixed operational-item parameter approach being applied starting in 2025. The Science MCA-IV field test items are calibrated using IRTPRO by fixing the non-CR operational item parameters. By fixing these operational item parameters, the calibrated field test items are placed on the same base scale as the operational items. The calibration of the science field test items is conducted in two rounds due to data availability for the procedures. In the first round, the machine scored field test items (i.e., MC, TE, and FIB items) are calibrated to the base scale. This round happens after the administration is over but before field test CR scoring is complete. In the second round, only the field test CR items are calibrated to the base scale once human-scoring is complete.

8.3. MTAS Equating

The commercial software package WINSTEPS (Linacre, 2023) is used for the Minnesota Test of Academic Skills (MTAS-III) performance tasks. As described in Chapter 7: Scaling, the IRT model used for calibration is the RPC 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 is 2014 for grade 11 mathematics, 2011 for grades 3–8 mathematics, and 2013 for grades 3–8 and 10 reading.

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.

8.4. Alt MCA Equating

Test construction for the new Science Alt MCA assessment began by selecting the operational items for the initial operational administration. To support test construction, Science Alt MCA field test items embedded in the MTAS-III test were equated to the MTAS-III base scale. Placing the new field test items on the temporary scale enabled evaluation of the new science tests' psychometric properties using the available field test data. Using the field test items available in the item pool, psychometricians and content specialists from Minnesota’s testing service provider and MDE constructed new operational forms by selecting items that meet the content specifications of the subject tested and targeted psychometric properties.

After the initial test builds were complete, Minnesota’s testing service provider and MDE content staff reviewed each grade-level test form, verified the match to the blueprint and other constraints, checked for content related issues such as clueing, and made any appropriate changes in the items selected for use. The summary of revisions and associated explications were passed to psychometrics for further review. Because the new test forms for the initial operational year were on the temporary MTAS-III scale, the MTAS-III cut scores and psychometric curves were used as reference targets during test construction. Reasonable flexibility was allowed during the evaluation process, acknowledging that a new assessment was being developed in alignment with new science content standards This iterative cycle between Minnesota's testing service provider and MDE staff continued until consensus was reached on a final operational test form. After approval of the operational forms, MDE content staff identified the locations on the test form where the FT items would be embedded. The composition of the operational forms were then finalized at the end of the test construction window between Minnesota's testing service provider and MDE content and psychometrics staff.

After the operational administration was complete, a statewide data file was produced and edited for exclusions. A statistical review of all operational and field test items was conducted using this data file before beginning the IRT calibrations. This key check process verified the correct answer key for all the MC items. After completing the data quality checks to ensure the scoring was correct and there were no printing or processing problems, the equating process was conducted for the operational and field test items.

As described in Chapter 7: Scaling, the Rasch IRT model was used for calibrations. The commercial software package WINSTEPS (Linacre, 2023) was used to equate the operational items and field test items using the spring 2025 student performance data with exclusion rules applied. First, all operational items on each grade-level science test were post-equated by freely calibrating the items to place them on the base scale. Next, the operational item parameters which are now on the base scale were then used to pre-equate the field test items on each grade-level test to the operational base scale. The field test items were calibrated and placed on the base scale by fixing the operational item parameters to their freely calibrated estimates. Once the operational item parameters were estimated, they were also used to create raw score-to-scale-score lookup tables for reporting purposes.

8.5. Item Pool Maintenance

After equating is complete, the next step is to update the item pool with the new statistical information. Item statistics and parameter estimates for the operational science items are added to the item pool database, as 2025 is the first year the MCA-IV Science items are equated. Because pre-equating has been used for Mathematics and Reading MCA-III, new parameter estimates are obtained only for the field test items and the eligible drifted operational items which were recalibrated.

8.6. 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).

8.6.1. 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.

8.6.2. 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 9: 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 5.4: Cautions for Score Use and Section 7.8: Scale Score Interpretations and Limitations.

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 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.

9.1. 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 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.

Committees 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 2019 version of the science standards). These committees 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.

A new set of alignment studies were conducted in August 2025 for the new Science MCA-IV and Alt MCA assessments. The alignment studies examined whether the assessment items represent the content and multidimensional nature of the state science standards across grades 5, 8, and high school. The studies applied a multidimensional framework reflecting the three interconnected science dimensions of the 2019 Minnesota Academic Standards: Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs). The external independent service provider convened one panel for each grade level composed of experienced Minnesota science educators and special education teachers to evaluate and rate all unique items from operational test forms. Overall, both alignment studies demonstrated that the Science MCA-IV and Alt MCA assessments have a strong alignment with the 2019 Minnesota Academic Standards in Science. Most criteria were fully met across grade levels, with some partially met designations. Recommendations were also provided to strengthen the alignment of the science assessments and the Minnesota Academic Standards in Science. A more detailed description of the alignment study activities for science can be found in the alignment study final reports[11].

[11] The testing vendor and MDE content specialists reviewed the items identified as having weak alignment in the final alignment study report and took appropriate actions.

Following the first administration of new assessments aligning with revised academic standards, MDE also develops PLDs, which provide a description of typical grade-level performance for the performance levels. The PLDs 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 PLDs and the knowledge and skills required to meet proficiency according to the academic standards. To ensure the PLDs have high validity, the PLDs 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 performance levels reported in Minnesota. The PLDs are used in standard setting and define the grade-level student performance in each level 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 10: Reliability. The *Yearbook* has tables reporting the SEM and coefficient alpha reliabilities for tests that report raw scores and tables reporting CSEM and marginal reliabilities for MCA adaptive tests, broken down by demographic groups. Additionally, the *Yearbook* also reports tables illustrating the inter-rater agreement for the Science MCA-IV CR items and comparing the CSEM of the MCA adaptive tests with previous administrations. These measures show the Minnesota statewide assessments to be reliable.

9.2. 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 item targets for each of three DOK levels. 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 Minnesota's statewide assessments 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 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 levels of DOK.

9.3. 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 subscores and overall mathematics scale.

One type of evidence for internal structure that is provided for the fixed-form 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 and MTAS/Alt MCA 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 MCAs and MTAS/Alt MCA assessments 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*, the MCAs and MTAS/Alt MCA 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. As unidimensionality analyses require the presence of a full data matrices, they are conducted only for fixed form assessments and not for the adaptive tests, where the data matrices contain systematically missing data.

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 tests generally show the first eigenvalue accounting for 10% to 20% of the total variance, while the MTAS/Alt MCA tests generally show much higher values for percent of total variance accounted for (greater than 30%). Although the MCA 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 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 and MTAS/Alt MCA tests measure an essentially unidimensional composite.

In addition to the PCA, the unidimensional composite for the assessments can be investigated at the item level through the point-biserial correlation (also called item-total correlation). The content measured by each item on the test should have a strong relationship with the content measured by the other items. Apoint-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 assessments and the total raw score for the MTAS/Alt MCA assessments. The subscores for the MCA assessments are represented by the subscore theta, and for the MTAS/Alt MCA assessments, the subscores are the raw scores for the Access Points. 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 MTAS/Alt MCA 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* 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.

9.4. 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.

9.5. 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's alternate tests (MTAS-III and Alt MCA) 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 alternate 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 Mathematics and Reading MTAS-III 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 Alt MCA 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 and Alt MCA 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 and Alt MCA Scale Scores."

In additional support of the validity of the alternate assessments, the relationships among the MTAS/Alt MCA 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 Alt MCA are consistent with expectations. Results from these analyses are provided in the *Yearbook* under the section "Correlation of LCI Variables with MTAS/Alt MCA Scale Scores."

9.6. 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 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 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, the evidence is less compelling if the criterion measures are only indirectly related to the standards. The same can be said of the MTAS/Alt MCA. 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 Mathematics MCA-III 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 Reading MCA-III 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 Mathematics MCA-III scale scores and the nationally recognized "precollege" entrance examination; similar findings were observed for reading, where the correlation between the grade 8 Reading MCA-III 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.

9.7. Additional Validity Evidence

9.7.1. 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.

9.7.2. Scoring of MTAS/Alt MCA Items

The auditing of the MTAS/Alt MCA administrations supplies validity evidence for the scoring of the assessments. The auditing procedure is described in Chapter 10: Reliability and results of the audits are provided in the *Yearbook*.

9.7.3. 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. Item fit statistic is calculated for each item and presented at the data review along with other IRT statistics. Item fit and other IRT statistics are used in conjunction to determine if an item should become an operational item. Item fit information also informs the selection and evaluation of items used in the equating procedures. Justification for the scaling procedures used for Minnesota's statewide assessments is found in Chapter 7: 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 MCAs is the mode of administration. Prior to 2015, the MCA 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 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 grades 3–8 Mathematics MCA-III .

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 10: 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 9: Validity.

10.1. 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:

$$Observed\ Score = True\ Score + Error \tag{10.1}$$

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:

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2} \tag{10.2}$$

where σ_T^2 is the true score variance, $\sigma_O^2$is the variance of the observed score, and $\sigma_E^2$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 10.1 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.

10.2. 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 9: 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.

10.2.1. 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.

10.2.2. 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.

10.2.3. 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:

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} S_{Y_i}^2}{S_X^2}\right) \tag{10.3}$$

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 the MTAS/Alt MCA assessments, which use a fixed form and report raw scores. 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 9: 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. The MTAS/Alt MCA coefficient alpha statistics for each grade and subject can be found in the Yearbook under the section "Summary Statistics Reports." 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. The MCA-III Mathematics and Reading accommodated form assessments and the MCA-III Science assessment, similar to the MCA-III Mathematics and Reading online adaptive assessments, calculated marginal reliability in lieu of coefficient alpha. Marginal reliability is discussed later in this chapter.

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 Mathematics and Reading MCA-III; all Science MCA-IV fixed-form assessments; and all MTAS/Alt MCA 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 the MCAs and MTAS/Alt MCA. 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. For fixed-form assessments, item-total correlation is calculated between an item score and the total raw score based on operational items. For adaptive forms, the item-total correlation is calculated between an item score and student's scale score from the CAT form. 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.

10.2.4. IRT-Based Reliability

Instead of reporting coefficient alpha for the MCA 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:

$$\hat{\theta} = \theta + \epsilon, \qquad (10.4)$$

where $\hat{\theta}$ 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:

$$R_{\theta} = \frac{var(\theta)}{var(\hat{\theta})} = \frac{var(\hat{\theta}) - var(\epsilon)}{var(\hat{\theta})}. \qquad (10.5)$$

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

$$Marginal\ Reliability = \frac{\sigma^2_{theta} - \overline{\mathrm{SE}_{\mathrm{theta}_i}{}^2}}{\sigma^2_{theta}}, \tag{10.6}$$

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 online strand and substrand scores, 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):

$$\text{EAP}\ Marginal\ Reliability = 1 - \overline{PSD^2}, \tag{10.7}$$

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

$$\text{PSD} = \text{Var}(\theta|\mathbf{u}) = \frac{\sum_{k=1}^{q}(\text{X}_k - \theta)^2\, L(\text{X}_k)W(\text{X}_k)}{L(\text{X}_k)W(\text{X}_k)} \tag{10.8}$$

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.

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 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 online accommodated forms, which are derived from the same base form that are used with different types of 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 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.

10.3. 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:

$$SEM = s_x\sqrt{1 - \rho_{xx}}, \tag{10.9}$$

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.

10.3.1. 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:

$$SEM = 10.00\sqrt{(1.00 - 0.84)} = 4.00 \tag{10.10}$$

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 total raw score and the raw scores for the subscores is reported for the Mathematics, Reading, and Science MTAS/Alt MCA. The SEM for the total scale score and the theta scores for the subscores is reported for Science MCA. The SEM values are listed in the *Yearbook* under the section "Summary Statistics Reports" for each respective grade and subject. The overall SEM for each MCA 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. SEM is not reported for assessments that use pattern scoring. It is often more realistic to use the CSEM instead of SEM for the assessments that apply pattern scoring because the CSEM can be made more specific to an individua's score.

10.3.2. 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 all MN assessments at the test level and for MCA-IV Science at the reporting category level, $CSEM(\theta_i)$ i calculated as the inverse of the square root of the test information function

$$CSEM(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{10.15}$$

where $I(\theta_i)$ is the test information, sum of information from all items on the test:

$$I(\theta_i) = \sum_{j=1}^{N} I_{ij}(\theta)$$

where N is the number of items on the assessment and $I(\theta_{ij})$ is the information for item j. The item information function for the 3PL IRT model is calculated as:

$$I_{ij}(\theta) = \frac{D^2 a_j^2 Q_j(\theta)(P_j(\theta) - c_j)^2}{P_j(\theta)(1 - c_j)^2} \tag{10.16}$$

where *a* is the discrimination parameter, *c* is the guessing parameter, and *D* is the scaling constant. This formula is also applicable to the 2PL IRT model, in which the *c*-parameter is zero, and to Rasch model, in which both *D* and a have a value of 1. The item information function for a CR item, which uses the GPCM model (Muraki, 1993), is computed as:

$$I_j(\theta) = 2.89 \cdot a_j^2 \cdot \sum_{c=1}^{m_j} [T_c - \overline{T}_l(\theta)]^2 \, P_{jc}(\theta)$$

where a_j is the discrimination parameter for item j, $P_{jc}(\theta)$ is the category response probability for a GPCM item, T_c is the linear integer scoring function and takes the value of 1, 2, 3, …m+1 for each of the score categories 0, 1, 2, …m, and $\overline{T}_l(\theta) = \sum_{c=1}^{m_j} T_c \, P_{jc}(\theta)$.

The CSEM that results from this calculation is in the theta metric, and thus should be multiplied by their respective scaling constant to place it in the scale score metric as shown in the following equation:

$$CSEM(Scale) = a \cdot CSEM(\theta)$$

where a is the scaling constant of each assessment.

For MCA-III Reading and Math, the strand level CSEM in the theta scale is calculated as follows: ,

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

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.

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 the MCAs and are reported by each raw score and scale score for MTAS and by each scale score for Alt MCA. 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 Mathematics and Reading MCA-III, 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 Mathematics and Reading MCA-III assessments to CAT resulted in improved measurement precision.

10.3.3. 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.

10.3.4. 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:

$$s_e = \frac{s_x}{\sqrt{N}}, \tag{10.17}$$

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:

$$s_e = \frac{10}{\sqrt{20}} = 2.2 \tag{10.18}$$

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).

10.4. Auditing of MTAS/Alt MCA Administrations

Reliability evidence primarily focuses on the amount of error involved in measurement. In an assessment such as the MTAS/Alt MCA an additional source of measurement error can come from the Test Administrator. To minimize the measurement error in the MTAS/Alt MCA, Minnesota test administrators strictly adhere to the procedures for administering and scoring the assessment. Because many students taking the MTAS/Alt MCA have unique communication styles that require significant familiarity with the student to understand their intended communication, the performance tasks for MTAS and items for Alt MCA 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 for MTAS. Minnesota conducts rater audits on test administrators for the MTAS. The MDE recruited Minnesota educators and administrators (current or retired) to serve as scoring auditors for MTAS and administration procedures auditors for Alt MCA. These auditors were trained in the administration and scoring of the MTAS/Alt MCA 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. For MTAS, 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."

10.5. 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 performance levels. For example, some students may have a true performance 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 performance level. As discussed in Section 10.3: 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 performance levels can be imperfect, especially for borderline students whose true scores lie close to performance level cut scores. For the MCAs and MTAS/Alt MCA, the performance levels are *Does Not Meet/Beginning*, *Partially Meets/Intermediate*, *Meets/Meets*, and *Exceeds/Advanced*. True level of achievement is based on the student's true score and cannot be observed; therefore, classification accuracy cannot be directly determined. However, it is possible 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 performance level (bounded by cut scores c and d) who are classified into a performance level category (bounded by cut scores a and b) can be obtained by:

$$P(Level_k) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{SE(\theta)}\right) - \phi\left(\frac{a-\theta}{SE(\theta)}\right) \right) f(\theta), \tag{10.19}$$

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 performance levels) is used to represent the true theta score (and performance level) distribution. Based on that distribution, Equation 10.19 is used to estimate the proportion of students at each performance level who are expected to be assigned to each possible performance level. To compute classification consistency, the percentages are computed for all cells of a true-versus-expected performance 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 10.1 presents an example classification table for an MCA assessment. The columns represent the true student performance level, and the rows represent the test-based performance level assignments expected to be observed, given Equation 10.19. 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/Beginning*. Similarly, 1.3% of students were incorrectly classified as *Does Not Meet/Beginning*, when their true score indicates *Partially Meets/Intermediate*.

Table 10.1. Example Classification Table

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

Note. D/B = *Does Not Meet/Beginning*, P/I = *Partially Meets/Intermediate*, M/M = *Meets/Meets*, E/A = *Exceeds/Advanced*

It is useful to consider decision consistency based on a dichotomous classification of *Does Not Meet/Beginning* or *Partially Meets/Intermediate* versus *Meets/Meets* or *Exceeds/Advanced* because Minnesota uses *Meets/Meets* 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/Beginning* with *Partially*

Meets/Intermediate and combining *Meets/Meets* with *Exceeds/Advanced*. For the example table above, this results in a classification accuracy of 92.2%. The percentage of students incorrectly classified as *Partially Meets/Intermediate* or lower, when their true score indicates *Meets/Meets* or above, is 2.3%.

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

10.6. Inter-Rater Agreement

To evaluate scoring consistency for the Science MCA CR operational items, inter-rater agreement was examined to determine the extent to which students would obtain the same CR score if scored by IEA and an independent human rater. Additional information regarding inter-rater reliability for the Science MCA CR items can be found in Chapter 4: Scoring.

Chapter 11: Quality Control Procedures

The Minnesota assessment program and its associated data play an important role in the state accountability system as well as in many local evaluation plans. Therefore, it is vital that quality control procedures are implemented to ensure the accuracy of student-, school-, and district-level data and reports. Minnesota's testing 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.

11.1. Quality Control for Test Construction

The Minnesota assessment program includes both fixed-form and computer adaptive assessments. For fixed-form assessments, the test construction process described in Chapter 2: Test Development is followed. In this process, items are selected and placed on a particular test form to provide a strictly parallel form across years in terms of content and statistics. Item and form statistical characteristics from the baseline tests 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 test form. Similarly, the baseline raw score-to-scale-score tables are used as the target tables that the test form (under construction) should match as close as possible at the performance level cut scores. This form is provided to Minnesota's testing service provider for form construction and typesetting.

For computer adaptive assessments, test construction involves identifying the item pool, establishing CAT settings, evaluating simulation results, and selecting the final operational CAT configuration. The content team of Minnesota's testing service provider collaborates with MDE's content team to determine all eligible items within the item banking system. The provider's psychometric team then independently extracts the item pool to confirm that all eligible items are included, and all non-eligible items are excluded. This team also develops multiple CAT settings, conducts simulations, and evaluates them using indices related to blueprint alignment, measurement accuracy, and item pool usage. Once the optimal settings are selected, a configuration file is generated for each assessment and sent to the provider's test administration team. A separate internal team reviews this file at multiple stages to ensure accuracy and current status. The file is also provided to MDE's psychometric team and to Minnesota's quality control vendor for independent verification before operational administration begins.

11.2. 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.

11.3. 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.

11.4. 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 MCA and Alt MCA 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 MCA 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.
- Minnesota's testing service provider and MDE perform a data validation analysis for Alt MCA. The analysis involves reviewing item responses to confirm the accuracy of the VS-ES score code and to identify any students who qualifies for VS-ES but did not receive the appropriate code.
- 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, remove, or recalibrate 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 measurement scale.

Glossary of Terms

Achievement Level Descriptor (ALD). A description of an achievement level by grade and subject that provides descriptive information of what typical students at the 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. The classification of a student's competency based on their test performance as reflected by their overall test results. Achievement levels group students into categories (e.g., *Beginning*, *Intermediate*, *Meets*, and *Advanced*) based on the cut scores that divide the score scale for a grade and subject.

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 Mathematics and Reading 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. Mathematics and Reading MCA-III are the only assessments in Minnesota's statewide assessment system administered as CATs.

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.

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, student responses from MCA paper accommodated test materials entered into data-entry forms in TestNav, and MTAS scores entered into PearsonAccess Next.. On-Demand Reports are available in PearsonAccess Next. Note: For 2024-25, On-Demand Reports were only available for Reading and Mathematics MCA-III and MTAS-III.

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: A score that takes the student's item response pattern (Reading and Mathematics MCA-III) or raw score (Science MCA-IV) and adjusts it for possible differences in test difficulty from one year to the next. For MTAS-III and Alt MCA: 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 and alternate assessments are based on the most recent academic content standards in Mathematics, Reading, and Science. They 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*.

Table of Contents

References
Appendix A: Benchmark Report Calculations Resources

References

Abedi, J. (2001). *Assessment and accommodations for English language learners: Issues and recommendations* (CRESST Policy Brief No. 4). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Abedi, J., Zhang, Y., Rowe, S. E., & Lee, H. (2020). Examining effectiveness and validity of accommodations for English language learners in mathematics: An evidence-based computer accommodation decision system. *Educational Measurement: Issues and Practice, 39*(4), 41-52.

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.

Almgren Bäck, G., Mossige, M., Bundgaard Svendsen, H., Rønneberg, V., Selenius, H., Berg Gøttsche, N., Dolmer, G., Fälth, L., Nilsson, S., & Svensson, I. (2024). Speech-to-text intervention to support text production among students with writing difficulties: a single-case study in nordic countries. *Disability and rehabilitation. Assistive technology, 19*(8), 3110–3129.

Alrawashdeh, G. S., Fyffe, S., Azevedo, R. F., & Castillo, N. M. (2024). Exploring the impact of personalized and adaptive learning technologies on reading literacy: A global meta-analysis. Educational Research Review, *42*, 100587.

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.

Anakwe, B. (2008). Comparison of student performance in paper-based versus computer-based testing. *Journal of Education for Business, 84*(1), 13-17.

Arcon, N., Klein, P. D., & Dombroski, J. D. (2017). Effects of dictation, speech to text, and handwriting on the written composition of elementary school English language learners. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 33*(6), 533–548.

Ashbaugh, C. A. (2018). The Enhancing Reading Comprehension Performance using the Strategy of Highlighting *Electronic Text* (Master's thesis, Stephen F. Austin State University).

Aspiranti, K. B., & Hulac, D. M. (2022). Using fidget spinners to improve on-task classroom behavior for students with ADHD. *Behavior Analysis in Practice, 15*(2), 454–465.

Barton, K. E. W. (2001). *Stability of constructs across groups of students with different disabilities on a reading assessment under standard and accommodated administrations* (Publication No. 3036177) [Doctoral dissertation, University of South Carolina]. ProQuest Dissertations and Theses Global.

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.

Bernard, J. L., & Witmer, S. E. (2025). Examining how students with ADHD use an extended time accommodation on a low-stakes math assessment. *Journal of Attention Disorders, 29*(10), 803–816.

Bertachini AL, Pupo AC, Morettin M, Martinez MA, Bevilacqua MC, Moret AL, Balen SA, Jacob RT. Frequency modulation system and speech perception in the classroom: A systematic literature review. *CoDAS*. 2015; *27*(3): 292–300.

Bertoni, S., Franceschini, S., Ronconi, L., Gori, S., & Facoetti, A. (2019). Is excessive visual crowding causally linked to developmental dyslexia?. Neuropsychologia, 130, 107-117.

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.

Bone, E. (2020). *Using Manipulatives to Support Secondary Students with High-Incidence Disabilities in Algebra*. Michigan State University.

Bone, E., Bouck, E., & Witmer, S. (2021). Evidence-Based Systematic Review of Literature on Algebra Instruction and Interventions for Students With Learning Disabilities. *Learning Disabilities: A Contemporary Journal, 19*(1), 1-22.

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.

Bouck, E. C., Flanagan, S., & Joshi, G. S. (2011). Speaking math—A voice input, speech output calculator for students with visual impairments. *Journal of Special Education Technology, 26*(4), 1–14.

Bouck, E., & Park, J. (2018). A systematic review of the literature on mathematics manipulatives to support students with disabilities. *Education and Treatment of Children, 41*(1), 65–106.

Bouck, E., Park, J., Shurr, J., Bassette, L., & Whorley, A. (2018). Using the virtual-representational-abstract approach to support students with intellectual disability in mathematics. *Focus on Autism and Other Developmental Disabilities, 33(*4), 239–248.

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.

Brown, T. , Reichel, P. and Quinlan, D. (2011) Extended time improves reading comprehension test scores for adolescents with ADHD. *Open Journal of Psychiatry, 1,* 79-87.

Buehler, K. L. (2001). *Standardized group achievement tests and the accommodation of additional time* (Publication No. 3050241) [Doctoral dissertation, Indiana State University]. ProQuest Dissertations and Theses Global.

Burch, M. A. (2002). *Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities* (Publication No. 3047429) [Doctoral dissertation, Vanderbilt University]. ProQuest Dissertations and Theses Global.

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.

Buzick, H., & Stone, E. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement: Issues and Practice, 33*(3), 17–30.

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.

Candrlic, S., Katic, M.A., & Dlab, M.H. (2014). Online vs. Paper-based testing: A comparison of test results. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* 657-662.

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).

Cawthon, S. W., Winton, S. M., Garberoglio, C. L., & Gobble, M. E. (2011). The effects of American Sign Language as an assessment accommodation for students who are deaf or hard of hearing. *The Journal of Deaf Studies and Deaf Education, 16*(2), 198–211.

Chang, W., & Shieh, R.-S. (2013). Exploration of the value of using a formula sheet for physics examinations. *European Journal of Physics, 34*(6), 1411–1423.

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.

Cohen, A. S., Gregg, N., & Deng, S. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*(4), 225-233.

Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education, 30*(4), 259-272.

Cohen, D. J., Ballman, A., Rijmen, F., & Cohen, J. (2020). On the reliable identification and effectiveness of computer-based, pop-up glossaries in large-scale assessments. *Applied Measurement in Education, 33*(4), 378-389.

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.

Cook, C. R., Bradley-Johnson, S., & Johnson, C. M. (2014). Effects of white noise on off-task behavior and academic responding for children with ADHD. *Journal of Applied Behavior Analysis, 47*(1), 160-164.

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.

Cox, M. E., Herner, J. G., Demczyk, M. J., & Nieberding, J. J. (2006). Provision of testing accommodations for students with disabilities on statewide assessments. *Remedial and Special Education, 27*(6), 346–354.

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.

Cullen, J. M., Richards, S. B., & Frank, C. L. (2008). Using software to enhance the writing skills of students with special needs. *Journal of Special Education Technology, 23*(2), 33-44.

Dalton, B. H., & Behm, D. G. (2008). Effects of noise and music on human and task performance: A systematic review. *Occupational Ergonomics, 7*(3), 143-152.

Dawson, K., Antonenko, P., & Lane, H. (2019). Assistive technologies to support students with dyslexia. *Assistive Technology, 31*(2), 65-74.

De Backer, F., Vantieghem, W., Slembrouck, S., & Van Avermaet, P. (2024). Science achievement of multilingual pupils: A study on the effectiveness of a read-aloud assessment accommodation. *Education Sciences, 14*(5), 494.

de la Mora Velasco, E., & Hirumi, A. (2020). The effects of background music on learning: A systematic review of literature to guide future research and practice. *Educational Technology Research and Development, 68*(6), 2817–2837.

Denton, T. F., & Meindl, J. N. (2016). The effect of colored overlays on reading fluency in individuals with dyslexia. *Behavior Analysis in Practice, 9*(3), 191–198.

Dhitareka, P. H., Firman, H., & Rusyati, L. (2018). The comparison between science virtual and paper based test in measuring grade 7 students' critical thinking. Journal of Physics: Conference Series, 1013(1), 012070.

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.

Evmenova, A. S., Graff, H. J., Jerome, M. K., & Behrmann, M. M. (2010). Word prediction programs with phonetic spelling support: Performance comparisons and impact on journal writing for students with writing difficulties. *Learning Disabilities Research & Practice, 25*(4), 170–182.

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.

Finizio, N. J., II. (2008). *The relationship between instructional and assessment accommodations on student IEPs in a single urban school district* (Publication No. 3313763) [Doctoral dissertation, University of Massachusetts Boston]. ProQuest Dissertations and Theses Global.

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.

Fox, D., Karic, B., & Ash, A. (2015). *The effect of white noise on psychological stress while performing cognitive tasks and its correlation to performance.* University of Wisconsin-Madison Minds Repository.

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.

Garrett, J. T., Heller, K. W., Fowler, L. P., Alberto, P. A., Fredrick, L. D., & O'Rourke, C. M. (2011). Using speech recognition software to increase writing fluency for individuals with physical disabilities. *Journal of Special Education Technology, 26*(1), 25–42.

Gilmore, C., Keeble, S., Richardson, K., & Cragg, L. (2015). The role of cognitive inhibition in different components of arithmetic. *ZDM – Mathematics Education, 47*(5), 773-782.

Goldstone, L., Lazarus, S. S., Olson, R., Hinkle, A. R., Ressa, V., & Rogers, C. M. (2021). *Student reads aloud to self: Research (NCEO Accommodations Toolkit #9a).* National Center on Educational Outcomes.

Gómezescobar, A., Guerrero, S., & Fernández-Cézar, R. (2020). How long is it? Difficulties with conventional ruler use in children aged 5 to 8. *Early Childhood Education Journal, 48*(6), 693–701.

Gómezescobar, A. (2025). Ruler by steps (RbySteps): Exploring length measurement with young children through a sequential ruler construction. *Early Childhood Education Journal, 54*.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Erlbaum.

Granquist, C., Wu, Y. H., Gage, R., Crossland, M. D., & Legge, G. E. (2018). How people with low vision achieve magnification in digital reading. *Optometry and Vision Science, 95*(9), 711-719.

Graziano, P. A., Garcia, A. M., & Landis, T. D. (2020). To fidget or not to fidget, that is the question: A systematic classroom evaluation of fidget spinners among young children with ADHD. *Journal of Attention Disorders, 24*(1), 163-171.

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.

Griffiths, P. G., Taylor, R. H., Henderson, L. M., & Barrett, B. T. (2016). The effect of coloured overlays and lenses on reading: a systematic review of the literature. *Ophthalmic and Physiological Optics, 36*(5), 519-544.

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.

Hallam, S., Price, J., & Katsarou, G. (2002). The effects of background music on primary school pupils' task performance. *Educational Studies, 28*(2), 111-122.

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.

Hahn, C., Goldstein, D., Hasher, L., Wiprzycka, U. N., & Zelazo, P. D. (2012). Circadian rhythms in executive function during the transition to adolescence: The effect of synchrony between chronotype and time of day. *Developmental Science, 15*(3), 408-416

Heeney, M. F. (2007). *Classroom sound field amplification, listening and learning* [Doctoral thesis, University of Newcastle]. The University of Newcastle Research Online.

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, stress, and performance: How students respond physiologically to high-stakes testing. *Education Finance and Policy, 16*(2), 183–208.

Helps, S. K., Bamford, S., Sonuga-Barke, E. J. S., & Söderlund, G. B. W. (2014). Different effects of adding white noise on cognitive performance of sub-, normal and super-attentive school children. *PLoS ONE, 9*(11), e112768.

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.

Henderson, L. M., Tsogka, N., & Snowling, M. J. (2013). Questioning the benefits that coloured overlays can have for reading in students with and without dyslexia. *Journal of Research in Special Educational Needs, 13*(1), 57-65.

Hokken MJ, Krabbendam E, van der Zee YJ, & Kooiker MJ (2023). Visual selective attention and visual search performance in children with CVI, ADHD, and Dyslexia: a scoping review. *Child Neuropsychology*, *29*(3), 357–390.

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.

Holmes, A., & Silvestri, R. (2019). Extra time or unused time? What data from a college testing center tells us about 50% extra time as an accommodation for students with learning disabilities. *Psychological Injury and Law, 12*(1), 7-16.

Hulac, D. M., Briesch, A. M., Chafouleas, S. M., & McCoach, D. B. (2020). A multisite study of the effect of fidget spinners on academic performance. *Contemporary School Psychology, 25*(2), 209-216.

Huurneman, B., Cox, R. F., Vlaskamp, B. N., & Boonstra, F. N. (2014). Crowded visual search in children with normal vision and children with visual impairment. *Vision research*, 96, 65–74.

Ilhan, C., Sekeroglu, M. A., Doguizi, S., & Yilmazbas, P. (2019). Contrast sensitivity of patients with congenital color vision deficiency. *International Ophthalmology, 39*(3), 571-576.

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.

Jakovljević, T., Janković, M. M., Savić, A. M., Soldatović, I., Čolić, G., Jakulin, T. J., Papa, G., & Ković, V. (2021). The relation between physiological parameters and colour modifications in text background and overlay during reading in children with and without dyslexia. *Brain sciences, 11*(5), 539.

Janáky, M., Borbély, J., Benedek, G., Kocsis, B. P., & Braunitzer, G. (2014). Achromatic luminance contrast sensitivity in X-linked color-deficient observers: An addition to the debate. *Visual Neuroscience, 31*(1), 59-65.

Jiang, Y., & Cayton-Hodges, G. A. (2023). Investigating problem solving on calculator items in a large-scale digitally based assessment: A data mining approach. *Journal for Research in Mathematics Education, 54*(2), 118-142.

Jiang, Y., Cayton-Hodges, G. A., Oláh, L. N., & Minchuk, I. (2023). Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the National Assessment of Educational Progress (NAEP). *Computers & Education,* 193, 1–21.

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.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.

Kieffer, M. J., & Lesaux, N. K. (2012). Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade. *Reading and Writing, 25*(2), 347-373.

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.

Kriescher, S. L. (2020). *The effects of fidgets on attention and learning of college students* [Doctoral dissertation, University of Northern Colorado]. Scholarship & Creative Works @ Digital UNC. https://digscholarship.unco.edu/dissertations/666/

Kulawiak, P. R. (2021). Academic benefits of wearing noise-cancelling headphones during class for typically developing students and students with special needs: A scoping review. *Cogent Education, 8*(1).

Kuti, L. M. (2011). *Accommodations for English language learners with disabilities in federally-mandated statewide English language proficiency assessment* (Publication No. 3468551) [Doctoral dissertation, Virginia Commonwealth University]. ProQuest Dissertations and Theses Global.

Kwon, M. K., Congdon, E., Ping, R., & Levine, S. C. (2024). Overturning Children's Misconceptions about Ruler Measurement: The Power of Disconfirming Evidence. *Journal of Intelligence, 12*(7), 62.

Laderach, S. (2017). *Student-Created Test Sheets* [Honors project, Bowling Green State University]. OhioLINK Electronic Theses and Dissertations Center. https://scholarworks.bgsu.edu/honorsprojects/220/

Lai, S. A., & Berkeley, S. (2012). High-stakes test accommodations: Research and practice. *Learning Disability Quarterly, 35*(3), 158-169.

Laitusis, C. C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of fatigue effects from extended-time accommodations on the SAT Reasoning Test.* ETS Research Report No. 07-31.

Latham, K. (2018). Benefits of low vision aids to reading accessibility. *Vision Research, 153*, 31-38

Lee, I. X. C. (2011). *The application of speech recognition technology for remediating the writing difficulties of students with learning disabilities* (Publication No. 3501541) [Doctoral dissertation, University of Washington]. ProQuest Dissertations and Theses Global.

Lehmann, J. A. M., & Seufert, T. (2017). The influence of background music on learning in the light of different theoretical perspectives and the role of working memory capacity. *Frontiers in Psychology, 8,* 1902.

Leonard, S., Stroud, M. J., & Shaw, R. J. (2021). Highlighting and taking notes are equally ineffective when reading paper or eText. *Education and Information Technologies, 26*(4), 3811–3823

Levine, S., Hsieh, H., Southerton, E., & Silverman, R. (2023). How high school students used speech-to-text as a composition tool. *Computers and Composition, 68*, 102775.

Lewandowski, L. J., Lovett, B. J., Parolin, R., Gordon, M., & Codding, R. S. (2007). Extended time accommodations and the mathematics performance of students with and without ADHD. *Journal of Psychoeducational Assessment, 25*(1), 17-28.

Lewandowski, L., Martens, B. K., Clawson, A., & Reid, T. (2021). Effects of a Private Room Versus Group Setting on Math Test Performance of College Students with ADHD. *Journal of Behavioral Education, 30*(2), 247–259.

Lewandowski, L., Wood, W., & Lambert, T. (2015). Private room as a test accommodation. *Assessment & Evaluation in Higher Education, 40*(2), 279-285.

Lewis, B. P. (2024). Writing Strategies for Elementary Multilingual Writers: A Systematic Review. *Education Sciences, 14*(7), 759.

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.

Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement: Issues and Practice, 33*(3), 3–16.

Li, H., & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education, 25*(4), 327–346.

Linacre, J. M. (2006). *A user's guide to WINSTEPS*. MESA Press.

Lindstrom, J. H., & Gregg, N. (2007). The role of extended time on the SAT® for students with learning disabilities and/or attention-deficit/hyperactivity disorder. *Learning Disabilities Research & Practice, 22*(2), 85–95.

Linn, R. L. (1993). Linking results in distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Liu, K. K., Lazarus, S., Thurlow, M. L., Stewart, J., & Larson, E. (2020). *A summary of the research on test accommodations for English learners and English learners with disabilities: 2010–2018.* National Center on Educational Outcomes (NCEO), University of Minnesota.

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.

Lovett, B. J., Lewandowski, L. J., & Carter, L. (2019). Separate room testing accommodations for students with and without ADHD. *Journal of Psychoeducational Assessment, 37*(7), 852–862.

Lusk, K. E. (2012). The effects of various mounting systems of near magnification on reading performance and preference in school-age students with low vision. *British Journal of Visual Impairment, 30*(3), 168–181.

Lutsic, S., & Zhang, M. (2023). Driving assessment scores with effective calculator implementation. *International Journal of Social Learning (IJSL), 3*(3), 361–377.

MacArthur, C. A., & Cavalier, A. R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43–58.

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.

Mallory, C., & Keehn, B. (2021). Implications of Sensory Processing and Attentional Differences Associated With Autism in Academic Settings: An Integrative Review. *Frontiers in psychiatry, 12*, 695825.

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

Matre, M. E., & Cameron, D. L. (2024). A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education. *Disability and Rehabilitation: Assistive Technology, 19*(3), 1103–1116.

McCollum, D., Nation, S., & Gunn, S. (2014). The effects of a speech-to-text software application on written expression for students with various disabilities. *National Forum of Special Education Journal, 25*(1), 1–13.

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.

Mealings, K. (2022). A review of the effect of classroom sound-field amplification on children in primary school. *American Journal of Audiology*, 31(2), 470–486.

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.

Mulloy, A. M., Gevarter, C., Hopkins, M., Sutherland, K. S., & Ramdoss, S. T. (2014). Assistive technology for students with visual impairments and blindness. In G. E. Lancioni & N. N. Singh (Eds.), *Assistive technologies for people with diverse abilities* (pp. 113–156). Springer Assistive Technology for People with Diverse Abilities.

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.

Myers, B., & Kopriva, R. (2015). *Decision trees linking individual student need to large-scale accommodations for English learners: A white paper.* Wisconsin Center for Education Research.

Nigg, J. T., Bruton, A., Kozlowski, M. B., Johnstone, J. M., & Karalunas, S. L. (2024). Systematic review and meta-analysis: Do white noise or pink noise help with task performance in youth with attention-deficit/hyperactivity disorder or with elevated attention problems? *Journal of the American Academy of Child & Adolescent Psychiatry, 63*(8), 778–788.

Noakes, M. A., Schmitt, A. J., McCallum, E., & Schutte, K. (2019). Speech-to-text assistive technology for the written expression of students with traumatic brain injuries: A single case experimental study. *School Psychology, 34*(6), 656–664.

Ogata, N. G., Boer, E. R., Daga, F. B., Jammal, A. A., Stringham, J. M., & Medeiros, F. A. (2019). Visual crowding in glaucoma. *Investigative ophthalmology & visual science, 60*(2), 538-543.

One Feather, M. (2010). *Test accommodations and standardized assessment for students with learning disabilities who are second language learners* [Publication No. 3423834] [Doctoral dissertation, The University of Arizona]. ProQuest Dissertations and Theses Global.

Öz, H., & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies, 14*(1), 67-85.

Paquin, J. D., Miller, M. L., & Barron, J. (2020, November). *Efficacy of Learning with Course-provided Equation Reference Sheets in Engineering Education.* Paper presented at the 2020 Fall ASEE Mid-Atlantic Section Meeting, Virtual.

Patros, C. H., Alderson, R. M., Kasper, L. J., Tarle, S. J., Lea, S. E., & Hudec, K. L. (2016). Choice-impulsivity in children and adolescents with attention-deficit/hyperactivity disorder (ADHD): A meta-analytic review. *Clinical psychology review, 43*, 162-174.

Pengelley, J., Whipp, P. R., & Rovis-Hermann, N. (2023). A testing load: Investigating test mode effects on test score, cognitive load and scratch paper use with secondary school students. *Educational Psychology Review, 35*(3), Article 67.

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* (Publication No. 8108266) [Doctoral dissertation, University of South Florida]. ProQuest Dissertations and Theses Global.

Prior, S. M., Boomer, J., Boomer, M., & McCallum, R. S. (2011). The effects of reading modality on reading comprehension in elementary students. *Reading Psychology, 32*(3), 263-284.

Prisacari, A., & Danielson, J. (2017). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior, 77,* 1-10.

Randler, C., Díaz-Morales, J. F., Rahafar, A., & Vollmer, C. (2016). Morningness–Eveningness-Stability-Scale improved (MESSi). *PsycTESTS*.

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.

Razuk, M., Perrin-Fievez, F., Gerard, C. L., Peyre, H., Barela, J. A., & Bucci, M. P. (2018). Effect of colored filters on reading capabilities in dyslexic children. *Research in developmental disabilities, 83,* 1-7.

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.

Richardson, R. T., Drexler, T. L., & Delparte, D. M. (2014). Color and Contrast in E-Learning Design: A Review of the Literature and Recommendations for Instructional Designers and Web Developers. *Journal of Online Learning and Teaching, 10*(4), 654–670.

Rios, J. A., Ihlenfeldt, S. D., & Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multi-study systematic review and meta-analysis. Educational Measurement: Issues and Practice, 39(4), 65-75.

Roach, N. W., & Hogben, J. H. (2007). Impaired filtering of behaviourally irrelevant visual information in dyslexia. *Brain : a journal of neurology*, 130(Pt 3), 771–785.

Roberts, B. R. T., Hu, Z. S., Curtis, E., Bodner, G. E., McLean, D., & MacLeod, C. M. (2024). Reading text aloud benefits memory but not comprehension. *Memory & Cognition, 52*(1), 57-72.

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.

Robinson, M. F., Meisinger, E. B., & Joyner, R. E. (2019). The influence of oral versus silent reading on reading comprehension in students with reading disabilities. *Learning Disability Quarterly, 42*(2), 105–116.

Robson, D. A., Johnstone, S. J., Putwain, D. W., & Howard, S. (2023). Test anxiety in primary school children: A 20-year systematic review and meta-analysis. *Journal of School Psychology, 98,* 39-60.

Rogers, C. M., Lazarus, S. S., Ressa, V. A., Goldstone, L., & Fleming, K. (2022). *Color contrast: Research* (NCEO Accommodations Toolkit #25a). National Center on Educational Outcomes.

Rohrberger, A. (2011). The efficacy of fidget toys in a school setting for children with attention difficulties and hyperactivity [Master's thesis/Unpublished manuscript]. Ithaca College.

Roohr, K. C., & Sireci, S. G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment, 22*(1), 35–53.

Rosalez, E., Johnson, C., Bradley-Johnson, S., & Kanouse, S.K. (2020). Effects of White Noise on Off-Task Behavior and Sleep for Elementary-Age Students with ADHD. *Child & Family Behavior Therapy, 42*, 20 - 36.

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, R. (2014). *The impact of using calculators as an accommodation on the math achievement of students with learning disabilities* (Publication No. 3630205) [Doctoral dissertation, Grand Canyon University]. ProQuest Dissertations and Theses Global.

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.

Sanjurjo, A. (2009). *Essays in decision making under cognitive load.* UC San Diego. ProQuest ID: Sanjurjo_ucsd_0033D_10578. Merritt ID: ark:/20775/bb71685334. Retrieved from https://escholarship.org/uc/item/26n5q2kv

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.

Schimmel, N., & Ness, M. (2017). The effects of oral and silent reading on reading comprehension. *Reading Psychology, 38*(4), 390–416.

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.

Sjöblom, A. M., Eaton, E., & Stagg, S. D. (2016). The effects of letter spacing and coloured overlays on reading speed and accuracy in adult dyslexia. *British Journal of Educational Psychology, 86*(4), 630–639.

Smart, J. L., Purdy, S. C., & Kelly, A. S. (2018). Impact of personal frequency modulation systems on behavioral and cortical auditory evoked potential measures of auditory processing and classroom listening in school-aged children with auditory processing disorder. *Journal of the American Academy of Audiology*, *29*(07), 568-586.

Smith, G. W., & Riccomini, P. J. (2013). The effect of a noise reducing test accommodation on elementary students with learning disabilities. *Learning Disabilities Research & Practice, 28*(2), 89–95.

Smyrnakis, I., Andreadakis, V., Rina, A., Boufachrentin, N., & Aslanides, I. M. (2021). Silent versus Reading Out Loud modes: An eye-tracking study. *Journal of Eye Movement Research, 14*(2).

Soares, J. S., & Storm, B. C. (2020). Putting a negative spin on it: Using a fidget spinner can impair memory for a video lecture. *Applied Cognitive Psychology, 34*(1), 277-284.

Spurlock. (2020). *Accommodations Toolkit: Math Charts/Tables Research*. National Center on Educational Outcomes (NCEO).

Stein J. (2014). Dyslexia: the Role of Vision and Visual Attention. *Current developmental disorders reports*, *1*(4), 267–280.

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.

Stone, E., Cook, L., Laitusis, C. C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education, 23*(2), 132–152.

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.

Svensson, I., Nordström, T., Lindeblad, E., Gustafson, S., Björn, M., Sand, C., Almgren/Bäck, G., & Nilsson, S. (2021). Effects of assistive technology for students with reading and writing disabilities. *Disability and rehabilitation. Assistive technology, 16*(2), 196–208.

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.

Thompson, W. F., Schellenberg, E. G., & Letnic, A. K. (2012). Fast and loud background music disrupts reading comprehension. *Psychology of Music, 40*(6), 700-708.

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., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *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.

Unwin, K. L., Powell, G., & Jones, C. R. (2022). The use of Multi-Sensory Environments with autistic children: Exploring the effect of having control of sensory changes. *Autism : the international journal of research and practice, 26*(6), 1379–1394.

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.

Wei, X. (2024). Text-to-speech technology and math performance: A comparative study of students with disabilities, English language learners, and their general education peers. *AERA Open,* 10.

Weis, R., & Beauchemin, E. L. (2020). Are separate room test accommodations effective for college students with disabilities? *Assessment & Evaluation in Higher Education, 45*(5), 724-739.

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.

Wilson, W. J., Harper-Hill, K., Armstrong, R., Downing, C., Perrykkad, K., Rafter, M., & Ashburner, J. (2021). A preliminary investigation of sound-field amplification as an inclusive classroom adjustment for children with and without Autism Spectrum Disorder. *Journal of communication disorders*, 93, 106142.

Witty, C. A. (2021). *The effects of white noise exposure on cognition: An examination of the impacts of white noise presentation on recall and cognitive load* [Master's thesis, Western Kentucky University]. TopSCHOLAR. https://digitalcommons.wku.edu/theses/3506/

Wolf, M. K., Kao, J. C., Rivera, N. M., & Chang, S. M. (2012). Accommodation practices for English language learners in states' mathematics assessments. *Teachers College Record, 114*(3), 1–26.

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).

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. Educational Assessment, 14(3-4), 139–159

Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities, 51*(1), 73–84.

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.

Yang, X. (2020). Assessment accommodations for emergent bilinguals in mainstream classroom assessments: a targeted literature review. *International multilingual research journal, 14*(3), 217-232.

Yu, D., & Legge, G. E. (2020). Effects of task on reading performance estimates. *Frontiers in Psychology, 11,* 2005.

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review, 27*(1), 69-78.

Zante, J. M. S., Faldas, O. K. A., Gayongo, E. J., Morante, I. F. C., Tesoro, J. C., & Agutaya, E. G. L. (2024). *The effectiveness of using a yellow highlighter on memory retention* [Unpublished bachelor's thesis].

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 “How To” Quick Guides (Reading and Mathematics). 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 and Mathematics assessments. Note for 2024-25: Benchmark Reports were not created for science. 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 and Mathematics 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:

$$\mathrm{LB_i} = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i (\theta_M - CSEM - b_i)}}, \tag{A.2}$$

$$\mathrm{UB_i} = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i (\theta_M + CSEM - b_i)}}, \tag{A.3}$$

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*:

$$ELB_{ob} = \frac{\sum_{i \in b} N_{io} \times LB_i}{\sum_{i \in b} N_{io}}, \tag{A.4}$$

$$EUB_{ob} = \frac{\sum_{i \in b} N_{io} \times UB_i}{\sum_{i \in b} N_{io}}, \tag{A.5}$$

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 and mathematics on the Additional Reporting Resources page (Minnesota Assessment Hub > Reporting > Additional Reporting Resources) for information about how you can use the reports in your district or school.