MICHIGAN Department of Education

Technical Report

Spring 2019

Michigan Student Test of Educational Progress

(M-STEP)

This document has been formatted to be ADA compliant.

TABLE OF CONTENTS

Executive Summary

In June 2014, the Michigan legislature required the Michigan Department of Education (MDE) to develop a new assessment to administer in the spring of 2015. MDE, in conjunction with its testing vendors, worked to create a new assessment system called the Michigan Student Test of Educational Progress, or M-STEP. M-STEP is designed to effectively measure student mastery and growth in comparison to Michigan state standards. The assessment program is made up of three content areas: English Language Arts (ELA), mathematics, and social studies. ELA and mathematics are assessed in grades 3–7, and social studies is assessed in grades 5, 8, and 11. The designs for the ELA and mathematics assessments are based on assessments provided by the Smarter Balanced Assessment Consortium (Smarter Balanced or SBAC) with Michigan-specific blueprints. The social studies assessments are designed specifically for Michigan. For spring 2019, the fourth content area of science assessments included only field test items aligned to the new Michigan state science standards. Since science was not administered operationally, the content area will not be addressed in this report.

This technical report addresses all phases of the testing cycle with the intention of providing evidence to support the validity of the M-STEP summative assessment program. All subsequent chapters of this report constitute evidence for the validity argument that M-STEP was developed with rigor, implemented with fidelity, and validated psychometrically.

E.1 ELA and Mathematics

MDE partners with Smarter Balanced, utilizing its ELA and mathematics test items, and Data Recognition Corporation (DRC) for the creation of M-STEP ELA and mathematics assessments in grades 3–7. Smarter Balanced member states retain flexibility regarding how to customize the assessments so that they may best be used as part of each state's approach to improving their local educational systems. Michigan has taken advantage of this in 2019 by not only customizing the M-STEP blueprints, but also adding passage-based writing (PBW) items to the ELA assessments to assess writing standards. This allowed Michigan to reduce the ELA and mathematics assessments to a legislatively mandated three-hour median testing time (combined) while retaining a writing task in all grades. As a hybrid of the Smarter Balanced assessments and Michigan-selected PBW prompts for ELA, the M-STEP assessments provide key feedback in preparing all Michigan students for success in college and career readiness.

E.2 Social Studies

M-STEP items for social studies are written and reviewed by Michigan educators. Teachers receive training in writing items for standardized assessment and write items testing specific Michigan content standards. Committees of educators review the items for content validity and potential bias issues. These reviews take place both before students see the items on a field test and using student data after they have been field tested. MDE staff and contractor content specialists provide guidance and review throughout this process, ultimately selecting the final items that appear on each test form to cover the full range of Michigan content standards.

E.3 MDE Office of Educational Assessment and Accountability (OEAA)

MDE's Office of Educational Assessment and Accountability (OEAA) has the responsibility of carrying out the requirements in state and federal statutes and rules for statewide assessments. The office oversees the planning, scheduling, and implementation of all major assessment activities and supervises MDE's testing contractors (i.e., DRC and Measurement Incorporated). In addition, the MDE staff from OEAA, in collaboration with outside contractors, conducts quality control activities for every aspect of the development and administration of the assessment program. For additional details for those groups, please refer to Appendix H. OEAA also actively monitors the security provisions of the assessment program.

E.4 Michigan Testing Contractors

DRC is MDE's item development contractor. DRC is responsible for providing test development content leads who work in conjunction with OEAA's content leads. DRC works with OEAA to develop test items. DRC is also a liaison between the Smarter Balanced item bank and OEAA test development staff. MDE administers online assessments to 99% of the students in grades 3–8 and 11. M-STEP is delivered through DRC's online test engine. DRC test development staff are responsible for rendering test items according to OEAA's style guide. Each item is reviewed by both DRC and OEAA content leads to ensure each student is presented with properly formatted test items that are clear and engaging and to ensure the content of each item replicates how the item appears in the item bank.

Measurement Incorporated is Michigan's contractor for paper/pencil materials, handscoring, and reporting. Measurement Incorporated is responsible for the development, distribution, and collection of all paper/pencil test materials as well as the monitoring of test security. Measurement Incorporated produces accommodated testing materials based on the test maps OEAA provides and in accordance with industry standards. Measurement Incorporated scores all the PBW prompts using Michigan-developed rubrics. Once testing is complete, Measurement Incorporated is responsible for developing and providing student results.

E.5 Michigan's Assessment System

Michigan's assessment system is a comprehensive, standards-based system. M-STEP is an accountability assessment, which means that it is used to evaluate school and district success in Michigan's accountability system. Other assessments exist for special populations of students, such as students with significant cognitive disabilities or English learners. All students in grades 3–8 and 11 are required to take Michigan's standards-based accountability assessments. Michigan's accountability assessments are listed in Table E-1 and are described in more detail in Section 3.3.

Table E-1. Michigan Accountability Assessments

Test	Content	Grades
M-STEP	Mathematics	3–7
M-STEP	ELA	3–7
M-STEP	Social Studies	5, 8, 11
PSAT 8/9	Mathematics	8
PSAT 8/9	ELA	8
SAT with Essay	Mathematics	11
SAT with Essay	ELA	11
MI-Access (alternate assessment)	Mathematics	3–8, 11
MI-Access (alternate assessment)	ELA	3–8, 11
MI-Access (alternate assessment)	Science	4, 7, 11
MI-Access (alternate assessment)	Social Studies	5, 8, 11
WIDA	Listening	1–12
WIDA	Reading	K–12
WIDA	Speaking	K–12
WIDA	Writing	1–12

E.6 Changes from Previous Administration

The most significant change to M-STEP between spring 2018 and spring 2019 was the substitution of the PSAT 8/9 for M-STEP in English language arts and mathematics in grade 8. As such, M-STEP grade 8 included only the content area of social studies.

With the assessment otherwise stable, MDE made improvements to supporting materials. A comprehensive *Assessment Coordinator Training Guide* was developed to improve training and reference materials for District and Building Assessment Coordinators. Test administration manuals were revised based on feedback from previous years. MDE conducted additional assessment monitoring. Improvements were made to student reports and the interpretive guides to reports.

E.7 Overview of This Report

Subsequent chapters of this technical report document the major activities of the testing cycle. This report provides comprehensive details confirming that the processes and procedures applied in the M-STEP program adhere to appropriate professional standards and practices of educational assessment. Ultimately, this report serves to document evidence that valid inferences about Michigan student performance can be derived from the M-STEP assessments. Note that part of this report is intended to be utilized in tandem with the *Smarter Balanced 2014–15 Technical Report* (2016) and *Smarter Balanced 2017–18 Technical Report* (2018), while providing additional Michigan-specific validity and reliability information.

Each chapter of this report details the procedures and processes applied in M-STEP as well as the results. Each chapter also highlights the meaning and significance of the procedures, processes, and results in terms of validity and the relationship to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014). Below is a brief overview of the contents of this report.

Chapter 1, "Background of Spring 2019 M-STEP Assessments," describes the background and history of M-STEP.

Chapter 2, "Uses of Test Scores," describes the use of the assessment scores and touches on the validity arguments the technical report intends to address.

Chapter 3, "Test Design and Item Development," describes the involvement of Michigan educators in the item and assessment development process. The assessment development process and the involvement of Michigan educators in that process formed an important part of the validity of M-STEP. The knowledge, expertise, and professional judgment offered by Michigan educators ultimately ensured that the content of M-STEP formed an adequate and representative sample of appropriate content, and that content formed a legitimate basis upon which to derive valid conclusions about student performance. This part of the technical report thus addresses Standard 4.6 of the *Standards* (AERA, APA, & NCME, 2014, p. 87). It shows that the assessment design process, and the participation of Michigan educators in that process, provides a solid rationale for having confidence in the content and design of M-STEP as a tool from which to derive valid inferences about Michigan student performance. This chapter also addresses AERA, APA, and NCME (2014) *Standards* 3.1,3.2, 4.0, 4.1, 4.2, 4.12, and 7.2.

Chapters 4 and 5, "Test Administration Plan" and "Test Delivery and Administration," describe the processes, procedures, and policies that guided the administration of M-STEP, including accommodations, security measures, and written procedures provided to assessment administrators and school personnel. These chapters address AERA, APA, and NCME (2014) *Standards* 4.15, 4.16, 6.1, 6.2, 6.3, 6.4, 6.6, 6.7, and 6.10.

Chapter 6, "Operational CAT," supports Chapter 3 in showing how assessment specification documents, derived from earlier developmental activities, guided the final phases of assessment development and ultimately yielded the assessments administered to students. This chapter thus addresses AERA, APA, and NCME (2014) *Standards* 1.11, 3.1, 3.2, 3.5, 4.0, 4.6, 4.7, 4.8, 4.10, 4.12, 7.2, 8.4, 12.4, and 12.8.

Chapter 7, "Scoring," explains the procedures used for scoring M-STEP autoscored items and handscored items. Chapter 7 adheres to AERA, APA, & NCME *Standards* 4.18, 4.20, 6.8, and 6.9.

Chapter 8, "Operational Data Analyses," describes the data used for calibration and scaling. For content areas for which they are appropriate, raw-score results and a classical item analysis were provided and served as a foundation for subsequent analyses. This chapter also describes the calibration and scaling processes, procedures, and results. Some references to introductory and advanced discussions of Item Response Theory (IRT) are provided. This chapter thereby demonstrates adherence to AERA, APA, and NCME (2014) *Standards* 1.8, 5.2, 5.13, and 5.15.

Chapter 9, “Test Results,” presents scale-score results and achievement level information. Scale-score results provide a basic quantitative reference to student performance as derived through the IRT models that were applied. This chapter thus addresses AERA, APA, and NCME (2014) *Standards* 5.1, 6.10, 7.0, and 12.18.

Chapter 10, “Performance-Level Setting,” provides background on the standard-setting activities and functions to address AERA, APA, and NCME (2014) *Standards* 5.21 and 5.22.

Chapter 11, “Fairness,” address validity evidence, specifically with respect to issues of bias. It demonstrates adherence to AERA, APA, and NCME (2014) *Standards* 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6.

The first half of Chapter 12, “Reliability and Evidence of Construct-Related Validity,” demonstrates adherence to the AERA, APA, and NCME (2014) *Standards* through several analyses of the reliability of the 2019 M-STEP. Information on reliability/precision, standard error of measurement (SEM), conditional standard error of measurement (CSEM), and a detailed examination of classification consistency and accuracy are provided. The first half of Chapter 12 thereby addresses AERA, APA, and NCME (2014) *Standards* 2.0, 2.3, 2.13, and 2.19.

The second half of Chapter 12 addresses validity evidence, including assessment content, response processes, issues of bias, dimensionality analysis, relations to other assessments, and consequences of assessment use. It demonstrates adherence to AERA, APA, and NCME (2014) *Standards* 3.16 and 4.3. This chapter ends with a section addressing the development of validity arguments for M-STEP.

MDE and its testing vendors have maintained an unwavering focus on the gathering of validity evidence in support of M-STEP throughout the development, administration, analysis, and reporting of the 2019 M-STEP administration.

Chapter 1: Background of Spring 2019 M-STEP Assessments

1.1 Background of M-STEP

The Michigan Department of Education (MDE), partnering with Smarter Balanced Assessment Consortium (Smarter Balanced or SBAC), utilizes the ELA and mathematics test items from Smarter Balanced and the passage-based writing (PBW) prompts from Data Recognition Corporation (DRC) for the creation of M-STEP ELA and mathematics assessments. MDE uses test items written by Michigan educators for the M-STEP social studies assessments, as well as for the M-STEP science assessments that were field tested in spring 2019. MDE also partners with DRC for all online delivery, item development, and some psychometric work for the program; and with Measurement Incorporated for the paper/pencil, handscoring, and reporting portions of the program.

In the spring 2019 administration of M-STEP, 99% of Michigan students took M-STEP online. Paper/pencil tests were available for accommodated testing for individual students and for MDE-approved schools that were unable to test online.

1.2 Purpose and Design of ELA and Mathematics M-STEP with Respect to the Smarter Balanced Assessment

Summative assessments measure students' progress toward college and career readiness in ELA and mathematics. These assessments are given at the end of the school year as a computer adaptive test (CAT).

Page ix of the *Smarter Balanced 2017–2018 Technical Report* (2018) details the purposes of the Smarter Balanced summative assessments. Represented in part for this report, the "assessments are to provide valid, reliable, and fair information about" the following:

- students' ELA and mathematics achievement with respect to those Common Core State *Standards* (CCSS) measured by the ELA and mathematics summative assessments,
- whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA and mathematics to be on track for achieving college readiness,
- students' annual progress toward college and career readiness in ELA and mathematics,
- how instruction can be improved at the school, district and state levels,
- students' ELA and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems, and
- students' achievement in ELA and mathematics that is equitable for all students and subgroups of students.

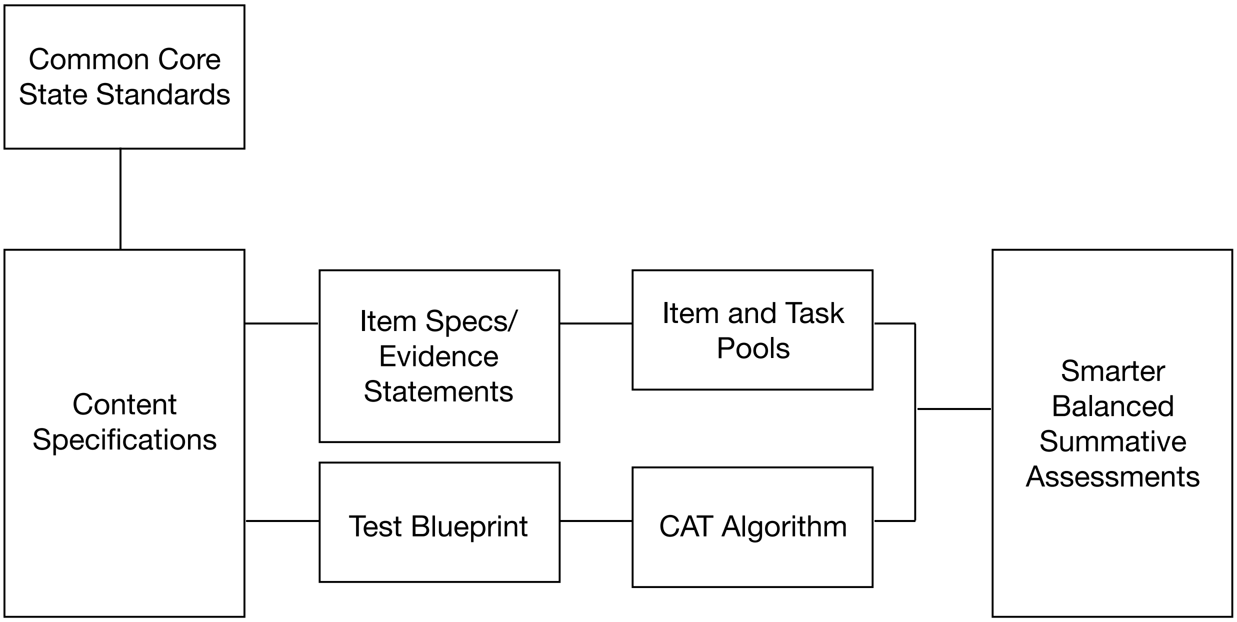
As stated on page 4-4 of the *Smarter Balanced 2017–2018 Technical Report* (2018) summative assessment scores will

- accurately describe both student achievement (i.e., how much students know at the end of the year) and student growth (i.e., how much students have improved since the previous year) to inform program evaluation and school, district, and state accountability systems.
- include writing at every grade and ask students to solve multistep, real-world problems in mathematics.
- capitalize on the strengths of CAT (i.e., efficient and precise measurement with a quick turnaround of results).
- provide valid, reliable, and fair measures of students' progress toward, and attainment of, the knowledge and skills required to be college- and career-ready.
- measure the breadth and depth of the CCSS across the full spectrum of student ability by incorporating a variety of item types (including items and tasks scored by expert raters) that are supported by a comprehensive set of accessibility resources.

The Smarter Balanced assessment system is a valid, fair, and reliable approach to student assessment that provides educators, students, and parents with meaningful results and actionable data to help students succeed.

In developing and maintaining a system of assessments, Smarter Balanced ensures that the assessments' measurement properties reflect industry standards for content, rigor, and performance. A key step in this direction is to ensure that the Smarter Balanced assessments are aligned with the CCSS, which Michigan adopted in 2010. Figure 1-1 (originally from the *Smarter Balanced 2017–2018 Technical Report*, 2018, p. 4-2), shows the components of the assessment.

Figure 1-1. Components of the Smarter Balanced Assessment Design

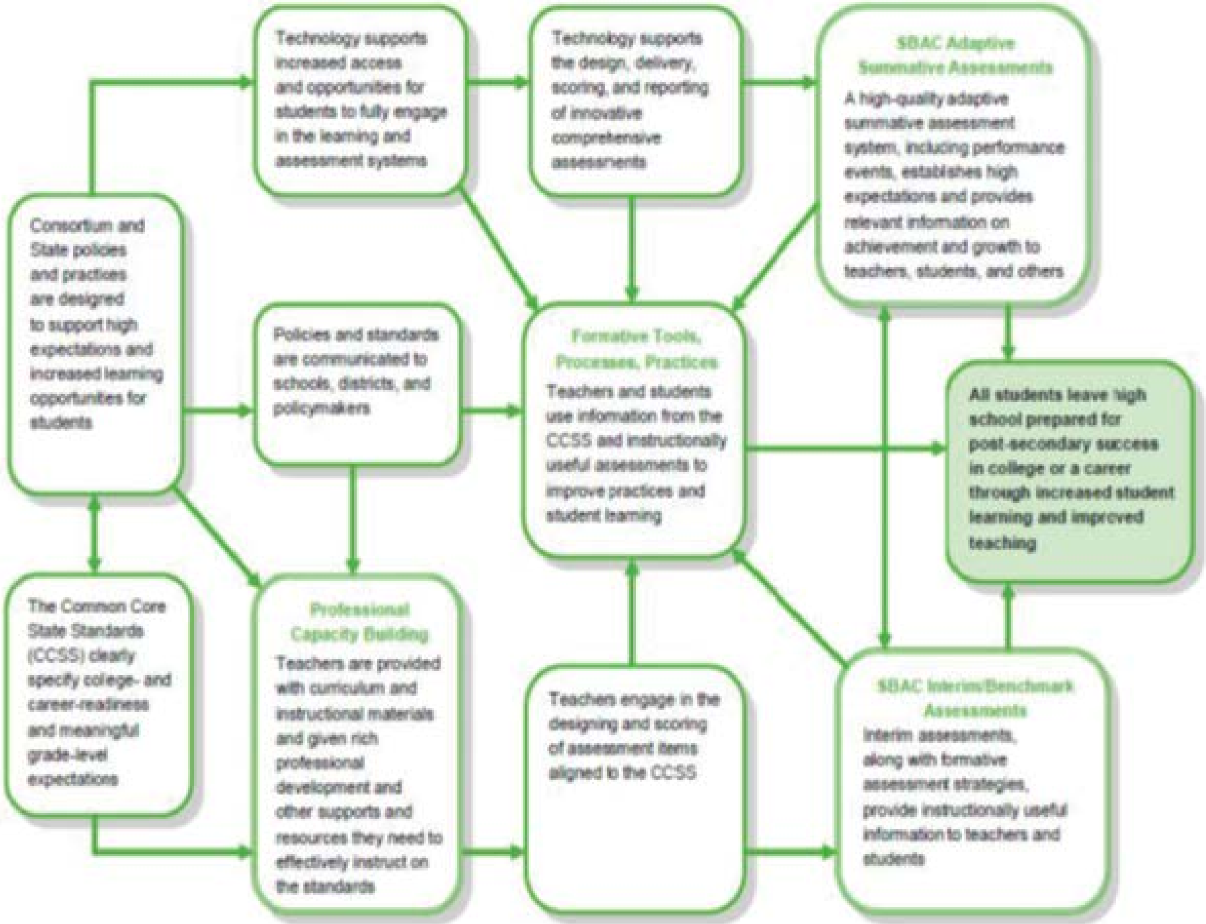

Common Core State Standards
Content Specifications
Item Specs/ Evidence Statements
Test Blueprint
Item and Task Pools
CAT Algorithm
Smarter Balanced Summative Assessments

1.2.1 Background on Smarter Balanced

Smarter Balanced supports the development and implementation of learning and assessment systems to reshape education in member states in order to improve student outcomes. Through expanded use of technology and targeted professional development, the Consortium's Theory of Action calls for the integration of learning and assessment systems, leading to more informed decision-making and higher-quality instruction and ultimately increasing the number of students who are well prepared for college and careers.

The ultimate goal of the Smarter Balanced assessment system is to ensure that all students leave high school prepared for postsecondary success in college or a career through increased student learning and improved teaching. This approach suggests that enhanced learning will result from high-quality assessments that support ongoing improvements in instruction and learning. A quality assessment system strategically balances summative, interim, and formative components (Darling-Hammond & Pecheone, 2010). An assessment system must provide valid measurement across the full range of performance on common academic content, including assessment of deep disciplinary understanding and higher-order thinking skills increasingly demanded by a knowledge-based economy. Figure 1-2 presents an overview of the Smarter Balanced Theory of Action (2011, pg. 7).

Figure 1-2. Overview of Smarter Balanced Theory of Action


Technology supports increased access and opportunities for students to fully engage in the learning and assessment systems
Technology supports the design, delivery, scoring, and reporting of innovative comprehensive assessments
SBAC Adaptive Summative Assessments
A high-quality adaptive summative assessment system, including performance events, establishes high expectations and provides relevant information on achievement and growth to teachers, students, and others
Consortium and State policies and practices are designed to support high expectations and increased learning opportunities for students
Policies and standards are communicated to schools, districts, and policymakers
Formative Tools, Processes, Practices
Teachers and students use information from the CCSS and instructionally useful assessments to improve practices and student learning
All students leave high school prepared for post-secondary success in college or a career through increased student learning and improved teaching
The Common Core State Standards (CCSS) clearly specify college- and career-readiness and meaningful grade-level expectations
Professional Capacity Building
Teachers are provided with curriculum and instructional materials and given rich professional development and other supports and resources they need to effectively instruct on the standards
Teachers engage in the designing and scoring of assessment items aligned to the CCSS
SBAC Interim/Benchmark Assessments
Interim assessments, along with formative assessment strategies, provide instructionally useful information to teachers and students

1.2.2 Test Blueprints

Part of the innovative aspect of the mathematics and ELA assessments is that the test blueprints sample the content domains using both a CAT engine and a PBW prompt. The test blueprints can be inspected to determine the contribution of the CAT and PBW components in a grade and content area toward the construct intended to be measured. Another aspect of the assessments is the provision of a variety of both autoscored and handscored item types. The contribution of these item types is specified in the Smarter Balanced test blueprints.

In February 2015, the governing members of Smarter Balanced adopted blueprints for the summative assessments of ELA and mathematics for grades 3–8 (Smarter Balanced, 2015a; Smarter Balanced, 2015b). These blueprints were fully implemented in the 2014–15 school year and were in effect in the 2018–19 school year for grades 3-7, with grade 8 changing to the PSAT 8/9 for spring 2019.

Since the 2017–18 school year, Michigan has slightly modified the Smarter Balanced blueprints for ELA and mathematics. To reduce testing time, the use of Performance Tasks was eliminated for both mathematics and ELA. In ELA, the PBW prompt was added to assess the writing standards. The net result was that, while the blueprints were modified, all students continued to receive a writing claim score. In mathematics, Michigan added items to Claims 2 and 4 to address any blueprint gaps caused by the removal of the PT items. This testing plan has been continued for the 2018–2019 school year. The ELA and mathematics blueprints are located in Chapter 3, Section 3.3. Due to the drift from the original Smarter Balanced blueprint, it should be noted that Michigan conducted a standards validation in July 2018 to review the M-STEP cut scores and determine if any changes needed to be made. More information can be found in Chapter 10 and Appendix E.

1.3 Purpose and Design of the Social Studies M-STEP

The summative assessments determine students' progress toward college and career readiness in social studies. These are given at the end of the school year. These assessments are primarily delivered online (99% of Michigan students took the test online) with paper/pencil and accommodated options. The social studies assessments are fixed forms. The summative assessments accurately describe student achievement (i.e., how much students know at the end of the year) to inform program evaluation and school, district, and state accountability systems.

The blueprints for social studies contain no constructed-response items, leading to a quick turnaround of results.

The social studies blueprints are located in Chapter 3, Section 3.3.

Chapter 2: Uses of Test Scores

Validity is an overarching component of M-STEP. The following excerpt is from the *Standards for Educational and Psychological Testing* (the *Standards*) (AERA, APA, & NCME, 2014):

> Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (p. 22)

As stated in the *Standards*, the validity of a testing program hinges on the use of the test scores. Validity evidence that supports the uses of M-STEP scores is provided in this technical report. This chapter examines some possible uses of the test scores.

As the *Standards* note, “validation is the joint responsibility of the test developer and the test user” (AERA, APA, & NCME, 2014, p. 13). For ELA and mathematics, the Smarter Balanced Assessment Consortium (Smarter Balanced) does not control aspects of test administration and test use. The Smarter Balanced members deliver the test, score operational items, and provide reports. Members use Smarter Balanced test scores in their own accountability models. In the *Smarter Balanced 2014–15 Technical Report* (2016)[1] and the *Smarter Balanced 2017–18 Technical Report* (2018), guidelines for administration and use are documented. Please see Chapter 1 of the *Smarter Balanced 2017–18 Technical Repor*t for the complete validity argument related to ELA and mathematics, member documentation on specific test administration procedures, reporting, and use for the Smarter Balanced assessments.

The following chapters of this technical report provide additional evidence for these uses as well as technical support for some of the interpretations and uses of test scores. The information in Chapters 3 through 12 also provides a firm foundation that M-STEP measures what it is intended to measure. However, this technical report cannot anticipate all possible interpretations and uses of M-STEP scores. It is recommended that policy and program evaluation studies, in accordance with the *Standards*, be conducted to support some of the uses of the test scores.

2.1 Uses of Test Scores

The validity of a test score ultimately rests on how that test score is used. To understand whether a test score is being used properly, the purpose of the test must first be understood. The intended uses of M-STEP scores include

- identifying Michigan students’ strengths and weaknesses;
- communicating expectations for all students;
- evaluating school-, district-, and/or state-level programs; and

[1] https://portal.smarterbalanced.org/library/en/v2.0/2014-15-technical-report.pdf

- informing stakeholders (i.e., teachers, school administrators, district administrators, Michigan MDE staff members, parents, and the public) on progress toward meeting state academic performance standards and meeting the requirements of the state's accountability program.

This technical report refers to the use of the test-level scores (i.e., scale scores and performance levels), claim-level scores, and claim performance indicators[2].

2.2 Test-Level Scores

At the test level, an overall scale score that is based on student performance on the entire test is reported. In addition, an associated performance level is reported. These scores indicate, in varying ways, a student's performance in ELA, mathematics, or social studies. Test-level scores are reported at four reporting levels: the state, school district, school, and student.

Items on the ELA and mathematics test forms were developed by Smarter Balanced. Items on the braille and enlarged print ELA and mathematics forms were also developed by Smarter Balanced. Final pencil/paper and accommodated forms were created using the items developed by Smarter Balanced, but the item selections were finalized by MDE and DRC content development staff. For social studies, all items and test forms were developed by MDE test development staff.

The following sections discuss two types of test-level scores that are reported to indicate a student's performance on M-STEP: (1) the scale score, and (2) its associated level of performance.

2.2.1 Scale Scores

A scale score indicates a student's total performance for each content area on M-STEP. A "scale score" is a statistical conversion of the "raw score" (numbers of questions that are answered correctly and incorrectly) that takes into account differences in questions students might see on different versions of the test across years, forms, or adapted versions of the test. In other words, the scale score represents the student's level of performance, where higher scale scores indicate higher levels of performance on the test and lower scale scores indicate lower levels of performance. Scaling scores permits comparison of assessment results across different test administrations within a particular grade and content area.

Scale scores are not comparable across grade levels or content areas. Scores are scaled within grade levels, so even if the same numbers are used in different grades, it does not mean that the scales form a single "vertical scale." M-STEP is a standards-based test that assesses the standards for each grade, so a very high score on grade 4 standards does not provide a valid estimate of how that student performs on grade 5 standards.

[2] Claim scores are only available for ELA and Math.

Details of the development of M-STEP scale scores are described in Chapter 10, Section 10.3. The scale score is stable because it allows for students' scores to be reported on the same scale regardless of which year the students took the assessment and which form of the assessments the student took. Schools can use scale scores to compare the performances of groups of students across years. These comparisons can then be used to assess the impact of changes or differences in instruction or curriculum. The scale scores can be used to determine whether students are demonstrating the same skill and ability across cohorts within a grade and content area.

2.2.2 Levels of Performance

A student's performance on M-STEP is reported in one of the four levels of performance: Not Proficient, Partially Proficient, Proficient, and Advanced. The cut scores for the ELA and mathematics performance levels were established by Smarter Balanced during the standard setting, which occurred in three phases: online panel, in-person workshop, and cross-grade review in October 2014. These cut scores were then evaluated and confirmed in the Michigan standards validation in July 2018 (see Appendix E).The cut scores for the social studies performance levels were established by MDE in August 2015.

M-STEP performance levels reflect the performance standards and abilities intended by the Michigan legislature, Michigan teachers, Michigan citizens, and MDE. Descriptions of each performance level in terms of what a student should know and be able to do are provided by MDE and are referenced in the M-STEP & MME Performance Level Descriptors.[3]

2.2.3 Use of Test-Level Scores

M-STEP scale scores and performance levels provide summary evidence of student performance. Classroom teachers may use these scores as evidence of student performance in these content areas. At the aggregate level, district and school administrators may use this information for activities such as planning curriculum. The results presented in this technical report provide evidence that the scale scores are valid and reliable indicators of student performance.

2.3 Claim-level Sub-scores for ELA and Mathematics

Claim-level sub-scores are scores on important domain areas within each content area. In most cases, sub-scores correspond to claims, but in mathematics, Claims 2 and 4 are so intertwined that they are reported as a single sub-score. The claims and reporting categories (sub-scores) are primary structural elements in test blueprints and item development. Figures 2.2 through 2.15 from the *Smarter Balanced 2017–18 Technical Report* (2018) provide information on the claims or sub-score reporting categories for ELA and mathematics.

The claim-level performance indicators are reported for ELA and mathematics for each student. A student's performance on each of the ELA and mathematics claims is reported in one of three levels of performance: *Adequate progress*, *Attention may be needed*, and *Most at risk of falling behind*. Performance-level indicator designations are based on the standard error of

[3] https://www.michigan.gov/documents/mde/M-STEP_and_MME_Performance_Level_Descriptors_671902_7.pdf

measurement of the claim-level sub-score and the distance of the claim sub-score from the proficient cut score. If the proficient cut score falls within a 1.5 SEM error band, it is designated as “Attention may be needed.” If the Level 2/3 cut score is above the error band, the sub-score is designated as “Most at risk of falling behind;” if the cut score is below the error band, the claim level sub-score is “Adequate Progress.”

The purpose of reporting claim-level sub-scores on M-STEP is to show for each student the relationship between the overall performance being measured and the skills in each of the areas delimited by the claims in ELA and mathematics. Teachers may use these sub-scores for individual students as indicators of strengths and weaknesses, but they are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observations. Chapter 12 of this technical report provides evidence of content validity and reliability that supports the use of the claim-level sub-scores. Chapter 12 of this technical report also provides evidence of construct validity that further supports the use of these sub-scores (for additional information see *Smarter Balanced 2017–18 Technical Report* (2018), p. 5-18)

Figure 2-1. ELA Claims

Claim #1—Reading

- Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.

Claim #2—Writing

- Students can produce effective and well-grounded writing for a range of purposes and audiences.

Claim #3—Speaking and Listening

- Students can employ effective speaking and listening skills for a range of purposes and audiences. At this time, only listening is assessed.

Claim #4—Research

- Students can engage in research/inquiry to investigate topics and to analyze, integrate, and present information.

Figure 2-2. Mathematics Claims

Claim #1—Concepts and Procedures

- Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.

Claim #2—Problem Solving/Claim #4-Modeling and Data Analysis

- Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. Students can analyze complex real-world scenarios and can construct and use mathematical models to interpret and solve problems.
- Students can analyze complex real-world scenarios and can construct and use mathematical models to interpret and solve problems.

Claim #3—Communicating Reasoning

- Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

Chapter 3: Test Design and Item Development

3.1 Overview

This chapter is particularly relevant to AERA, APA, & NCME (2014) *Standards* 4.0, 4.1, and 4.7. It also addresses *Standards* 3.1, 3.2, 3.9, 4.12, and 7.4, which will be discussed in pertinent sections of this chapter. *Standards* 4.0, 4.1, and 4.7 are from Chapter 4 of the AERA, APA, & NCME (2014) *Standards*, "Test Design and Development." AERA, APA, & NCME (2014) Standard 4.0 states the following:

> Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (p. 85)

The purpose of this chapter is to document the test design and item development process used for M-STEP. This chapter describes steps taken to create M-STEP, from the development of test specifications to the selection of operational items.

3.1.1 A Brief Description of Smarter Balanced Content Structure for ELA and Mathematics

The Common Core State *Standards* (CCSS) are the content standards in ELA and mathematics that many states have adopted. Because the CCSS were not specifically developed for assessment, they contain extensive rationale descriptions and information concerning instruction. Adopting previous practices used by many state programs, Smarter Balanced content experts produced content specifications in ELA and mathematics, which distill assessment-focused elements from the CCSS. The Smarter Balanced *Content Specifications for the Summative Assessment of the CCSS for English Language Arts/Literacy* (2015a) and *Content Specifications for the Summative Assessment of the CCSS for Mathematics* (2015b) were expressly created to guide the structure and content of assessment development. Within each of the two content areas in grades 3–8, there are four broad claims. Within each claim, there are a number of assessment targets. The claims in ELA and mathematics are given in Table 3–1 (from the *Smarter Balanced 2017–18 Technical Report* (2018), p. 5–18).

Table 3-1. Claims for ELA and Mathematics

Claim	ELA	Mathematics
1	Reading	Concepts and Procedures
2	Writing	Problem Solving
3	Speaking/Listening	Communicating Reasoning
4	Research	Modeling and Data Analysis

Currently, only the listening part of ELA Claim 3 is assessed. In mathematics, Claims 2 and 4 are reported together as a single sub-score, so there are only three reporting categories for mathematics but four claims.

Because of the breadth in coverage of the individual claims, targets within each claim were needed to define more specific performance expectations. The relationship between targets and CCSS elements is made explicit in the Smarter Balanced content specifications (2015a; 2015b).

The *Item and Task Specifications* (Smarter Balanced, 2015c) for ELA and mathematics provide guidance on how to translate the Smarter Balanced content specifications into assessment items. In addition, guidelines for bias and sensitivity issues, accessibility and accommodations, and style help item developers and reviewers ensure consistency and fairness across the item bank. The specifications and guidelines were reviewed by member states, school districts, higher education representatives, and other stakeholders. The item specifications describe the evidence to be elicited and provide sample task models to guide the development of items that measure student performance relative to the target.

The Smarter Balanced assessment blueprints found in the *Smarter Balanced 2017–18 Technical Report* (2018) describe the content of the ELA and mathematics summative assessments for grades 3–7 administered in the 2018–19 school year and how that content was assessed. The blueprints also describe the composition of the assessment and its scoring. Specific items administered to each student are uniquely determined based on an item-selection algorithm that includes content constraints that correspond to the test blueprint. Developed with broad input from member states, partners, and stakeholders, the summative test blueprints reflect the depth and breadth of the performance expectations of the CCSS. Smarter Balanced governing members adopted the preliminary test blueprints in 2012. The summative test blueprints that were subsequently developed contain refinements and revisions based on the analyses of the pilot and field tests.

3.1.2 Evidence-Centered Design in Constructing Smarter Balanced Assessments

The *Smarter Balanced 2017–18 Technical Report* (2018) discusses the concept of evidence-centered design:

> Evidence-centered design (ECD) is an approach to the creation of educational assessments in terms of reasoning about evidence (arguments) concerning the intended constructs. The ECD process begins with identification of claims, or inferences, users want to make concerning student achievement. Evidence needed to support those claims is then specified, and finally, items/tasks capable of eliciting that information are designed (Mislevy, Steinberg, & Almond, 2003). Explicit attention is paid to the potential influence of unintended constructs. The ECD process accomplishes this in two ways. The first is by incorporating an overarching concept of assessment as an argument made from imperfect evidence. This argument makes explicit the claims (i.e., the inferences that one intends to make based on scores) and the nature of the evidence that supports those claims (Hansen & Mislevy, 2008; Mislevy & Haertel, 2006). The second is by distinguishing the activities and structures involved in the assessment enterprise to exemplify an assessment argument in operational processes. By making the underlying evidentiary argument more explicit,

> the framework makes operational elements more amenable to examination, sharing, and refinement. Making the argument more explicit also helps designers meet diverse assessment needs caused by changing technological, social, and legal environments (Hansen & Mislevy, 2008; Zhang, Haertel, Javitz, Mislevy, Murray, & Wasson, 2009). The ECD process entails five types of activities. The layers focus in turn on the identification of the substantive domain to be assessed; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and the way they function in an operational assessment, as described below. For Smarter Balanced, a subset of the general ECD elements was used. (p. 4-4)

3.1.3 A Brief Description of Content Structure for Social Studies

M-STEP content in social studies is defined by the knowledge and skills identified in the Michigan state standards. Michigan state standards were approved by the Michigan State Board of Education after consultation and collaboration with educators and the general public, representing consensus of the essential content for Michigan learners. Evidence of validity based on test content includes information about the test specifications, including the test design and test blueprint. Test development involves creating a design framework from the statement of the construct to be measured. The M-STEP social studies test specifications evolve from the tension between the constraints of the assessment program and the benefits sought from the examination of students. These benefits and constraints mix scientific rigor with policy considerations.

The M-STEP test specifications consist of a blueprint and test maps for each grade level and content area. For social studies, the 2019 M-STEP test selection specifications were finalized by MDE and its psychometricians and vendors in 2018.

The key structural aspect is the test blueprint, which specifies the target score points for each discipline in social studies, as shown in Table 3-6. The blueprint represents a compromise among many constraints, including the target weights for each discipline, availability of items from field testing, and results of multiple reviews by content specialists. Test design includes such elements as number and types of items for each of the scores reported. The 2019 M-STEP operational forms matched the test blueprint that was intended for this assessment.

3.2 Test Blueprints

Test specifications and blueprints define the knowledge, skills, and abilities intended to be measured on each student's test event. A blueprint also specifies how skills are sampled from a set of content standards (i.e., the CCSS or Michigan state standards). Other important factors, such as Depth of Knowledge (DOK), are also specified. Specifically, a test blueprint is a formal document that guides the development and assembly of an assessment event/form by explicating the following types of essential information:

- content (i.e., claims/disciplines and assessment targets) that is included for each assessed content area and grade across various levels of the system (i.e., student, classroom, school, district, and state levels)
- the relative emphasis of content standards generally indicated as the number of items or percentage of points per claim and assessment target
- item types used or required, which communicate to item developers how to measure each claim and assessment target and communicate to teachers and students about learning expectations
- DOK, indicating the complexity of item types for each claim and assessment target

The test blueprint is an essential guide for both assessment developers and for curriculum and instruction. For assessment developers, the blueprint and related test-specification documents define how the test will ensure coverage of the full breadth and depth of content and how it will maintain fidelity to the intent of the CCSS and/or Michigan state standards on which the assessments are based. Full content alignment is necessary to ensure that educational stakeholders can make valid, reliable, and unbiased inferences about student, classroom, school, and state performance. At the instructional level, the test blueprint provides a guide to the relative importance of competing content demands and suggests how the content is demonstrated, as indicated by item type and DOK. In summary, an assessment blueprint provides clear development specifications and signals to the broader education community both the full complexity of the standards and how performance on these standards is substantiated.

3.2.1 Test Specifications

AERA, APA, and NCME (2014) Standard 4.1 states the following:

> Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (p. 85)

The purpose of M-STEP is discussed in Sections 1.2 and 1.3 of Chapter 1. M-STEP tests the knowledge and skills that are identified within Michigan's standards-based accountability system. This framework, in turn, is based on prior consensus among MDE staff, Michigan educators, and experienced content-matter experts that the framework represents content that is important for teachers to teach and students to learn.

The test specifications are discussed in accordance with AERA, APA, and NCME (2014) Standard 4.12, which states the following:

> Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (p. 89)

Item and test development are guided by sets of specifications. Details on these specifications for ELA and mathematics can be found in the *Smarter Balanced 2017–2018 Technical Report* (2018), the *Item and Task Specifications* (Smarter Balanced, 2015c), and the *Content Specifications for the Summative Assessment of the Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects* (2015a). While MDE reviews all Smarter Balanced operational items, MDE utilizes the Smarter Balanced documentation for the technical details of item and test development. The remainder of this section will focus on the details for Michigan-developed assessments and items (operational items and test maps for social studies and field test items for all content areas).

A general description of development activities applying to Michigan-developed assessments (i.e., M-STEP social studies) is provided below. OEAA staff, contractors, and Michigan educators work together to develop these state assessments. Specifically, the development cycle includes the following steps:

- Item writer training
- Item development
- Item review
- Field-testing
- Field-test data review
- Operational test construction

3.2.2 Item Writer Training

Once item specifications are finalized, Michigan's item development contractor uses customized materials approved by OEAA to train item writers to author items specifically for M-STEP. Item writer training can last anywhere from three to five days and is conducted by contractor staff in conjunction with OEAA test development staff. The process of item writing includes cycle(s) of feedback from contractor and OEAA staff and can take between 4 and 8 weeks for an item to move from initial assignment to accepted status. All item writers are Michigan educators who have curriculum and instruction expertise for the grade and content for which they are writing items. In addition, prospective item writers are required to submit three original test items aligned to grade-specific content standards, which OEAA test development staff review and possibly approve for item authoring. Michigan's item writers possess relevant degrees and experience, and many have previous experience in item writing that is M-STEP specific.

3.2.3 Item Development

Item development is discussed in this section in compliance with the AERA, APA, and NCME (2014) *Standards*. Standard 4.7 states the following:

> The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (p. 87)

For ELA and mathematics, development of item content for the operational test was completed by Smarter Balanced from 2012 to 2014. Smarter Balanced tested items and refined its approach to item development through three steps: small-scale tryouts in fall 2012, a large pilot test in 2013, and a field test in spring 2014. Items administered for the 2019 M-STEP operational test complied with Smarter Balanced content specifications and with the item and task specifications that were refined after the pilot test and before the field test. Further details can be found in Chapter 3 in the Item Development section of the *Smarter Balanced 2017–2018 Technical Report*.

For social studies items and Michigan-developed ELA and mathematics field test items, Michigan item writers drafted test items in accordance with item specifications approved by OEAA test development staff. Contractor staff reviewed items and OEAA test development staff conducted additional review. Sections 3.2.6 and 3.3 discuss how the items are selected for field-testing or operational use. The review consisted of meeting the following criteria:

Skill:

- Item measures one skill level.
- Item measures skill in manner consistent with specifications.
- Item assesses appropriate (i.e., realistic) level of skill.
- Item makes clear the skill to be employed.

Content:

- Item measures one primary academic standard.
- Item measures academic standard in a manner consistent with specifications.
- Item taps appropriate (i.e., important) aspect of content associated with the academic standard.
- Item makes clear the benchmark or problem to be solved.

Relevance:

- Item is not contrived.
- Item is appropriate for the grade level to be tested.
- Item groups reflect instructional emphasis.

Accuracy:

- Item is factually accurate.
- Multiple-choice (MC) items contain only one correct or best response.
- Multi-select items contain answer choices that are clearly correct or best responses.
- Technology-enhanced (TE) items follow approved style guidelines for each grade and content area.
- If item pertains to disputed content, context for correct answer is clearly defined.
- Item is unambiguously worded.
- Item contains no extraneous material, except as required by the standard.
- Vocabulary is grade-level appropriate and clear.
- Item contains no errors of grammar, spelling, or mechanics.
- Item responses are parallel and related to the stem.
- Item responses are independent.
- Item contains no clues or irrelevant distractors.
- Item is clearly and conveniently placed on the page.
- Physical arrangement of item is consistent with OEAA style guide.
- Keys for sets of MC items are balanced (e.g., equal numbers of As, Bs, Cs, and Ds).

Bias:

- Item is free of race and gender stereotypes.
- Item contains no material known or suspected to give advantage to any group.
- Item is free of insensitive language.
- Item sets that identify race or gender either directly or indirectly are balanced with reference to race and gender.
- Item content and format are accessible to students with disabilities.
- Item content and format are accessible to students with limited English proficiency.

3.2.4 Graphics Creation

For all Michigan-developed items, MDE has an internal team of media designers who use the graphic descriptions submitted by the item writers through Michigan's Item Bank System (IBS) to create the pictures, graphs, maps, artwork, etc. that are needed for online test items. MDE and DRC staff review and approve the completed artwork in preparation for the item review.

3.2.5 Item Review

Continuing from Standard 4.7 above, AERA, APA, and NCME (2014) Standard 3.2 is particularly relevant to fairness in item development:

> Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (p. 64)

The Bias and Sensitivity Review Committees (BSC) are composed of representatives from various backgrounds whose purpose was to screen the items for racial, socioeconomic, gender, and other sensitivity issues. This follows AERA, APA, and NCME (2014) Standard 3.1, which states the following:

> Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)

Panels of educators, including those from Michigan, under Smarter Balanced patronage, reviewed all Smarter Balanced items and item stimuli for accessibility, bias/sensitivity, and content. (Item stimuli include the reading passages used on the ELA assessments and the figures and graphics used on the Mathematics assessments.) During the accessibility reviews, panelists identified issues that could negatively affect a student's ability to access stimuli or items or to elicit valid evidence about an assessment target. During the bias and sensitivity review, panelists identified content in stimuli and items that could negatively affect a student's ability to produce a correct response because of their background. The content review focused on developmental appropriateness and alignment of stimuli and items to the content specifications and appropriate depths of knowledge. Panelists in the content review also checked the accuracy of the content, answer keys, and scoring materials. Items flagged for accessibility, bias/sensitivity, and/or content concerns were either revised to address the issues identified by the panelists or removed from the item pool. The final and approved selection by Smarter Balanced educators became the Smarter Balanced computer adaptive item pool and was used for M-STEP ELA and mathematics tests.

For Michigan-developed items, after the internal reviews take place, all M-STEP items are reviewed by Michigan educators through the Content Advisory Committee (CAC) and BSC. Contractor staff trains the CAC and BSC participants using OEAA-approved materials and facilitates the committee meetings under the leadership of OEAA test development staff. All newly written test items are typically reviewed first by the BSC and then by the CAC.

An item rejected by the BSC may or may not get passed on to the CAC for review. Each review is led by experienced contractor staff, with test development staff in attendance, using the following prescribed guidelines to indicate the final status of each item:

- **Accept:** The criteria outlined in the review were met in all areas (i.e., skill, content, relevance, accuracy, and bias), and the item appears suitable for field-testing.

- **Revise:** One or more of the criteria have not been met or the item needs minor changes to make it acceptable. Reviewers recommend changes to be made to the item that will make the item suitable for field-testing.
- **Reject:** Several category conditions have not been met, are suspect, or need radical changes to make the item acceptable. In such cases, the item may be vague or ambiguous, inappropriate, or not clearly related to the text or the standard. Without extensive revisions, it is unlikely to be salvaged. Reviewers provide comments to explain why the item should be rejected.

Items that have passed bias/sensitivity and content reviews are eligible for field-testing.

3.2.6 Field-Testing

Before an item can be used on an operational test or added to the operational item pool, it must be field-tested. OEAA uses two approaches to administer field-test items: embed field-test items in an operational administration or use field-test items in a stand-alone field-test administration. Items that have passed bias/sensitivity and content review are eligible for field-testing.

OEAA embeds field-test items in multiple forms of operational fixed-form assessments or randomly assigns field-test items to students across the state during the computer adaptive test (CAT) administrations. Administering field-test items this way ensures that they are randomly distributed, and this allows a large representative sample of responses to be gathered under operational conditions for each item. Enough field-test items are administered annually to replenish and improve the item pools.

When MDE implements testing at new grade levels, for new content areas, or for revised academic standards, it can be necessary to conduct a separate stand-alone field test to obtain performance data. In 2019, MDE administered a stand-alone field test in science and for Michigan-developed passage-based writing (PBW) items.

3.2.7 Range-Finding

After the student responses to the field-tested PBW prompts are collected, a range-finding is conducted to determine scoring guidelines and score-point ranges for the different score points for each field-tested writing item. This information is then used in the preparation of materials to guide the handscoring of the PBW item student responses by a trained team of readers, as described in Chapter 7 of this report.

3.2.8 Data Review

After field-testing, MDE psychometric staff analyze results. Contractor staff and test development staff convene data review committee meetings with Michigan educators. Significant effort goes into ensuring that these committee members represent the state demographically with respect to ethnicity, gender, school district size, and geographical region. These committees receive training on interpreting the psychometric data compiled for each field-test item by OEAA psychometric staff. Content experts (usually teachers) and group facilitators apply this training to the data review process. During these data review meetings, participants review the items with field-test statistics. Data provided to the data review committees are separated by BSC and CAC. The data that are reviewed during BSC include

- *N*-count;
- adjusted *p*-value (i.e., adjusted item mean in the range of 0–1 for all items);
- Differential Item Functioning (DIF) flag;
- favored group; and
- percentage of students who choose each option (option-total correlation), omit a response (omit-total correlation), and in paper/pencil tests, submit multiple marks (multiple marks-total correlation).

The data that are reviewed during CAC include

- overall *N*-count;
- adjusted *p*-value;
- difficulty flag;
- item-total correlation;
- item-total flag; and percentage of students who choose each option (option-total correlation), omit a response (omit-total correlation), and in paper/pencil tests, submit multiple marks (multiple marks-total correlation).

As mentioned above, specific directions are provided on the use of the statistical information and how to use Michigan's IBS. BSC members evaluate each test item for fairness issues with respect to culture, ethnicity, gender, geographic location, and economic status, using the data listed above for this group. CAC members evaluate each test item with regard to alignment to the academic content standard, grade-level appropriateness, and level of DOK, using the data information listed above for this group. Both committees then recommend that the item be accepted, revised for additional field-testing, or rejected.

After new items have survived all reviews and field-testing, they are saved in the Michigan IBS as "Ready for Operational," meaning they are now eligible for operational use.

3.2.9 Development of the Passage-Based Writing Items

A Passage-based Writing (PBW) item requires text-dependent analysis, based on a passage or a multiple-passage set that each student reads during the assessment. Both literary and informational texts are addressed through this item type. Students must draw on basic writing skills while inferring and synthesizing information from the passage(s) in order to develop a comprehensive, holistic essay response. The reading and writing skills required of a student in response to a PBW item coincide with the skills required for a student to be college and career ready. The current PBW items are selected from a bank of items developed internally at DRC.

Table 3-2. Item Development Process for Passage-Based Writing (PBW) Items

Step	Description
Analyze Item Bank	Existing test items in DRC's current college-and career- ready (CCR) bank are reviewed for technical psychometric quality as well as for their match to the college- and career- ready standards. During this phase, test development specialists also make a tally of the test items by content and standard expectations—including test development specialists' best thinking regarding the number of usable test items in the existing item bank. A tally is also made of the number of usable passages.
Refine Test Item Development Plan to Include Writers and Subcontractors	Item and test development specialists identify the writers who will write the test items (test development specialists or other professional item writers, subcontractors, etc.), the estimated number of writers needed, the qualifications of writers, and the approximate number of test items to be submitted by each source.
Passage Development	Passages are developed by published authors as well as technical writers that have previous publishing experience. DRC trains the writers for passage development specific to assessment ensuring that bias, sensitivity, text complexity and difficulty, as well as technical quality are adhered to as passages are being submitted for review. Once received, DRC sends the passages through a rigorous quality review process. Upon acceptance and completion of the internal processes, passages are reviewed by external educators to validate use and appropriateness for the grade level being assessed. As most passages are commissioned texts, adjustments are made as needed post review and prior to field test administration.
Train Item Writers	Item and test development specialists train item writers to write PBW items based on the big ideas in the passage or passages, as needed. Item writers receive the college-and career-ready standards as well as all other technical quality material to ensure alignment to the expected content for the CCR bank. Item writers who have written for DRC in the past receive updated information, as needed.
Write and Review Items	Test items and sample responses are written by item writers after training is complete. Feedback is provided by the item and test development specialists to item writers on a regular basis. As test items are written, they are reviewed and edited in a series of internal reviews. Item and test development specialists review and edit items to include, but not limited to, the following: match to the standard, relevance to purpose, accuracy of content, item difficulty, interest level, grade appropriateness, depth of knowledge and cognitive complexity, adherence to the principles of Universal Design, and freedom from issues of bias/fairness/sensitivity. Sample responses are reviewed and edited to ensure they are accurate representatives of student responses for the PBW items. At the same time, the process of procuring permissions (if applicable) also begins, including securing permissions for passages, art, etc.
Enter Test Items into Database	Upon acceptance from item writers, test items are entered into the item management system, IDEAS (*Item Development and Educational Assessment System*). Item data stored in the system database includes, but is not limited to, the following: readability, cognitive level, estimated level of difficulty, alignment to standards, and correlation to stimulus prompts and passages. The sample responses are also included with the items in IDEAS.

Step	Description
Review and Revise Items	All PBW items are reviewed again by test development specialists in preparation for field testing. Several rounds of internal reviews are conducted to ensure quality and adherence to standard test development practices. Item and test development specialists incorporate all review feedback, and requested edits to items are made.
Field Testing	The PBW items are all field or pilot tested in multiple state programs. In most cases, the field test is a stand-alone field test to allow for shorter testing times on the overall summative assessments. Once field tested, the items undergo a rigorous range-finding process with groups of educators or internal DRC test development content personnel in order to gather and create anchor, training and qualifying sets as needed for scoring.
Creation of Scoring Materials	The PBW items in the CCR bank are scored using a 4-point holistic rubric. It is the holistic rubric that DRC uses for all PBW items within the CCR bank. Key evaluation material including evidence and elaboration, structural organization, language demands, and grammar and mechanics are the overall focal points on the rubric. The rubric has been vetted through the scoring process as well as the quality review process to ensure accuracy and success during scoring. During the range-finding process, DRC's Performance and Assessment Scoring department create all necessary materials for scoring the items including anchor, training, and qualifying sets.
Review Results of the Field Test	Following the administration of a field test form and the subsequent range-finding and field test scoring processes for field test items, performance data for all field test items are analyzed by DRC psychometricians and test development specialists. Test item performance data that meet certain triggering criteria are flagged for additional reviews by test development specialists. Flagged field-test items with extreme performance data are considered psychometrically unusable and are removed from future operational consideration.

3.3 Operational Test Construction

OEAA test development staff build test maps that meet the test specifications (i.e., blueprint and psychometric specifications) inside Michigan's IBS. All test maps are reviewed for correct answer keys, accurate content standards, and appropriate statistical/psychometric information for each item. In addition, comparability of the overall test across forms and across adjacent years is also examined for social studies. Corresponding details for the content areas are presented below.

3.3.1 ELA

M-STEP ELA is based on Michigan's ELA academic content standards, which were adopted by the State Board of Education in 2010. M-STEP ELA consists of four claims: Reading, Writing, Listening, and Research. The assessment is administered in grades 3–7.

M-STEP ELA is a CAT using Smarter Balanced items, all of which are reviewed and approved by OEAA staff for use in Michigan's CAT. Also, each CAT form distributes one PBW prompt per student. The PBW prompts were developed by DRC and reviewed by OEAA staff. In addition, Michigan embeds five ELA field-test items in each form for grades 3–7.

In the CAT at all grades, Claim 1 (Reading) consists of both informational and literary passages, each with related items. Passages are assessed using MC items and a variety of technology-enhanced items, such as hot text, drop-down menus, and multi-select items. Claim 2 (Writing) includes student writing samples with a set of associated items, some independent items, and one PBW item. PBW prompts cover all Claim 2 content categories. Claim 3 (Speaking/Listening) consists of 3 or 4 listening passages, each with 2 or 3 associated items. Claim 4 (Research) consists of 8 or 9 independent items. The ELA assessment structure is summarized in Table 3-3.

Table 3-3. ELA Structure for Grades 3–7

Claim/Score Reporting Category	Content Category	CAT Stimuli	PBW Stimuli	CAT Items	PBW prompts
1. Reading	Literary	2	0	7–8	0
1. Reading	Informational	2	0	7–8	0
2. Writing	Organization/Purpose and Evidence/Elaboration	0	1	6–8	1
2. Writing	Conventions	0	0	5	0[1]
3. Speaking/Listening	Listening	3–4	0	8–9	0
4. Research	Research	0	0	8–9	0

3.3.2 Mathematics

M-STEP mathematics is based on Michigan’s mathematics academic content standards, which were adopted by the State Board of Education in 2010. M-STEP mathematics consists of four claims: Concepts and Procedures, Problem Solving, Communicating Reasoning, and Modeling and Data Analysis. The assessment is administered in grades 3–7.

There are non-calculator portions of the mathematics assessment embedded throughout the online test. All items in grades 3–5 are non-calculator items.

M-STEP mathematics is a CAT using Smarter Balanced items, all of which are reviewed and approved by OEAA staff for use in Michigan’s CAT. Michigan embeds five mathematics field-test items in the CAT in each form in grades 3–7.

In the mathematics assessment, the Claim 1 (Concepts and Procedures) section consists of 20 items (MC or TE) in the CAT. Details of the various TE types can be found in Section 3.7. The Claim 2 (Problem Solving) section consists of 4 items. The Claim 3 (Communicating Reasoning) section consists of 8 items. The Claim 4 (Modeling and Data Analysis) section consists of 4 items. Claims 2 and 4 are combined in the blueprint and reporting structure because of content similarity and to provide flexibility for item development. There are still four claims, but only three claim scores are reported with the overall mathematics score. The mathematics assessment structure is summarized in Tables 3-4 and 3-5.

[1] PBW prompts cover all Claim 2 content categories but are listed under only one in Table 3-2 to avoid double-counting.

Table 3-4. Mathematics Overall Structure: Number of Items Claim/Reporting Category

Claim/Score Reporting Category	Grades 3–7
1. Concepts and Procedures	20
2. Problem Solving and 4. Modeling and Data Analysis	8
3. Communicating Reasoning	8

Table 3-5. Mathematics Structure for Grades 3–7

Claim/Score Reporting Category	Content Category	CAT Items
1. Concepts and Procedures	Priority Cluster	15
1. Concepts and Procedures	Supporting Cluster	5
2. Problem Solving and 4. Modeling and Data Analysis	Problem Solving, Modeling and Data Analysis	8
3.Communicating Reasoning	Communicating Reasoning	8

3.3.3 Social Studies

M-STEP social studies is based on Michigan's social studies academic content standards, which were adopted by the State Board of Education in 2007. The assessment is administered in grades 5, 8, and 11. The M-STEP social studies assessment in grade 5 consists of five domains: History, Geography, Civics and Government, Economics, and Public Discourse. There are 45 operational items and 15 embedded field-test items. The M-STEP social studies assessment in grade 8 consists of four domains: History, Geography, Civics and Government, and Economics. There are 44 operational items and 22 embedded field-test items. The M-STEP social studies assessment in grade 11 social studies assessment consists of four domains: U.S. History and Geography, World History and Geography, Civics, and Economics. There are 38 operational items and 16 embedded field-test items. The social studies assessment structure is summarized in Table 3-6.

Table 3-6. Social Studies Structure for Grades 5, 8, and 11

Grade	Domain	# of Operational Items
5	History	19
5	Geography	7
5	Civics and Government	10
5	Economics	7
5	Public Discourse	2
8	History	21
8	Geography	14
8	Civics and Government	4
8	Economics	5
11	U.S. History and Geography	12
11	World History and Geography	12
11	Civics	7
11	Economics	7

3.3.4 Science

M-STEP science is based on Michigan's science academic content standards, which were adopted by the State Board of Education in 2015. Because science is tested once in a three-year grade band, it would not be appropriate to test under the new standards until schools have had three years of instruction under the new standards, nor would it be appropriate to test under the previous standards. Instead, M-STEP science was a statewide field test for spring 2019.

3.3.5 Accommodations

Michigan is committed to ensuring all students, including English learners and students with disabilities, have access to a wide array of tools across M-STEP. Sections 4.1–4.3 in this report detail the Universal Tools, Designated Supports, and Accommodations Michigan provides. It is important to note that M-STEP is available to students who require Accommodations according to their Individualized Education Program (IEP) or 504 plan. Paper/pencil accommodated tests are available in many different forms to meet the needs of Michigan's students by providing the tests in contracted and uncontracted braille, enlarged print, and translated forms of the tests. Students may also test online with many different options such as video sign language (American Sign Language and Signed Exact English), stacked Spanish, English text-to-speech, and closed captioning (in the Listening claim). Whether students take a test with or without Universal Tools, Designated Supports, and Accommodations, the M-STEP assessments are administered during the same testing window as standard operational tests.

3.4 Sources of Items and Metadata

3.4.1 ELA and Mathematics

M-STEP ELA and mathematics have three sources for test items:

1. The Smarter Balanced Assessment Consortium
2. The Michigan Item Bank System (IBS)
3. The DRC Item Development and Educational Assessment System (IDEAS)

Smarter Balanced worked with a variety of assessment vendors, state education departments, and educators throughout 2012 to create a pool of ELA and mathematics test items in preparation for pilot testing. In the process of creating the test items, the item writers were provided trainings in evidence-centered design, universal design, DOK, accessibility, and issues of bias and sensitivity. The item writers also received content and item specifications to guide their development. Each test item passed through approval from a content committee, an accessibility committee, and a bias and sensitivity committee before being added to the item pool.

In 2013, Smarter Balanced conducted a pilot test in a small number of schools across states participating in Smarter Balanced, using items from the existing item pool. Smarter Balanced used feedback from this pilot test in preparation for further item development and testing. In 2014, Smarter Balanced administered a field test of the existing item pool to more than 4 million students across states participating in the consortium, including Michigan. Smarter Balanced conducted a subsequent data review using educator committees to evaluate the performance of the test items across the country and to ensure that the items met the quality levels required in terms of content, accessibility, and issues of bias and sensitivity to be included in the operational item pool. The items were then made available to Michigan for inclusion in the CAT item pool. Each year, additional items are field-tested to replenish the general item pool.

The Michigan IBS contains items that have been developed and reviewed by Michigan teachers using processes described earlier in this chapter. The items from both sources (i.e., Smarter Balanced and Michigan IBS) contained a mixture of MC and TE item types, with a DRC-developed PBW prompt added for each student.

Passage-Based Writing (PBW) items are developed by DRC as described in 3.2.9 and stored in the DRC IDEAS item bank. PBW items are reviewed by DRC and OEAA content leads before being included in M-STEP.

3.4.2 Social Studies

The item development process for M-STEP social studies utilizes the Michigan IBS as its main resource. The Michigan IBS is a secure, web-based application that allows users to create contexts and test items. It leads users through all the steps of the item development process, including context review, item review, and data review as described in Section 3.2.

3.5 Import into DRC INSIGHT Test Engine

M-STEP is administered through the DRC INSIGHT test engine. The test items must be imported into INSIGHT from the various sources noted earlier. Once the items are loaded into INSIGHT, they can be rendered for review in the identical formatting structure in which a student would see the item in a test. After the items have been formatted and rendered, they can be assembled into online test forms based on the sequence and information provided in the test maps.

3.6 Psychometric Review During Assessment Construction

Content specialists and psychometricians both from MDE and from Smarter Balanced followed psychometric guidelines and targets for operational forms construction. The foremost guideline was for item content to match the test blueprint for the given content. Both groups used item flagging criteria (discussed below) to guide the assessment construction. Items with flags were avoided when possible.

Details for psychometric reviews are described below by content area groups. Such reviews for ELA and mathematics are done by the Smarter Balanced psychometrician(s), while social studies reviews are carried out by an MDE psychometrician.

3.6.1 ELA and Mathematics

The psychometric review for the items in the M-STEP CAT pool and fixed forms was conducted by Smarter Balanced. Smarter Balanced flagged items based on the following content criteria (Smarter Balanced, 2016, p. 4–22):

- The following items were flagged based on item difficulty and score distribution:
  - items with a low average item score (i.e., less than .10)
  - items with a high average item score (i.e., greater than .95)
  - items with a proportion obtaining any score category less than 0.03

- The following items were flagged based on item discrimination:
  - items with a low item-total correlation (i.e., less than .30)
  - items with a higher mean criterion score for students in a lower score-point category

- The following multiple-choice items were flagged:
  - items where higher ability students (i.e., those in the top 20% on overall score) select a distractor more often than the key
  - items with a higher criterion score mean for students choosing a distractor than the mean for those choosing the key
  - items with a positive correlation between distractor and total score

Items are also classified into three Differential Item Functioning (DIF, for corresponding details please see Chapter 11) categories of A, B, or C. The focus group was indicated by a positive value (e.g., C+), and the reference group was noted with a negative value (e.g., C-). The positive and negative values were reported for items with C DIF. DIF comparison was not done if the sample size for either group was less than 100 or if the combined sample size for the groups being compared was less than 400 (Smarter Balanced, 2017, p. 3–15.)

DIF was evaluated for eight subgroup comparisons, shown here with the focal groups listed first and the reference groups listed second.

- Gender: Female – Male
- Race/Ethnicity: Asian – White
- Race/Ethnicity: Black – White
- Race/Ethnicity: Hispanic – White
- Race/Ethnicity: Native American – White
- Individualized Education Program: Yes – No
- Limited English Proficiency: Yes – No
- Title 1: Yes – No

Items with C+ or C- DIF were flagged for data review.

Items that were not flagged for content or bias statistical issues were eligible for use in the operational pools. Flagged items became eligible for the operational pools if they were approved by a multidisciplinary panel of experts during data review.

3.6.2 Social Studies

For social studies, the following analyses were carried out for psychometric review (note that the listed analyses are routine annual procedures):

1. Content standard distribution check: This check is to ensure that operational (OP) items on each form have the desired content coverage (i.e., the reporting categories are the same as depicted in the test blueprint; and within each reporting category, the content standards have as much variety as possible.)

2. Item position check: Equating items and common items (i.e., non-equating items that appear on multiple forms) need to appear in the same test positions across forms. Moreover, equating items are checked to make sure they are within +/-2 position change from the previous year's positions.

3. Across year comparability check: For this check, distributions of item difficulty and item discrimination (*p*-values and adjusted item-total correlations, see Section 8.3.1.2 for details) are checked across adjacent years for unique items to make sure they are comparable. Moreover, when Item Response Theory (IRT) item difficulty and item discrimination (b-parameters and a-parameters) (see Section 6.2 and Equation 6-2 for details) are available for all OP items, test characteristic curves (TCCs), test information function (TIF) curves, and test standard error (TSE) curves are plotted to check the comparability across years.

4. Across mode comparability check: Comparability of OP items across modes (paper/pencil, online) is checked using the same approaches listed above in the across year comparability check.

5. Comparability of equating items and other OP items per form: Two analyses are involved in this comparison on each form: (1) content coverage homogeneity test (to make sure that equating items and other OP items have comparable content coverage) and (2) distributions of item difficulty and adjusted item-total correlation comparability check. These analyses are conducted to make sure that the equating items function as a mini-test (i.e., they are representative of the overall test, both statistically and in terms of content).

6. Item key distribution check: This check involves all multiple-choice (MC) items on the test (i.e., OP and field-test items). Here the desired result is for all four key options to appear relatively equally on each test map, with no same key option appearing three times consecutively. Although it is desirable to have unique field-test items on each form, if a field-test item must be repeated on multiple forms, a check is carried out to ensure that it appears in the same test position across forms.

7. Overall OP item set quality check: This check ensures that no OP items have problematic flags. Specifically, DIF results are checked to make sure that no OP items are with “B” or “C” DIF flags. All OP items that appear on the final form have been scrutinized to make sure that there are no bias or sensitivity issues involved. Moreover, adjusted item-total correlations, various item statistics flags (e.g., key option-total correlation being negative, distractor option-total correlation being positive, omit-total correlation being positive, key option percentage not being the highest), and IRT item parameters are also checked to see if items are free of concerns (i.e., adjusted item-total correlation should be ≥0.2, a-parameter should be >0, b-parameter should be in the range of [-3, +3], and there should be no item statistics flags).

All identified problems are documented and communicated to the corresponding content leads. Content leads then revise and resubmit test maps for another round of review. This iterative process continues until all issues have been resolved or the problematic item selections are proven to be the best selections given various constraints (e.g., content coverage considerations, and the need to avoid possible clueing).

3.7 Item Types Included

In addition to the traditional MC items, TE items were included in M-STEP. The following is a list of the TE item types used:

- Drag and Drop—Students drag pictures or words into boxes or “drop zones” to indicate an answer.
- Choice Interaction—This is similar to an MC item, but the item can have more than four options, and any number of the options can be correct.
- Hotspot (Count or Selection)—Students answer by selecting graphics, either a particular number of hotspots (Count) or a specific hotspot (Selection).
- Matching Interaction—Students select areas of an interaction grid to match options in rows and columns.
- Matching—Students make line connections between options from two sets.
- Keypad Input—Students use an embedded keyboard with mathematical functions to answer mathematics questions.
- Drop-Down—Students select options from a drop-down list.
- Hot Text Highlight (Line and Paragraph)—Text is selectable and, once selected, will become highlighted for the students. Students select one or more lines of text (Line) or words or sentences from a block of text (Paragraph).
- Order—Students answer by rearranging a list of items or sentences.
- Coordinate Graph Input—Students plot points, lines, and shapes on a coordinate grid.
- Number Line Graph—Students plot points on a number line.
- Text Input—Students enter values in a response box.
- Bar Graph—Students answer by selecting amounts to complete a bar graph.

Not all the TE item types are used in every content area.

3.8 Field-Test Selection and Administration

3.8.1 Field Test Item Selection

The OEAA content leads are tasked with selecting field-test items. The blueprints specify the number of field-test items by grade level and content area. The content leads work within Michigan’s IBS to monitor the number of operational items available for each content standard. Where there are gaps in the numbers available, content leads may decide to field-test items assessing that standard. The content leads also monitor the number of items that may be overexposed and need replacement items as one way to select field-test items.

Responses on field-test items do not contribute to a student’s score on the operational tests. The specific locations of the embedded items in the assessment are not disclosed. These data are free from the effects of differential student motivation that might characterize stand-alone field-test designs since the items are answered by students taking operational tests under standardized test administration procedures.

3.8.2 Field Test Administration

3.8.2.1 Mathematics and ELA

MDE-developed field-test items are embedded within the ELA and mathematics CAT assessments at all grade levels. The items are not designated as field-test items to the students, so the field-test items are not distinguishable from the operational items. This ensures that the students give the same effort to the field-test items as the operational items. All the students taking the CAT receive the same number of field-test items, and the selection and delivery of the field-test items are not affected by a student's performance on the test or the difficulty level of the field-test items. To avoid complications due to position placement, the field-test items are not distributed in the first five or the last five positions on the test.

For mathematics, the field-test items are placed at the same sequence positions throughout the CAT testing experience. For ELA, the field-test items are positioned in similar locations; however, due to the inclusion of passage sets in ELA, the field-test positions are shifted as necessary to accommodate a preceding passage set. Students receive either stand-alone field-test items or a field-test passage set containing associated items.

3.8.2.2 Social Studies

Social studies assessments consist entirely of MDE-developed operational and embedded field-test items for all grade levels.

The operational item set is the same across all online forms in a grade level, appearing in the same test positions. The remaining form positions are used for field-test items, which are unique to each form. The three online forms at each grade level are randomly administered to the student population.

The paper/pencil forms for social studies share all the equating items with the online forms. However, since TE items cannot be presented on paper/pencil forms, items in those positions are replaced by items assessing the same content standards and having similar item statistical profiles that were presentable on paper/pencil forms and in braille format.

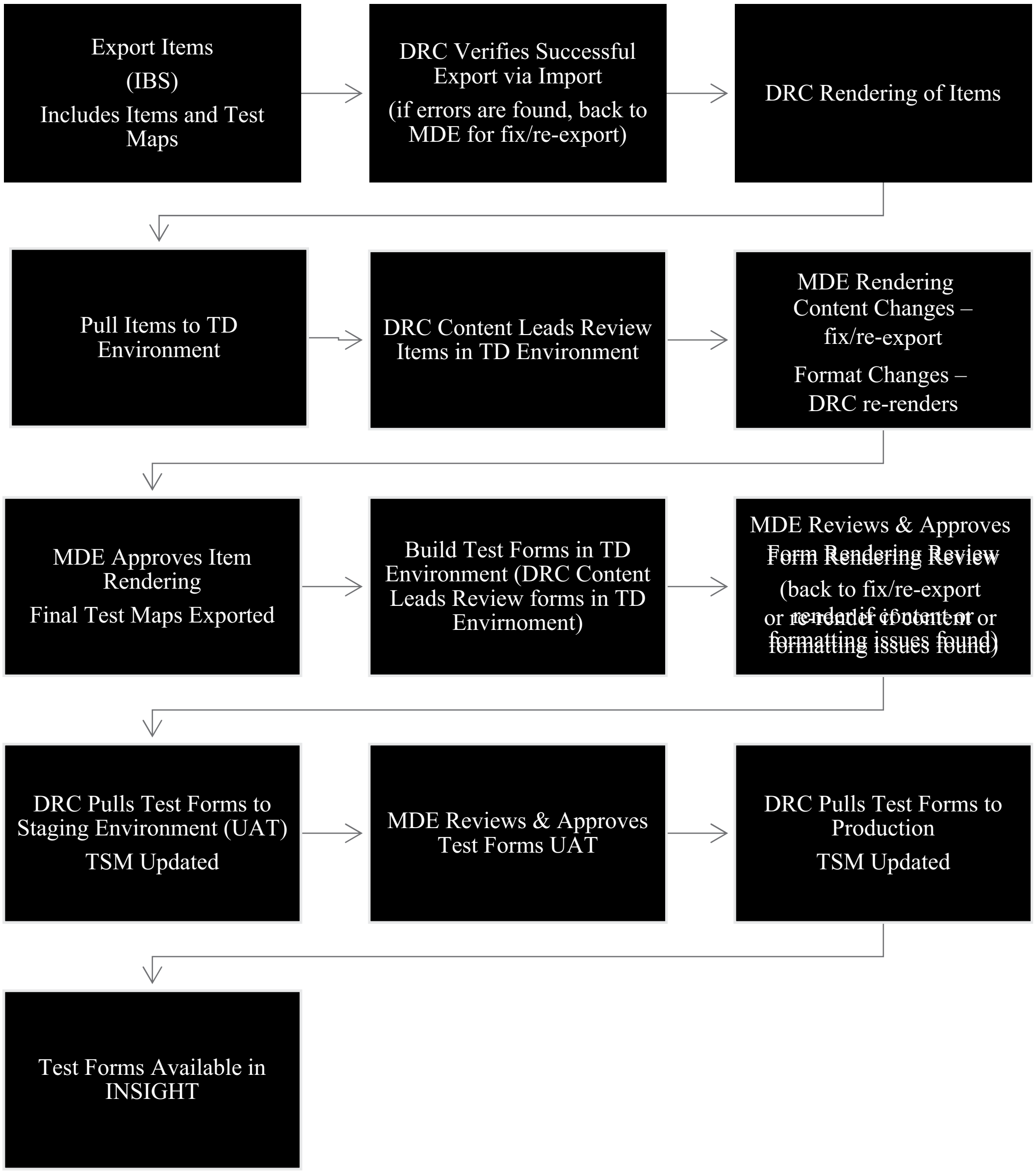
Details on constructing forms and follow in Sections 3.9 and 3.10.

3.9 Online Form Building and Rendering Process

3.9.1 Overview of Rendering Process

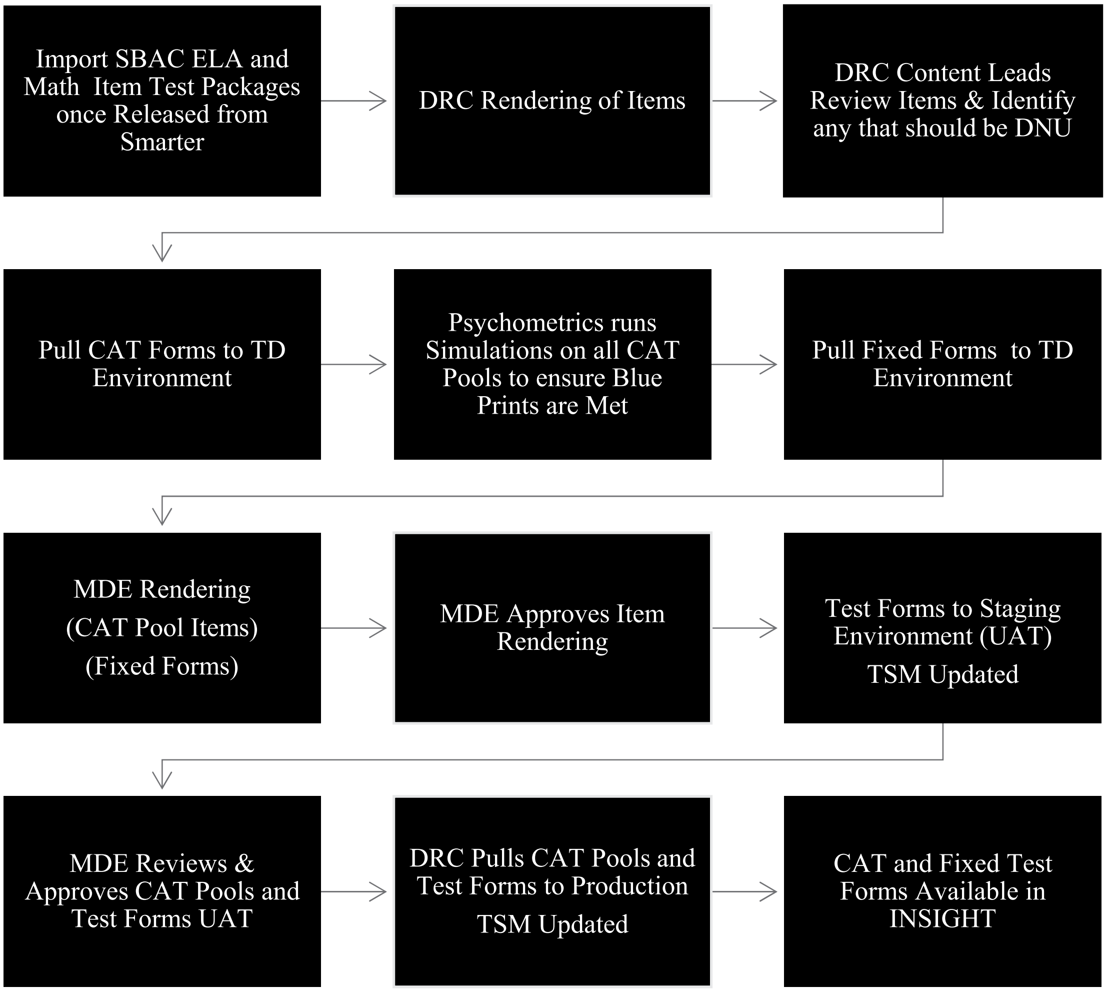
DRC and MDE follow a very rigorous rendering process for all items on the 2019 M-STEP. Using the web-based application LeanKit, DRC and MDE monitor the progress of each grade and content batch. The process begins right after the import of items from the Michigan and Smarter Balanced item banks. All parts of the rendering process are completed a month prior to the start of testing to ensure time for User Acceptance Testing (UAT) of all grades and contents. Figure 3-1 below shows the entire process for M-STEP field-test items and social studies items that are imported from the Michigan IBS.

Figure 3-1. Rendering Process of Michigan-Built Items


Export Items
(IBS)
Includes Items and Test Maps
DRC Verifies Successful Export via Import
(if errors are found, back to MDE for fix/re-export)
DRC Rendering of Items
Pull Items to TD Environment
DRC Content Leads Review Items in TD Environment
MDE Rendering Content Changes – fix/re-export
Format Changes – DRC re-renders
MDE Approves Item Rendering
Final Test Maps Exported
Build Test Forms in TD Environment (DRC Content Leads Review forms in TD Envirnoment)
MDE Reviews & Approves Form Rendering Review
(back to fix/re-export or re-render if content or formatting issues found)
DRC Pulls Test Forms to Staging Environment (UAT)
TSM Updated
MDE Reviews & Approves Test Forms UAT
DRC Pulls Test Forms to Production
TSM Updated
Test Forms Available in INSIGHT

The rendering process for the Smarter Balanced items is slightly different. Figure 3-2 shows the process followed for all items that are imported from Smarter Balanced to use for M-STEP ELA and mathematics.

Figure 3-2. Rendering Process of Smarter Balanced Items


Import SBAC ELA and Math Item Test Packages once Released from Smarter
DRC Rendering of Items
DRC Content Leads Review Items & Identify any that should be DNU
Pull CAT Forms to TD Environment
Psychometrics runs Simulations on all CAT Pools to ensure Blue Prints are Met
Pull Fixed Forms to TD Environment
MDE Rendering
(CAT Pool Items)
(Fixed Forms)
MDE Approves Item Rendering
Test Forms to Staging Environment (UAT)
TSM Updated
MDE Reviews & Approves CAT Pools and Test Forms UAT
DRC Pulls CAT Pools and Test Forms to Production
TSM Updated
CAT and Fixed Test Forms Available in INSIGHT

Requirements are established and reviewed with MDE prior to the imports of the 2019 M-STEP items. The requirements include the QTI 2.2 import specifications between the IBS and DRC’s IDEAS system as well as specific rules when importing each type of item. Detailed rendering requirements are also documented and reviewed.

3.9.2 Form Preparation and Rendering in INSIGHT

For all fixed forms, after the individual items are formatted and rendered, online test forms are assembled in the INSIGHT test engine based on the sequence and information provided in the test maps created by MDE. The test maps provide test-form data, item form sequence location, and metadata (e.g., content standard, DOK, item position, *p*-value, IRT parameters, answer key, points possible) for each test form for each test type (i.e., program, content, grade). DRC applies the appropriate styles and formatting to the fixed forms based on the previously set style and formatting guidelines.

The assembled fixed forms are then reviewed by content leads at DRC and MDE in a UAT setting to ensure that the forms match the exact design and data displayed in the test maps and that the forms, features, and functionality of INSIGHT appear and operate correctly. The UAT is conducted using the same INSIGHT test delivery system that the students use so the forms appear and function just as the students see them. The forms include features such as the online tools provided for each item, test directions, help files, calculators, and reference materials. Detailed information on student tools can be found in Chapter 4.

3.10 Paper/Pencil Form Building and Review Process

Although approximately 99% of Michigan students test online, there will always be paper/pencil forms available for those students who may not be able to test online and for student groups that require specific Accommodations or tests in other languages. Michigan offers the following Accommodations for students with disabilities and the following accessibility features for English learners delivered through paper/pencil assessments: enlarged print; braille; audio supports, such as reader scripts for teacher read-aloud Accommodations; audio CDs; and DVDs in Arabic and Spanish. The ELA and mathematics paper/pencil tests are provided by Smarter Balanced and align to Michigan's ELA and mathematics blueprints. OEAA's composition unit assembles the test booklets. There are several rounds of reviews conducted by OEAA content leads, OEAA assessment specialists, and OEAA's editor. Once the initial test booklets are approved, they are posted for printing by Measurement Incorporated, and the paper/pencil test maps are provided to Measurement Incorporated for use in creating braille and enlarged print forms using the American Printing House (APH) for the Blind.

The social studies paper/pencil tests are developed by OEAA's content leads using Michigan's IBS. They mirror their online counterparts with modifications, i.e., only TE items are replaced. The content leads review each item in the test map to check for text and/or graphic errors, clueing, correct answer keys, and a balance of answer keys. Once the test map is approved by the content lead, the psychometric lead reviews the test map in a similar way as mentioned above for online forms, but with more focus on comparability of paper/pencil forms to their online counterparts. Once the test maps are approved by both the content lead and the psychometric lead, the composition unit creates one item per page (i.e., "one-per") for review by both the OEAA content lead and the OEAA editor. A one-per is created for each item on the test map, showing how each item will appear in a test booklet. Content leads ensure the one-per matches the item as it is in the IBS, which is the source of truth. The item as it appears on the one-per must also follow OEAA's style guide and be free of errors. After the content lead approves the one-pers, they are reviewed by OEAA's editor. Once the editor approves the one-pers, test booklets are created. The draft printed test booklets are reviewed first by the editor and then by the content lead. Both the content leads and the editor use OEAA's Proofing Tools Guide and its task checklists to ensure each step is followed. Once the test booklet has final approval, the test maps and approved test booklets are sent to Measurement Incorporated for mass printing and accommodated format production of enlarged print, braille, reader scripts, audio CDs, and DVDs in Spanish and Arabic.

3.11 Summary

In summary, the overall purpose of this chapter is to explicate the procedures used in the development of M-STEP. The efforts by MDE and its vendors address multiple best practices of the test industry, particularly the following AERA, APA, and NCME (2014) *Standards*:

- Standard 3.1—Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
- Standard 3.2—Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.
- Standard 4.0—Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.
- Standard 4.1—Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).
- Standard 4.7—The procedures used to develop, review, and try out items and to select items from the item pool should be documented.
- Standard 4.12—Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

Chapter 4: Test Administration Plan

Chapter 4 reviews the test administration process for both the online and paper/pencil administrations of the M-STEP assessment. Detailed information on supports, accommodations, test materials, and training and test security practices can be found throughout this chapter. According to the AERA, APA, & NCME *Standards* (2014), "[t]he usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions" (p. 111). Chapter 4 examines how test administration procedures implemented for M-STEP strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations.

The online platform components of eDIRECT and INSIGHT, which were necessary for all online test administrations, are discussed in Section 4.4. The web-based application known as eDIRECT was used for all test preparation and test monitoring, while INSIGHT was the online test delivery system used by students when taking online assessments. More information on the online components can be found in Chapter 5.

4.1 Universal Tools, Designated Supports, and Accommodations

To allow all students the ability to fully demonstrate their knowledge and skills on the statewide assessments, a variety of tools are made available across all grades, content areas, and modes of testing. The variety of tools offered attempts to ensure that an equal opportunity for a student to demonstrate what he or she knows on a test is not negatively impacted by the student's disability or English language proficiency.

MDE categorizes tools into three levels: Universal Tools, Designated Supports, and Accommodations. Universal Tools can be used by students at their own discretion. Use of a Designated Support requires an educator identify that support type for a student because of an instructional need. Tools listed as Accommodations require that a student has an Individualized Education Program (IEP) or 504 plan and that the need to use that support is identified within that document.

Regardless of the level of the tool type, MDE requires educators to make decisions about use on an individual basis. The decision for use should be based on the individual student's instructional needs for each content area. Some tools may be classified as nonstandard, as described in the Supports and Accommodations documentation, in which case the use of those tools by students may result in invalid test scores. 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 and discussions with MDE's assessment content leads.

MDE's policies related to the use of Accommodations are in compliance with AERA, APA, and NCME (2014) Standard 6.2, which states the following:

> When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing (p. 115).

Additional information about Michigan’s accommodations framework and a list of which Universal Tools, Designated Supports, and Accommodations are considered allowable and valid for students to use can be found in the Supports and Accommodations Guidance Document.[1]

4.1.2.1 Educator Guidelines

Many of the allowable Designated Supports and Accommodations require educators to perform an action for the student or on behalf of the student. For example, a student needing a scribe may have one provided to them as long as the educator is using the scribing protocol outlined in MDE’s *Supports and Accommodations Guidance Document*. This manual collects guides
to ensure educators are providing these Designated Supports and Accommodations in a consistent and reliable manner. Additional guidelines include read-aloud guidelines for English and other languages.

4.1.2.2 Research Base for Supports and Accommodations

Smarter Balanced has published multiple literature reviews that support the use of MDE’s Universal Tools, Designated Supports, and Accommodations. Because MDE uses Smarter Balanced test content, the framework upon which the assessments have been built was based on the development efforts of Smarter Balanced. These Smarter Balanced Literature Reviews address research related to tools for students with disabilities and English learners.

4.1.2.3 Monitoring the Use of Designated Supports and Accommodations

MDE monitors Designated Supports and Accommodations used by students to ensure high reliability and validity of test results. Data audits include verification that students receiving Accommodations on the assessment had an Individualized Education Program or 504 plan. In the event that students received accommodations without an IEP or 504 plan, schools are contacted and asked to verify the use of Accommodations and make a plan to improve their process for future student use of Designated Supports and Accommodations. Interviews are conducted with schools after assessment monitoring to verify the decision-making processes used in providing Designated Supports and Accommodations to students for use on the assessment.

4.2 Online Accommodations

Appropriate Universal Tools, Designated Supports, and Accommodations were available for students to use while taking the assessment. Students with an IEP or 504 plan are required to have their assessment needs formalized in those documents prior to using any Universal Tools, Designated Supports, or Accommodations. Some embedded Designated Supports and the assessments with embedded Accommodations were delivered via fixed forms. Embedded refers to supports that were provided within the online delivery platform and non-embedded refers to supports that were provided externally from the online delivery platform. The Designated Supports and Accommodations that were embedded in the online fixed forms and used for the spring 2019 M-STEP were as follows.

- Audio Sign Language (applicable for ELA and Math) was available to students at grades 3–8. For ELA, Audio Sign Language video was available for only Listening passage stimuli and items.
- Stacked Translation (applicable for Spanish Math) was available to students at grades 3–8.
- Closed Captioning (applicable for ELA) was available to students at grades 3–8. ELA Closed Captioning was available for only Listening passages stimuli.

Embedded and non-embedded Designated Supports were also available and could be selected for each student via eDIRECT.

Embedded Designated Supports and embedded Accommodations were available within the CAT assessments and the fixed-form assessments. The available embedded Designated Supports and Accommodations are listed below.

- Text-to-Speech (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11. This Designated Support reads aloud items only. Text-to-Speech with Passages (applicable for ELA only) was available to students at grades 6–7 as an Accommodation. This embedded Accommodation reads aloud both items and passages.
- Masking (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11 (Designated Support).
- Color Choice (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11 (Designated Support).
- Contrasting Color (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11 (Designated Support).

In addition to the Designated Supports and Accommodations delivered by the test engine, there are a number of non-embedded Designated Supports and Accommodations available to students. The use of these non-embedded Designated Supports and Accommodations can be indicated in eDIRECT.

The non-embedded Designated Supports and Accommodations that are listed in eDIRECT can be found below. This is not a full list of allowable non-embedded Designated Supports and Accommodations but is only a list of what MDE considers the most frequently used non-embedded Designated Supports and Accommodations.

- Administered Individually/Small Group (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Noise Buffers (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Oral Translated Test Directions (applicable for mathematics) was available to students at grades 3–8.
- Read Aloud (Human Reader) (applicable for ELA and mathematics) was available to students at grades 3–8.
- Bilingual Word-to-Word Dictionary (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Auditory Amplification (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Visual Aids (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Scribe (non-writing items) (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Scribe (PBW prompt) (applicable for ELA) was available to students at grades 3–8.
- OEAA Multiplication Table (applicable for mathematics) was available to students at grades 4–7.
- Abacus (applicable for mathematics and social studies) was available to students at grades 3–8 and 11.
- Non-embedded Calculator (applicable for mathematics and social studies) was available to students at grades 6-7 in mathematics and 5, 8, and 11 in social studies.
- Administrator Sign Test Directions in ASL (applicable for ELA, mathematics, and social studies) was available to students at grades 3–8 and 11.
- Administrator Sign Test Content in ASL (applicable for social studies) was available to students at grades 5, 8, and 11.
- Alt Communication Devices (ACD) (applicable for ELA, mathematics, and social studies) was available to students at grades 3-8 and 11.

Table 4-1 below presents more details for DRC INSIGHT student tools. The following tools are available only on some fixed forms or in certain content areas.

Table 4-1. DRC INSIGHT Student Tools by Grade, Content Area, and Test Type

Assessment	Gr	Pointer	Crossoff	Highlighter	Magnifier	Line Guide	Sticky Notes	Ruler	Protractor	Calculator	Graphing Tool	Dictionary/ Thesaurus	Periodic Table	Help	Flag for Review	Pause	Writing Tools
ELA – CAT	3	x	x	x	x	x	x					x		x		x	x
ELA – CAT	4	x	x	x	x	x	x					x		x		x	x
ELA – CAT	5	x	x	x	x	x	x					x		x		x	x
ELA – CAT	6	x	x	x	x	x	x					x		x		x	x
ELA – CAT	7	x	x	x	x	x	x					x		x		x	x
Math – CAT	3	x	x	x	x	x	x							x		x	
Math – CAT	4	x	x	x	x	x	x		x					x		x	
Math – CAT	5	x	x	x	x	x	x							x		x	
Math – CAT	6	x	x	x	x	x	x			x				x		x	
Math – CAT	7	x	x	x	x	x	x			x				x		x	
Social Studies	5	x	x	x	x	x	x							x	x	x	
Social Studies	8	x	x	x	x	x	x							x	x	x	
Social Studies	11	x	x	x	x	x	x							x	x	x	

Figure 4-1 provides descriptions of system tools that help with navigation, may be Universal Tools, and some that are made available based on the item type and/or when a Designated Support or Accommodation is enabled.

Figure 4-1. DRC INSIGHT Student Tools Descriptions

TOOL	DESCRIPTION/FUNCTION
Navigation Tools	
Back Next	**Back and Next**—Move to the next question or a previous question. (**Back** is only available in CAT within passage and listening sets.)
Question 2 ▾	**Go To Question**—Jump to any item or passage set on the test by choosing the item from a drop-down list (only available in fixed forms).
Pause	**Pause**—Pause the test for a short period of time (e.g., restroom break) and resume upon return.
Flag	**Flag**—Mark a question for review at a later point (only available in fixed forms).
Test Review	**Test Review**—Review and change answers by section and indicate whether the test is ready to be scored (only available in fixed forms).
Standard Test-Taking Tools (available at all times)	
	Pointer—Select, change, or unselect an answer option; select other user tools; and navigate through the test. When moved over an answer choice, the pointer converts to a pencil image.
	Cross-Off Tool—Cross out an MC answer selection believed to be incorrect. This tool includes an eraser to remove the cross off if a student changes his or her mind.
	Highlighter—Highlight a portion of text or a graphic and remove highlights.
	Magnifier—Magnify/enlarge a portion of the screen (i.e., object, image, or text) by two times for better viewing.
Line Guide	**Line Guide**—Movable, straightedge line used to follow along with each line of text. Student can drag the guide up or down on the screen as an aid in reading an item or passage.
?	**Help**—The Help Library provides information on tool usage, test directions, helpful hints, and other topics. Also includes a “What’s This?” feature that allows a student to access contextual help for a specific tool or button.
	Sticky Note—Creates and places a small note in which a student can type a short message for later reference (multiple notes can be created for each item or passage).
	Calculator—Basic four-function and scientific options are available as required, either individually or together.
	Measurement Tools—Includes a **Protractor** for measuring angles that can be moved over any object on the screen and rotated.
	Graphing Tool—Used to graph one or several functions. Includes zoom and trace features.
Click to Respond	**Click to Respond**—Allows for placing various types of response areas in a snapshot view that a student expands to respond to the question. For example, a large graphing item can be placed in an item where it might not normally fit.
Enlarge	**Click to Enlarge**—Allows for large graphics by using a thumbnail image of the graphic that can be enlarged for viewing. Student can interact with the test item and other tools simultaneously.

TOOL	DESCRIPTION/FUNCTION
Accommodations Tools (determined at the student level)	
🎧 ⏹ ⏸ ▶	**Audio/Video tools**—Includes a **Text-to-Speech Synthesizer** that allows all test-related information (e.g., test directions, questions and answers, formula sheets) to be read aloud to the student. VSL fixed forms provide video for **sign language administration**.
Options	**Display Options**—Can be made available for all students or just those with a specific accommodation, such as **Color Overlays**, that allows a student to change the background color for text, graphics, and response areas.

4.3 Paper/Pencil Universal Tools, Designated Supports, and Accommodations

As noted earlier, OEAA provides a multitude of opportunities for students to demonstrate their knowledge on the M-STEP assessment with appropriate Universal Tools, Designated Supports, and Accommodations on the paper/pencil forms as well. Below is a list of available paper/pencil Designated Supports and Accommodations that require a specific form of the assessment.

- Braille, contracted and uncontracted for all content areas and grades
- Stacked Spanish, available for Mathematics in all grades
- Arabic, Spanish, and English DVD, available for Social Studies in all grades, to be used with form 1
- Reader Script, available for Social Studies in all grades, to be used with form 1
- English Audio CD, available for Social Studies in all grades, to be used with form 1

Referenced in Table 4-2 is the Designated Support and Accommodation information that is tracked (i.e., bubbled in) on each content area’s booklet. This is not a full list of allowable Universal Tools, Designated Supports, and Accommodations but is only a list of what MDE considers the most frequently used Designated Supports and Accommodations.

Table 4-2. Paper/Pencil Accommodations

Accommodation	ELA	Math	Social Studies
Directions Read in Native Language	✓	✓	
Oral Translation in Native Language		✓	✓
Spanish Booklet		✓	
Enlarged Print	✓	✓	✓
Multiple-Day Testing	✓	✓	✓
Audio CD			✓
English DVD			✓
Spanish DVD			✓
Arabic DVD			✓
Reader Script			✓
Alternate Response	✓	✓	
American Sign Language (ASL)	✓	✓	
Noise Buffers	✓	✓	
Read-Aloud (see Supports and Accommodations Guidance Document for specifics)	✓	✓	
Scribe	✓	✓	
Speech-to-Text	✓	✓	
Abacus		✓	
L1 Glossary		✓	
Other	✓	✓	✓
Nonstandard Accommodation/Support	✓	✓	✓

4.4 Online Test Platform

The secure web-based test engine DRC INSIGHT Online Learning System is loaded onto computers that students access for all online assessments. Test items and forms can only be accessed using a valid test ticket. INSIGHT automatic updates are suggested to be turned to “Enable” for the software to be automatically updated as needed. From the INSIGHT landing page, students have access to the test via the “Test Sign In” link and to the sample item sets via the “Online Tools Training” link.

DRC’s client portal, eDIRECT, is used to manage the test setup functions of student assessments and provide the downloads available for installation. The INSIGHT secure browser software is downloaded from eDIRECT and installed on student testing devices. The secure browser can be installed on computers individually, or it can be downloaded to a central location, copied, and simultaneously distributed to multiple computers using common network distribution tools. Everything needed for testing is found within the secure browser, eliminating the need for districts to coordinate updates to third-party software.

Technology coordinators install local caching servers (a testing site manager (TSM) or Central Office Services (COS) Service Device) to manage the content (test content and audio files) and regulate traffic between testing sites and DRC’s servers. The System Readiness Check helps troubleshoot any issues that may have occurred during or since INSIGHT installation. This application is installed when INSIGHT is installed, and the System Readiness Check performs a series of tests that can be used to diagnose and prevent or correct most errors.

The Load Simulation Tool is also available for sites to use for pre-planning purposes. The software is used by technology coordinators to perform load simulation tests that help estimate the amount of time needed to download tests and upload responses based on the number of students testing at the same time, the current network traffic, the amount of available bandwidth, and other site-specific factors.

The local caching software features Load Balancing, which allows the ability to monitor content caching availability. Load Balancing solutions also allow a district to quickly add or remove content servers when required without reconfiguring testing clients or redirecting or reassigning addresses. This tool also allows for an easier method to distribute testers between servers; each testing client is not dependent on a single server having enough capacity.

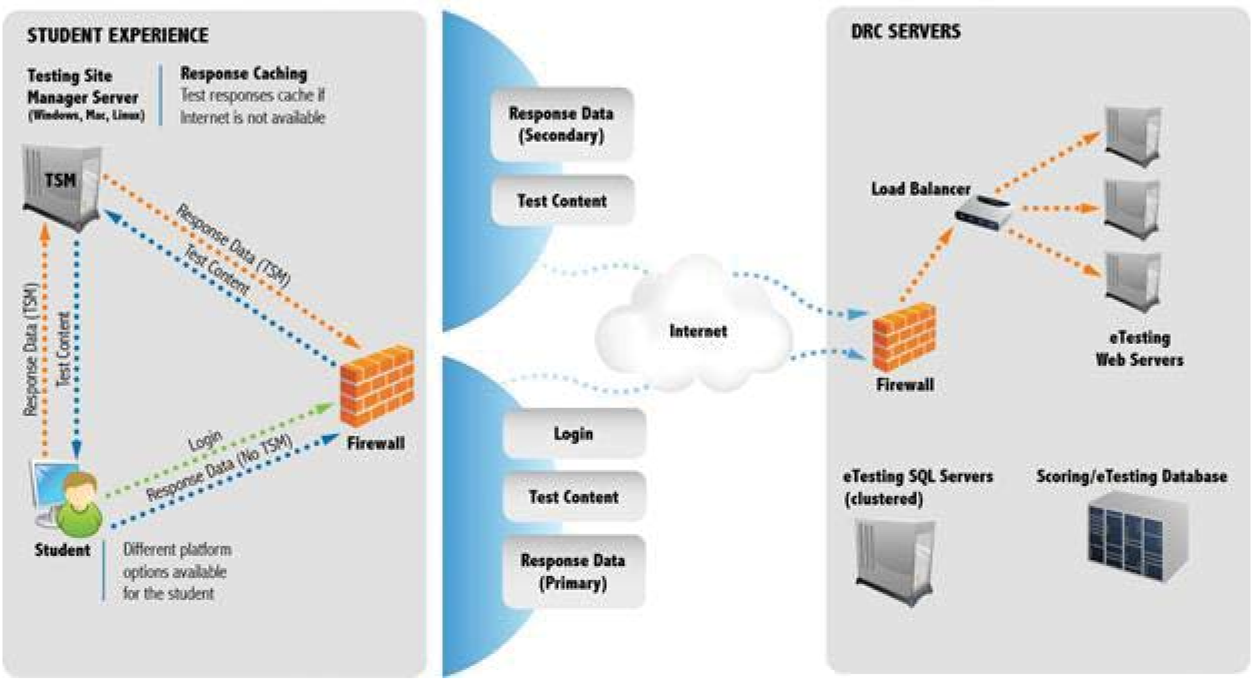
Prior to an assessment’s operational use, DRC’s quality assurance staff perform full system-level tests in an independent test environment that simulates the production configuration. Tests are run on all supported computer platforms and browsers and include a comprehensive review of system functionality, usability, reliability, security, and overall performance. Test content is also validated during this process.

Multiple methods are used to ensure secure data transfer, including encryption technologies and Secure Sockets Layer (SSL) protocol through Hypertext Transfer Protocol Secure (HTTPS). Test content is encrypted at the host server and remains encrypted throughout all network transmissions; content is decrypted only after the student login is validated. Decrypted test content on the student workstation is stored in memory only during each test session. After the session has ended (i.e., the test is completed, or the student logs out), computer memory is purged to ensure the security of test content.

During testing, responses are sent to a DRC server each time the student navigates away from an item or clicks the “Next” button to submit an answer. Responses are saved automatically every 45 seconds during testing, when the student navigates away from an item, or when the student answers a selected-response item, depending on whichever comes first. If an item takes the student longer than 45 seconds to answer, then the partial, incomplete response is submitted at 45-second intervals until the student completes the item. This autosave helps safeguard against students losing their work on longer items, such as Passage-Based Writing items. When the student returns to the test after a break or interruption, the student is returned to the point at which he or she left off without having to navigate through all previously answered questions.

Figure 4-2 illustrates the secure transfer of online test responses between the student and DRC.

Figure 4-2. Architecture of the Student Testing Experience



4.5 Test Administration Training

All staff involved in the administration of M-STEP are required to receive training based on the role they will serve during the test administration. Districts provide training for Building Assessment Coordinators, and District or Building Assessment Coordinators provide training for Test Administrators and Proctors. MDE provides test administration training resources for District Assessment Coordinators, Building Assessment Coordinators, and Test Administrators.

DRC, in conjunction with MDE, held a WebEx training presentation on March 5, 2019, with the District and Building Coordinators and Test Administrators. The presentation included pertinent information for all M-STEP online testing. The presentation was recorded and posted to eDIRECT for Michigan users to reference throughout the testing window.

MDE held a New Assessment Coordinator Preconference Workshop for both paper/pencil and online M-STEP administrations at the 2019 Michigan School Testing Conference on February 12, 2019. This presentation provided detailed information for new assessment coordinators administering both the paper/pencil assessment and the Online assessment. This training was structured into before-, during-, and after-testing activities and included the following:

- Before Testing
    - Universal Tools, Designated Supports, and Accommodations
    - Pre-identification of students
    - Materials ordering
    - Providing training to test administrators and proctors
    - Scratch paper and calculator policies
    - How to prepare students for testing (M-STEP tutorials, Online Tools Training (OTTs))
    - Off-Site testing requirements and requests
    - eDIRECT training
    - Test security and the Assessment Integrity Guide (AIG)
    - Test materials and handling of secure materials
    - Test schedules and test session setup
    - How to address a testing irregularity

- During Testing
    - Test directions
    - Testing irregularities
    - Active monitoring during testing
    - Materials allowed/not allowed in a test session

- After Testing
    - Materials return
    - Preliminary reports
    - Data files
    - Final reports

MDE also provided three webcasts with accompanying PowerPoint presentations organized into sections that discuss what administrators should do before, during, and after M-STEP administration. These webcasts followed the format used in the New Assessment Coordinator Workshop. These presentations are available on the MDE YouTube channel.[2]

Training materials are provided to districts to use for training purposes. These materials include the following:

- Guide to State Assessments
- M-STEP Test Administration Manual (TAM)
- Assessment Coordinator Training Guide
- Secure Site training and resource materials—provide training on pre-identification for testing, materials ordering, student scores and reporting, and using each function in the OEAA Secure Site (OEAA Secure Site Training)
- Test Directions—offered for each test mode (online and paper/pencil) and for each grade
- M-STEP List of Important Dates and Grade 8 List of Important Dates
- Supports and Accommodations Guidance Document
- eDIRECT mini-modules—provide training for all functions used in eDIRECT
- eDIRECT User Guide
- INSIGHT Tools poster—displays the tools available for students and describes how to use each tool
- Assessment Integrity Guide (AIG)
- Incident Reporting Guide (also included as an appendix to the Test Administration Manual)
- Scratch Paper Policy
- Calculator Policy

All these materials were available to schools during the 2018–2019 academic year on the M-STEP Home Page.

Additionally, OEAA publishes a weekly online newsletter called “Spotlight on Assessment and Accountability” throughout the year. The newsletter takes a two-week break in late December/early January, and no issues are published in July. The newsletter provides districts and schools with timely information regarding the M-STEP assessments and test administration, including training opportunities, document availability, and date reminders.

4.6 Test Security

The primary goal of test security is to protect the integrity of the assessment and to ensure that results are accurate and meaningful. OEAA uses four test security goals to maintain the integrity of the State of Michigan assessment system. These goals are

1. to provide secure assessments that result in valid and reliable scores,

2. to adhere to high professional test administration standards,

3. to maintain consistency across all testing occasions and sites, and
4. to protect the investment of resources, time, and energy.

4.6.1 Prevention

Prevention of breaches in test security includes following standards and best practices for test integrity and ensuring security aspects of the design, development, operation, and administration of M-STEP are met to prevent irregularities from occurring. Operational and administrative security policies and procedures apply to both online and paper/pencil test administrations.

Online testing uses DRC's INSIGHT Online Learning System. This is a secure browser that locks the student into the testing environment, preventing access to other applications or websites. The software must be installed on each device used for testing. Test content is held securely in an encrypted local caching server. All students are assigned to test sessions and need an individual test ticket for every online test session. Each ticket has a username and a unique password. Access to test tickets is controlled through DRC's eDIRECT site, and eDIRECT access is controlled through locally administered permissions in the OEAA Secure Site.

For the paper/pencil test administration, OEAA and Measurement Incorporated design forms to assist the district and building assessment coordinators with the successful receipt and return of test materials. These forms provide security and accountability during fulfillment and distribution, test administration, and collection processes. Secure packaging and distribution of materials for M-STEP are provided to ensure prompt, accurate, and secure delivery of test materials to districts and schools. All materials that contain test questions or student responses are considered secure materials and must be handled in a way that maintains their security before, during, and after testing. Handling of secure materials for paper/pencil and online testing is discussed at length in Chapter 5. As part of professional test administration practices, OEAA provides test security resources for state, district, and school personnel to use in the prevention of testing irregularities. These include the Assessment Integrity Guide (AIG), TAM, online and paper/pencil administration directions, test security training modules, and incident reporting guides.

All school staff members involved in testing are required to be trained in test administration and security prior to the opening of the assessment window. Training resources are available on a statewide basis. Districts and schools can customize trainings by role and location, using state-provided materials and including local plans. The AIG is intended to be used by districts and schools in the fair and appropriate administration of state assessments. It includes guidelines on the expected professional conduct of educators who administer state assessments to ensure proper test administration and academic integrity. Four assessment security training modules are available as a supplement to the AIG. The modules are intended to be used as an online training program for district and building assessment coordinators, test administrators, and test proctors. These modules explain why test security is important, describe different staff roles in test administration, and detail how to plan for and handle incidents that compromise test security. The M-STEP TAM helps staff administering the assessment understand the administration process, key dates for specific assessment activities, the roles of school personnel in the administration process, and the ways to use available Universal Tools, Designated Supports and Accommodations. Test administrators have online and paper/pencil test directions to follow when administering M-STEP. District assessment coordinators are required to file an incident report in the case of any testing irregularity. The incident reports are filed on the OEAA Secure Site. The test security specialist and other MDE assessment administrative staff review the incidents and determine what the required remediation will be through the use of internal and independent investigations.

4.6.2 Detection

Detection practices include guidelines for assessment monitoring, testing, and reporting irregularities. Detection resources and practices include the AIG, incident reporting, random/targeted test administration monitoring, administration observation, Universal Tools, Designated Supports and Accommodations monitoring, social media monitoring, and data forensic analysis. Districts are instructed to monitor test sessions for proper test administration and to enforce the policies and guidelines in the AIG to promote fair, approved, and standardized practices. OEAA uses random and targeted assessment monitoring to ensure the security and confidentiality of state assessments and to ensure testing personnel adhere to proper procedures. Targeted assessment monitoring is used when schools have had a previous irregularity or show unusual results from previous state assessment data analyses. Random assessment monitoring uses a sample of schools that are randomly selected for quality and integrity checks. Specific requirements of assessment monitoring are documented in the *Assessment Observation Requirements Document* created with OEAA's vendor Measurement Incorporated. The AIG details the process for monitoring district and school personnel. Internet and media monitoring occurs during testing windows. The goal of this monitoring is to combat breaches and disclosure of secure assessment materials. These monitoring activities include monitoring comments on the internet for test items captured and shared either from testing computer screens or from paper/pencil test booklets. Social media sites are also monitored for posts discussing or exposing test material. Requirements for social media monitoring are documented in the *Social Media Monitoring Requirements Document* created with OEAA's vendor Measurement Incorporated. The AIG details the process for monitoring the social media sites of district and school personnel.

DRC provides MDE with online forensic telemetry data via a secure table data load. The table below references the data that are captured and sent to MDE on a weekly basis during the testing windows.

Table 4-3. INSIGHT Forensic Data

Attribute of Forensic Data	Description
Test Interrupted Stopped Flag	Test was interrupted/stopped
Test Interrupted Stopped Count	Number of times the test was interrupted/stopped
Total Item Time	Total time spent on an item
Item Visit Count	Total number of times the item was visited
Wrong to Right	Item's response was changed from wrong to right (within or across item visits)
Wrong to Right Count	Total number of times the item's response was changed from wrong to right (within or across item visits)
Right to Wrong	Item's response was changed from right to wrong (within or across item visits).
Right to Wrong Count	Total number of times the item's response was changed from right to wrong (within or across item visits)
Wrong to Wrong	Item's response was changed from wrong to wrong (within or across item visits).
Wrong to Wrong Count	Total number of times the item's response was changed from wrong to wrong (within or across item visits)
Total Enters Net Total Exits	Records total enters are greater than or less than total exits.

During and after online and paper/pencil test administrations, OEAA conducts multiple analyses on student assessment results. These statistical analyses help in flagging potential testing irregularities. The types of data forensic analyses conducted in spring 2019 included unusual score gains and losses, online right-to-wrong changes, proficiency level gains, occurrence of perfect scores, and response time analysis.

4.6.3 Investigation and Remediation

District assessment coordinators are required to notify OEAA as soon as they are made aware of an alleged or suspected violation or misadministration of M-STEP. Testing irregularities are reported to OEAA via an online incident report form. The M-STEP TAM and AIG provide an incident reporting guide for districts and schools. Each testing irregularity report is reviewed by MDE test administration staff, and corrective action is taken based on policies and procedures for test administration outlined in the M-STEP TAM and AIG.

OEAA also has a phone and online "tip line" to report unethical behavior. Reports can be made anonymously. This provides a means for school staff members to report test integrity issues within their chain of command when they do not feel comfortable reporting the issues to their superior.

All incident reports and supporting documentation are reviewed by MDE, and a determination is made regarding the disposition of each incident. If OEAA determines that the irregularity caused no consequences affecting security, validity, or fraud and that the school took appropriate actions to correct the situation, OEAA may consider the issue resolved and log or close the case. If OEAA determines that questions remain regarding the security, validity, or authenticity of the test administration, OEAA will request either a school self-investigation or, if the problem is considered potentially severe, an independent investigation.

After investigations have taken place, OEAA will create a summary report of the findings. Determination of the investigation is provided in the report.

Remediation of the incidents reported and investigated differ based on the severity of a confirmed allegation or misadministration. Minor mistakes receive recommendations of best practices. Isolated security incidents or negligence provide good candidates for targeted monitoring the next year. Individual student tests tainted by misadministration are typically invalidated. More serious incidents can lead to invalidating entire classes of tests, retraining staff, or barring staff from participating in statewide testing. When possible, remediation happens within the testing window so that students with invalidated tests can be retested if appropriate. Further information on remediation is available in the AIG.

4.7 Summary of M-STEP Administration Security Best Practices

The elements discussed in previous sections align with MDE's prevention practices that help maintain the integrity of the assessment and also adhere to the testing practices and AERA, APA, & NCME (2014) *Standards* relevant to test administration. The previous sections also demonstrate how information in the MDE trainings and manuals addresses *Standards* 4.15 and 6.1:

Standard 4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented. (p. 90)

The M-STEP TAM, Test Directions, and AIG provide instructions for before-, during-, and after-testing activities with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the TAM, Test Directions, and AIG describe the following: general rules of online and paper/pencil testing; pause and break rules; test scheduling; assessment duration, timing, and sequencing information and recommendations; handling of secure materials; and materials that the test administrator and students need for testing.

Standard 6.1 Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (p. 114)

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it is essential that M-STEP is administered according to the directions provided in the TAM, Test Directions, and AIG.

MDE's protocol, discussed in Section 4.6, stresses incident reporting and adheres to *Standards* 6.3, 6.6, and 6.7.

Standard 6.3 Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (p. 115)

Incident reporting by district assessment coordinators is required when there is any type of misadministration or problem with test administration. MDE provides an Incident Reporting Guide within the TAM that details incident categories and subcategories that are used in the Secure Site Reporting tool and provides sample scenarios for each category or subcategory. MDE staff review the incident reports and respond with corrective action when appropriate. MDE also documents local variations on standardized test administration procedures such as alternate testing hours (for schools that have regular hours of instruction outside the usual 7:00 AM to 4:00 PM window) and locations (for schools with virtual students, students who cannot reach the school building, or off-premises testing sites).

Standard 6.6 Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (p. 116)

MDE requires that the testing environment for both online and paper/pencil testing be conducive to a proper test environment. All information regarding the content being measured or test-taking strategies displayed in the testing room must be removed or covered. Students must be seated so there is enough space between them to minimize opportunities to view each other's work. Test administrators and Proctors are encouraged to frequently move through the testing room and monitor the students' work areas during testing. Only staff involved in administering the test and students taking the test can be in the testing room. Students are not permitted to access any electronic devices used for communication, for capturing images of the test or testing room, or for data storage during testing. Testing materials are required to be kept secure at all times before, during, and after the testing sessions. Certain secure materials are required to be returned to Measurement Incorporated or securely destroyed.

Standard 6.7 Test users have the responsibility of protecting the security of test materials at all times. (p. 117)

The AIG and TAM describe the ethical practices that testing staff and students must follow during test administration. Students are reminded at the start of the testing session that in order for their results to be valid, they must not talk to or help other students; look at or copy other students' answers; ask for or accept any help from other students; use their cell phones or any other electronic devices, including an eBook; take pictures or make copies of any test materials; cause a disturbance; remove a test booklet, test ticket, or answer document from the room; or post or chat about any part of the test through social media. All staff who participate in a state assessment or handle secure assessment materials must be fully trained and sign an Assessment Security Compliance Form. By signing the Assessment Security Compliance Form, staff certify that they will follow test administration directions, maintain security and confidentiality of the tests, and report any suspected violations of test security.

Throughout the manuals, test coordinators and examiners are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted. Detailed information about test security procedures are presented in Section 4.6.

4.8 Test Materials

A list of available test materials can be found below in Table 4-4.

Table 4-4. M-STEP Paper Test Materials

Material Description	Product Type
Blank Labels	Ancillary
DVD Information Sheet	Ancillary
FedEx Return Air Bills	Ancillary
Instruction for Materials Return	Ancillary
OEAA Security Compliance Form	Ancillary
Outgoing Box Labels (M-STEP Materials Label)	Ancillary
Packing List Enclosed Label	Ancillary
PreID Labels	Ancillary
Return Kit Cover Sheet	Ancillary
Scorable Labels	Ancillary
Special Handling Envelopes	Ancillary
ELA Answer Document	Answer Document
ELA Emergency Answer Document	Answer Document
Mathematics Answer Document	Answer Document
Mathematics Emergency Answer Document	Answer Document
Social Studies Answer Document	Answer Document
ELA AABB	Braille
ELA Braille—Contracted Test Booklet	Braille
ELA Braille—Uncontracted Test Booklet	Braille
ELA Braille—Uncontracted Print to Braille Correspondence Document	Braille
Mathematics AABB	Braille
Mathematics Braille—Contracted Test Booklet	Braille
Mathematics Braille—Uncontracted Test Booklet	Braille
Mathematics Braille—Uncontracted Print to Braille Correspondence Document	Braille
Social Studies AABB	Braille
Social Studies-Contracted Braille Test Booklet	Braille
Social Studies-Uncontracted Braille Test Booklet	Braille
Social Studies-—Uncontracted Print to Braille Correspondence Document	Braille
ELA Listening Audio CD	CD
Social Studies Audio CD	CD
Social Studies Arabic DVD	DVD
Social Studies English DVD	DVD

Material Description	Product Type
Social Studies Spanish DVD	DVD
ELA Enlarged Print Test Booklet	Enlarged Print
Mathematics Enlarged Print Test Booklet	Enlarged Print
Social Studies Enlarged Print Test Booklet	Enlarged Print
Glossary Reference Sheets	Glossary
Graph Paper	Graph Paper
ELA Listening Script	Listening Script
ELA Listening Script, Emergency	Listening Script
M-STEP Test Administration Manual	Manual
M-STEP Paper/Pencil Test Directions	Manual
M-STEP Online Test Directions	Manual
M-STEP Emergency Test Administration Directions Addendum	Manual
Social Studies Emergency Reader Script (English)	Reader Script
Social Studies Reader Script (English)	Reader Script
ELA Emergency Test Booklet	Test Booklet
ELA Test Booklet	Test Booklet
Mathematics Emergency Test Booklet	Test Booklet
Mathematics Spanish Test Booklet	Test Booklet
Mathematics Test Booklet	Test Booklet
Social Studies Test Booklet	Test Booklet
Social Studies Emergency Test Booklet	Test Booklet
Math Test Booklet	Test Booklet
Social Studies Test Booklet	Test Booklet
Social Studies Emergency Test Booklet	Test Booklet

4.9 Summary

In summary, the overall purpose of the test administration documentation and training opportunities is to keep districts informed about policies and procedures related to testing in general and the M-STEP program. The information imparted is clearly related to maintaining the integrity of the administration of M-STEP, maintaining the security of the assessment, allowing access to the assessments for special populations by clearly delineating appropriate Universal Tools, Designated Supports or Accommodations, and providing guidance on appropriate interpretations of the test results. These communication and training efforts by MDE and its test vendors are in alignment with multiple best practices of the testing industry, particularly the following standards (AERA, APA, & NCME, 2014):

- Standard 4.15—The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.
- Standard 6.1—Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.
- Standard 6.2—When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.
- Standard 6.3—Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user.
- Standard 6.6—Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.
- Standard 6.7—Test users have the responsibility of protecting the security of test materials at all times.

Chapter 5: Test Delivery and Administration

5.1 Online Administration Details

MDE delivered 99% of M-STEP online via DRC's online testing platform, INSIGHT, in Spring 2019. 858 Michigan school districts administered M-STEP online to 3,151 Michigan schools.

For the fourth consecutive administration, M-STEP ELA and mathematics were administered as computer adaptive tests (CATs). M-STEP social studies was administered as a fixed form for a fifth consecutive year. Additionally, accommodated forms in all content areas were delivered as fixed-form assessments.

The Spring 2019 M-STEP was administered to enrolled students in grades 3–8 and 11, along with grade 12 students who missed testing in 2018. Table 5-1 presents content areas tested by grade.

Table 5-1. Content Areas Tested by Grade

Grade Tested	Content Areas Tested
Grade 3	ELA and Mathematics
Grade 4	ELA and Mathematics
Grade 5	ELA, Mathematics, and Social Studies
Grade 6	ELA and Mathematics
Grade 7	ELA and Mathematics
Grade 8	Social Studies
Grade 11 & 12	Social Studies

The number of students tested online for the spring 2019 M-STEP can be found in Table 5-2 below. Note that counts may vary across chapters due to differing definitions or inclusion rules. For example, a student could have an incomplete or invalid test, which might be included for some purposes but excluded for others.

Table 5-2. Number of Students Tested Online

Grade	Subject	Online Students Tested
3	ELA	100,234
4	ELA	101,739
5	ELA	104,443
6	ELA	108,407
7	ELA	108,654
3	Mathematics	100,425
4	Mathematics	101,977
5	Mathematics	104,590

Grade	Subject	Online Students Tested
6	Mathematics	108,539
7	Mathematics	108,737
5	Social Studies	104,379
8	Social Studies	107,415
11 & 12	Social Studies	104,928

5.1.1 Online Administration Reports

MDE and DRC defined requirements for all online administration reporting prior to administering the 2019 assessments. Administration reports were delivered to MDE daily or weekly based on the established requirements. Table 5-3 shows the types of administration reports that were delivered to MDE during the 2019 M-STEP testing windows.

Table 5-3. Online Administration Reports

Report Name	Delivery Frequency	Description of Report
Form Assignment Report	Weekly throughout the testing window	Shows fixed-form assignments for monitoring equal distribution of fixed forms per grade and content area
Cumulative Student Status	Daily throughout the testing window	Status of student testing by site; allows MDE to monitor how students are progressing with testing by grade and content area
Excessive Logins Report	Daily throughout the testing window	Shows online tests that have been logged into more than four times
Accommodations and Supports Report	Daily throughout the testing window	Shows number of students using each accommodation/support by grade and content area
Test Sessions - Content Area Compare	Daily starting in March through end of testing window	Shows number of students assigned to the same content area in both M-STEP and MI-Access administrations

5.1.2 Online User Manuals and Reference Documents

To help assist with the administration of the online M-STEP, numerous manuals and documents were created. Some of these include the test administration manuals, test directions by grade and test mode, the Technology User Guide, and many additional reference documents.

The M-STEP Test Administration Manual (TAM) is available for all test modes, grades, and content areas of M-STEP tests. It provides an overview of the assessments; important testing dates; information on when and how to assign and use Universal Tools, Designated Supports, and Accommodations; guidelines on who must test; testing policies and procedures including scratch paper and calculator policies; and resources for assessment coordinators and administrators.

The TAM provides detailed information regarding the roles involved in administering a test and responsibilities for each role: District Assessment Coordinator, Building Assessment Coordinator, and Test Administrator.

Information provided in the M-STEP TAM includes the following:

- Testing Schedules
- Overview
  - M-STEP Assessments
  - What's New
  - Content Area- and Grade-Specific Sections
  - Scratch Paper Guidelines
  - Supports and Accommodations
  - Resources for Students to Prepare for Testing
  - Valid, Equitable, and Ethical Assessment
  - Call Center Contact Information
  - OEAA Communications with Schools and Districts
  - Valid, Equitable, and Ethical Assessment
  - Assessment System Access

- Roles and Responsibilities
  - District Coordinators
  - Building Coordinators
  - Test Administrators
  - Technology Coordinators

- Supports and Accommodations
  - What Are Supports and Accommodations?
  - Supports and Accommodations Tracking Sheet
  - Ordering Accommodated Materials
  - Embedded and Non-embedded Supports and Accommodations
  - Turning on Supports and Accommodations
  - Verifying Test Tickets
  - Where to Find More Information on Supports and Accommodations
  - Filling Out Supports and Accommodations Information on Answer Documents
  - Using Accommodated Versions of the Tests
  - Read-Aloud Guidelines
  - Scribing Protocol
  - English Learner Supports
  - Returning Accommodated Materials and Answer Documents

- Student Pre-ID and Test Eligibility
  - Students to be Tested
  - Student Populations
  - Student Grade Considerations
  - Students in Different Locations
  - Students in Unique Circumstances

  - Accountability Considerations
  - Unique Identification Codes
  - Test Administration Windows
  - Computer Adaptive Testing for ELA and Mathematics
  - INSIGHT Availability
  - Incident Reporting
  - Testing Irregularities

- Before Testing
  - OEAA Assessment Security Compliance Forms
  - Training Tools
  - Security
  - Materials Permitted or Required During Testing
  - Software and Testing Device
  - Important Tasks Before testing

- During Testing
  - Important Tasks During Testing

- After Testing
  - Important Tasks After Testing

- Appendices:
  - Calculator Policy
  - Scratch Paper Policy
  - Incident Reporting Guide
  - eDIRECT User Guide
  - List of Important Dates
  - Administration Resources
  - Checklists
  - Change Log

Online Test Directions documents are provided for each grade level. These documents provide information for testing including materials needed during testing, items permitted in testing rooms, test scratch paper and calculator policy information, and test directions for each content area test in the grade level.

Information provided in the Online Test Directions for each grade includes the following:

- Introduction
  - Key
  - Online Tools Training (OTT) and Student Tutorials

- Before Testing
  - Test Materials Needed for M-STEP
  - Before Testing Checklist

- During Testing
  - Permitted Items in Testing Room
  - Procedures for Testing Breaks, Interruptions, or Pauses
  - Test Directions – Introduction
    - Test Sign-In
    - Welcome Screen
    - System Check and Test Security
    - Introduction
    - Answering Questions
  - Test Directions by grade and content area, including directions for accommodated assessments
  - Monitoring During Testing
  - Testing Irregularities
- After Testing
  - Completing the Test Session
  - Exiting the Test Engine

5.2 Paper/Pencil Administration Details

MDE delivered paper/pencil assessments to meet individual students' needs and for buildings that applied and were approved for a waiver of online testing.

Online testing waivers were available for the following reasons:

- Buildings that were not technologically ready
- Buildings that were under construction or otherwise had a disrupted technological environment
- Center-based programs
- Juvenile justice facilities
- Buildings that do not use technology in instruction

In addition to the standard version, the paper/pencil test was available in Enlarged Print and in both contracted and uncontracted braille versions. A Spanish language paper/pencil test was also available for mathematics in each grade. Enlarged print and Spanish versions were based on form 1.

There were three forms for each test, including the braille form. These forms are listed in the table below.

Table 5-4. Paper/Pencil Test Forms by Content Area

Content Area	Paper Pencil Forms Available
ELA	Form 1—administered to all students testing paper/pencil
ELA	Form 2—Emergency form
ELA	Form 88—Braille form
Mathematics	Form 1—administered to all students testing paper/pencil
Mathematics	Form 2—Emergency form
Mathematics	Form 88—Braille form
Social Studies	Form 1—administered to all students testing paper/pencil
Social Studies	Form 2—Emergency form
Social Studies	Form 88—Braille form

The paper/pencil tests were provided for the same grades and content areas that had online counterparts (see Table 5-1).

The M-STEP Test Administration Manual (TAM) is common for all test modes, grades, and content area M-STEP tests. It provides an overview of the assessments, important testing dates, information on when and how to assign and use Universal Tools, Designated Supports and Accommodations, guidelines on who must test, testing policies and procedures including scratch paper and calculator policies, and resources for assessment coordinators and administrators. See Section 5.1.2 for information about the content included in the M-STEP TAM.

Paper/Pencil Test Directions documents are provided for each grade level. These documents provide information for testing including materials needed during testing, items permitted in testing rooms, test scratch paper and calculator policy information, and test directions for each content area test in the grade level.

Information provided in the Paper/Pencil Test Administrations Directions for each grade includes the following:

- Paper/Pencil Test Schedule
- Introduction
  - Ensuring Test Security
  - Verifying Student Information
  - Barcode Label Directions
- Student Data Grid Information and Administration Directions
  - Directions for Completing the Student Demographic Page
  - Administration Directions for Completing the Student Data Grid
- Test Directions by Content Area for Each Grade
  - Student Participation
  - Preparation for the Assessment
  - General Rules for the Paper/Pencil Assessment
  - Testing Time Estimates
- After Testing
  - Assemble Materials for Return
  - Checklist for Test Administrators

The number of students tested using the spring 2019 paper/pencil M-STEP can be found in Table 5-5, below.

Table 5-5. Number of Students Tested with Paper/Pencil

Grade	Content Area	Number of Students Tested with Paper/Pencil
3	ELA	559
4	ELA	588
5	ELA	635
6	ELA	541
7	ELA	321
3	Mathematics	594
4	Mathematics	625
5	Mathematics	682
6	Mathematics	569
7	Mathematics	335
5	Social Studies	737
8	Social Studies	396
11 & 12	Social Studies	364

5.3 OEAA Secure Site

The OEAA Secure Site is a web-based application used for state assessments and accountability. The functions of the Secure Site include pre-identification of students for both paper/pencil and online assessments; ordering paper/pencil materials (see Chapter 4.8), including accommodated versions of the assessments; incident reporting; review of accountable students and test verification; and retrieval of data score files and score reports.

The Secure Site is available to authorized district and school personnel only. The MDE Secure Site training page[1] includes a complete list of Secure Site functions and how to use them.

The Secure Site takes student information from the Michigan Student Data System (MSDS) and provides a secure, centralized interface for using student information in testing. District-identified permissions for school staff carry through to vendor systems. To prepare for testing, students can be pre-identified for tests, moved between general and alternate assessments, arranged in online testing sessions, and have test materials ordered. During testing, the Secure Site provides tools for correcting issues missed during preparation and reporting incidents. After testing, the Secure Site provides access to student test score reports and data files, as well as providing tools for accountability and test verification.

[1] https://www.michigan.gov/securesitetraining

5.4 eDIRECT

5.4.1 Michigan Users

DRC uses MDE's Secure Site to pull and load Michigan users to eDIRECT based on Secure Site Test Cycle IDs. For the 2018–19 school year, the M-STEP *Test Cycle ID* was 175. Users were identified by their *Security Role ID* and pulled into eDIRECT according to the established requirements. The mapping of users from the Secure Site to eDIRECT can be found below in Table 5-6.

Table 5-6. Mapping of Building Users from Secure Site to eDIRECT

Security Role ID	eDIRECT Role and Permission Set
17—Public School Administrator	School
20—District Administrator	School
40—Public Online Test Administrator	School
31—Nonpublic School Administrator	School
41—Private School Online Test Administrator	School
42—District Test Administrator	School
45—State	State
38—District Technology Coordinator	District Technology Coordinator
39—School Technology	District Technology Coordinator
43—Public School Technology	District Technology Coordinator
44—Private School Technology	District Technology Coordinator

All users were identified by the site code(s) they had access to within eDIRECT. Users were only able to access student and test information based on their site permissions in the MDE Secure Site.

5.4.2 Administrative Functions

Online administration is managed through the DRC eDIRECT client portal that provides tiered, secure access to all required administrative functions. Within eDIRECT, users manage student information and create test sessions.

Student information for M-STEP is imported into eDIRECT via automatic loading of data. DRC utilizes the MDE Secure Site to pull new and updated student records for import into eDIRECT. Student data is pulled three times a day so that any new student records or updated student records are loaded in a timely manner. Building users can view all the demographic information associated with the students from the Secure Site before placing them in test sessions for test tickets.

Once the student data is loaded into the Test Setup application within eDIRECT, users organize students into test sessions. Test sessions can be created by content area, class, grade, or school. Through Test Setup, users can also update student Designated Supports and Accommodations, print test tickets, and monitor student testing status.

The student login ticket contains unique login credentials that students use to access the testing software. For a selected test session, assessment coordinators can download and print a PDF document containing instructions, a roster of student tickets, and the actual test tickets. Student test tickets are considered secure materials, and assessment coordinators keep printed tickets in a locked and secure location to be distributed to test administrators and collected daily for secure storage.

5.4.3 Online Testing Resources

eDIRECT houses an assortment of testing resources available to the district and school users as well as the technology coordinators. The INSIGHT installables and requirements are maintained on eDIRECT, as are all technology guides and information necessary for setting up schools' computers and servers.

Video tutorials containing mini-chapters on how to use eDIRECT applications are available to help users familiarize themselves with the different administrative applications within eDIRECT. An eDIRECT user guide is also available for reference.

Student-facing online testing resources include M-STEP tutorials for students and Online Tools Training (OTTs). OTTs are practice questions that demonstrate examples of each tool in the online system and each item format a student might encounter. OTTs are not complete practice tests.

For more information on M-STEP-specific online testing resources, visit the M-STEP website.[2]

5.5 Return Material Processing

Each box of materials shipped to schools contains a box list, which shows each item in the box. Each order contains a packing list, which shows a complete list of items, quantities, and box locations for the entire order. When an order contains secure materials, a security list is also included that shows a complete list of secure items and the associated shrink-wrapped pack barcodes.

All M-STEP scorable and non-scorable secure testing materials that are not destroyed after testing are to be returned via FedEx Express Saver to Measurement Incorporated to be processed.

When boxes of returned materials arrive at Measurement Incorporated, the warehouse team scans the boxes into the Measurement Incorporated tracking system database, where they are checked against the tracking numbers that are assigned to each school. FedEx also scans each of its tracking barcodes to record each box as it was delivered to Measurement Incorporated. This provides immediate information on the number of boxes received and points of origin of the boxes. Once this procedure is completed, the boxes are opened and all materials are sorted.

[2] http://www.michigan.gov/mstep

Scorable and non-scorable materials are securely scanned in using Measurement Incorporated’s Security Barcode Check-In Application. This application allows IT Operations to scan the security identifier on individual secure materials or the security identifier located on the outside of an intact pack of shrink-wrapped documents using Measurement Incorporated’s automated security scanning process. Scanning the security identifier on the shrink-wrapped pack is equivalent to scanning all the individual security identifiers included in the shrink-wrapped pack and is more efficient than scanning each individual test booklet in the shrink-wrapped pack.

As each security identifier is securely scanned, it is checked against the original list of identifiers that were entered into the Measurement Incorporated database. Any discrepancies are noted, and a security report is generated for MDE.

For scorable answer documents, the same scanning process that captured the security identifier information also captures information from the student pre-identification label, bubbled demographic information on the answer document cover, bubbled student responses, and images of constructed responses to be sent on to handscoring.

All loose (i.e., individual) test booklets are securely scanned into the Measurement Incorporated database by IT Operations using Measurement Incorporated’s automated security scanners.

Warehouse personnel securely scan in all returned accommodated materials using a human-operated computer station equipped with a barcode reader and enter those materials into the ObjectTracker database.

The accommodated materials include CDs, DVDs, braille test booklets, Enlarged Print test booklets, and Reader Scripts. ELA Listening CDs and Reader Scripts for M-STEP are also scanned in.

After all returned secure materials are checked in, Measurement Incorporated’s IT team prepares the initial security report data by comparing the security barcodes of checked-in materials with the barcodes of all secure materials.

The initial missing materials and security report data are provided to MDE in a spreadsheet. All schools that were sent materials by Measurement Incorporated are included in the summary, regardless of whether the schools are active or inactive entities.

For public school districts that are missing secure materials, district coordinators are shipped security reports to be further distributed to building coordinators.

For public school academies and nonpublic schools that are missing secure materials, each building coordinator is shipped a security report.

Missing materials previously reported by the school as destroyed or never received are not included on the security report sent to the district or school. Missing materials reported as lost remain on the security report, and the comment “Reported Lost” is added to the comment section of the security report.

FedEx Ground Package Returns Program labels are provided in case any secure materials need to be returned. Schools that find no additional secure materials are directed to return the summaries of missing secure materials and any additional information.

The Measurement Incorporated IT team updates the security report data using the spreadsheet of issues reported to the Call Center, which includes materials that were lost, destroyed, or never received. This spreadsheet is maintained by the Measurement Incorporated management team. MDE staff forwards to the Measurement Incorporated management team any information collected via phone calls or incident reports regarding materials that were lost, destroyed, or never received.

If a summary of missing secure materials is accompanied by a corresponding explanation letter, the two are stapled together. All summaries of missing secure materials are checked in using the district/building code barcode and are filed in order by assessment, district code, and building code. Any returned secure materials are checked in by security barcode and are stored with the other secure materials.

After the initial response window ends and the returned letters and secure materials are processed, the IT team refreshes the security report data for each assessment, indicating schools that responded with newly returned secure materials and/or letters and schools that did not respond. Follow-up security reports are generated.

A second round of cover letters and security reports is sent to districts and schools that still have outstanding missing materials and have not returned a letter or a security report with comments. This procedure is the same as the ones used for the first round of security reports. Schools that return a letter, materials, or both in the first round are not included in the second round.

Measurement Incorporated checks in and files any returned summaries of missing secure materials, secure materials, and additional information received. When MDE determines that schools have had sufficient time to respond, Measurement Incorporated generates and provides to MDE a final missing materials report.

The final security report spreadsheet sent from Measurement Incorporated to MDE includes all schools and districts that were tested. The Excel filter feature is used to list those that still have outstanding missing materials. The “Returned Letter or Additional Items or Both” column reflects letters and items returned in response to both the initial round and the second round of security reports.

Tables 5-7 through 5-9 show shipped M-STEP material information. The amount of material shipped was and should be expected to be higher than the number of students testing on paper/pencil. Each student testing on paper needs at least two secure materials (booklet and answer document), plus additional secure materials for accommodated testing.

Table 5-7. Count of Secure M-STEP Materials Shipped

Grade	ELA	Mathematics	Social Studies
3	2,076	2,017	N/A
4	2,162	2,066	N/A
5	2,991	2,602	1,653
6	2,261	2,059	N/A
7	1,795	1,620	N/A
8	N/A	N/A	1,147
11	N/A	N/A	1,621

Table 5-8. Count of Secure M-STEP Materials Returned

Grade	ELA	Mathematics	Social Studies
3	2,060	1,972	N/A
4	2,135	2,044	N/A
5	2,752	2,484	1,638
6	2,179	2,006	N/A
7	1,678	1,511	N/A
8	N/A	N/A	1,128
11	N/A	N/A	1,541

Table 5-9. Count of Secure M-STEP Materials Not Returned

Grade	ELA	Mathematics	Social Studies
3	16	45	N/A
4	27	22	N/A
5	239	118	15
6	82	53	N/A
7	117	109	N/A
8	N/A	N/A	19
11	N/A	N/A	80

5.6 Testing Window and Length of Assessment

Online testing was originally scheduled for four-week testing windows for each set of grades. Due to inclement weather, an additional week was added to the end of each testing window. The five-week testing windows for the online 2019 operational M-STEP were as follows:

- Grades 5 and 8 were administered the ELA, mathematics, and social studies assessments from April 8 through May 10, 2019.
- Grade 11 was administered the social studies assessment from April 8 through May 10, 2019.
- Grades 3, 4, 6, and 7 were administered the ELA and mathematics assessments from April 29 through May 31, 2019.

All online accommodated and standard assessments were administered in these time frames; there were no specific make-up windows for online assessments.

Paper/pencil testing dates for grades 5 and 8 were as follows:

- ELA Days 1 and 2: April 9 and 10, 2019
- Math: April 16, 2019
- Social Studies: April 17, 2019
- Makeup days:
  - ELA: April 11, 12, and 15, 2019
  - Any content area: April 19–26, 2019

Paper/pencil testing dates for grade 11 were as follows:

- Social Studies: April 11, 2019
- Makeup days: April 12–26, 2019

Paper/pencil testing dates for grades 3, 4, 6, and 7 were as follows:

- ELA Days 1 and 2: April 30 and May 1, 2019
- Math: May 7, 2019
- Makeup days:
  - ELA: May 2, 3, and 6, 2019
  - Any content area: May 8–17, 2019

The spring 2019 M-STEP was not timed and was paced by students. Schools scheduled test sessions and determined the appropriate amount of time for students to spend testing in a single test session. Any students needing more time were able to complete the test in a later test session during the five-week grade-level testing windows. Further information on test session timing is provided on pages 4–9 of the 2018–2019 Guide to State Assessments.

Chapter 6: Operational CAT

This chapter mainly covers elements of the CAT algorithm, including entry point, ability estimation and standard error of measurement (SEM), passage selection, test navigation, test termination, and forced submission. M-STEP CAT configurations and simulations for ELA and mathematics are reported toward the end of this chapter. Information on the Smarter Balanced Summative CAT configurations can be found in the *Smarter Balanced 2017–2018 Technical Report* (2018), and the 2017 Smarter Balanced Summative CAT simulations can be found here.[1]

Before a CAT administration, the configurations and the item pool need to be loaded into the CAT engine. The configurations define the operational test blueprint with different content rules (e.g., Min and Max number of items in one or more content standards and/or item types), field-test blueprint (e.g., number of items in each claim and item position), scoring algorithm (e.g., theta estimation method, scaling constants, highest obtainable scale score [HOSS], and lowest obtainable scale score [LOSS]), and passage-selection criteria (e.g., Min and Max number of items in a passage, passage Min percentage, distinct passage Min for ranking, passage ranking criteria, passage randomization options, and options to fulfill rules). The details of the configurations for each grade and content are presented after the descriptions of the processes. Note that specific information related to the psychometric background can be found in the *Smarter Balanced 2017–2018 Technical Report* (2018).

6.1 Entry Point

The M-STEP CAT algorithm for ELA and mathematics is designed to administer items targeted for an individual student based on his or her performance. However, students' performance is unknown at the beginning of the test. With no prior information about a student, DRC has determined, based on simulation studies prior to operational administrations, that using a starting point one standard deviation (SD) below the average item difficulty of the M-STEP ELA and mathematics CAT pools provides students with a better test-taking experience at the beginning of the test, particularly for those who are at the lower end of the achievement continuum. Table 6-1 lists the initial values used in the 2018–19 M-STEP CAT.

The M-STEP CAT algorithm includes a randomization component when selecting items to control item exposure. That is, one item is selected from among a set of items that is near the targeted item difficulty. This is especially important at the beginning of the test when no prior information is available. Randomization of items and rules defined by the test blueprint ensure that students will not see the same set of items in the same order even when all the students are assumed to perform the same or have the same initial theta at the beginning of the test.

[1] https://portal.smarterbalanced.org/library/en/2016-17-summative-assessments-simulation-results.pdf

Table 6-1. Initial Thetas for the CAT

Content	Grade	Initial Theta
ELA	3	-1.637
ELA	4	-1.258
ELA	5	-0.882
ELA	6	-0.420
ELA	7	-0.233
Mathematics	3	-1.991
Mathematics	4	-1.198
Mathematics	5	-0.544
Mathematics	6	-0.329
Mathematics	7	0.577

6.2 Theta Estimates and Standard Error of Measurement

After each item response, the theta estimate and SEM are calculated via the maximum likelihood estimation (MLE) for the total test and each claim. Note that only responses to autoscored items are accounted for in the theta estimate used by the CAT algorithm. The items in the item bank are calibrated based on the Generalized Partial Credit Model (GPCM) (Muraki, 1992) (see Equation 6-1).

$$P_{im}(\theta_j) = \frac{\exp[\sum_{k=0}^{m} Da_i(\theta_j - b_{ik})]}{\sum_{v=0}^{M_i-1} \exp[\sum_{k=0}^{v} Da_i(\theta_j - b_{ik})]}, \text{ (6-1)}$$

where $a_i(\theta_j - b_{i0}) \equiv 0$; $P_{im}(\theta_j)$ is the probability of an examinee with ability θ_j getting score m on item i; M_i is the number of score categories of item i with possible scores as consecutive integers from 0 to $M_i - 1$; D is the scaling constant, 1.7; a_i is the discrimination parameter of item i; b_{ik} is the location parameter or threshold of category k. The GPCM is equivalent to the 2 Parameter Logistic (2PL) Model (Birnbaum, 1968) (see Equation 6-2) when the item is scored dichotomously.

$$P_i(\theta_j) = \frac{1}{1 + \exp[-Da_i(\theta_j - b_i)]}, \text{ (6-2)}$$

where $P_i(\theta_j)$ is the probability of an examinee with ability θ_j answering item i correctly; D is the scaling constant, 1.7; a_i and b_i are the discrimination and difficulty parameters of item i.

For a general MLE, the likelihood combines both dichotomously and polytomously scored items as shown below:

$$\mathrm{L}(\theta_j \mid \mathrm{U}) = \left(\prod_{i=1}^{n} P_i(\theta_j)^{u_i} Q_i(\theta_j)^{1-u_i} \right) \cdot \left(\prod_{i=n+1}^{N} \prod_{m=0}^{M_i-1} P_{im}(\theta_j)^{u_{im}} \right), \quad (6\text{-}3)$$

where $Q_i(\theta_j)$ is $1 - P_i(\theta_j)$ and the response matrix U contains the response of dichotomously scored items

$$u_i = \begin{cases} 1, if\ correct, \\ 0, otherwise, \end{cases}$$

for $i = 1,\ldots,$ n, and the responses of polytomously scored items

$$u_{im} = \begin{cases} 1, & if\ scored\ m, \\ 0, & otherwise, \end{cases}$$

for $i = n + 1$, …, N and m = 0, 1, …, $M_i - 1$.

The modified version of the Newton-Raphson equation used by DRC for estimating theta at iteration t is given as below:

$$[\hat{\theta}]_t = [\hat{\theta}]_{t-1} + \frac{L_1' + L_2'}{ABS(L_1'' + L_2'')}. \quad (6\text{-}4)$$

where *ABS* stands for the absolute value. L_1' and L_1'' are the first and second derivatives of the likelihood function of polytomously scored items:

$$L_1' = \sum_{i=1}^{m} Da_i(u_i - p_i) \quad \text{and (6-5)}$$

$$L_1'' = \sum_{i=1}^{m} \frac{D^2 a_i{}^2 (-p_i{}^2)(1 - p_i)}{p_i}, \quad (6\text{-}6)$$

where is u_i is the score a student gets from a dichotomously scored item and the possible values are 1 or 0. L_2' and L_2'' are the first and second derivative of the likelihood function of polytomously scored items:

$$L_2' = \sum_{i=n+1}^{N} Da_i \sum_{m=0}^{M_i-1} u_{im} \left(m - \sum_{m=0}^{M_i-1} mP_{im}(\theta_j) \right) \quad \text{and (6-7)}$$

$$L_2'' = -\sum_{i=n+1}^{N} D^2 a_i^2 \left[\sum_{m=0}^{M_i-1} m^2 P_{im}(\theta_j) - \left(\sum_{m=0}^{M_i-1} m \cdot P_{im}(\theta_j) \right)^2 \right], \quad (6\text{-}8)$$

where u_{im} is the value 1 or 0.

During the M-STEP CAT administration process, in the case of scores that are zero (i.e., all items are incorrect) and perfect (i.e., all items are correct), a correction factor is applied before computing the relevant MLEs because the corresponding thetas cannot be estimated. The correction factor can be configured as any fractional value between 0 and 1 (e.g., 0.3). However, for the final scoring, the LOSS and the HOSS are assigned to the "all incorrect" and "all correct" cases according to the scoring specifications.

For each theta estimate, the corresponding SEM is calculated. SEM is the inverse of the square root of the test information function (TIF), which is the sum of the item information functions (IIFs). The IIF for dichotomously and polytomously scored items can be calculated by using the following equations, respectively:

$IIF_i = D^2 {a_i}^2 (1 - P_i) P_i$ (6-9)

and

$$IIF_i = D^2 {a_i}^2 [{}_{M_i - 1 \; M_i - 1} \sum_{m=0}^{M_i - 1 \; M_i - 1} m^2 P_{im} - (\sum_{m=0}^{M_i - 1 \; M_i - 1} m P_{im})^2] \text{ (6-10)}$$

6.3 Item Selection

After the initial item set is administered, the M-STEP CAT algorithm is designed to administer items targeted at an individual student's current performance, given content coverage boundaries. Specifically, the M-STEP CAT algorithm makes selection decisions each time based on the interim theta estimates while also taking many other factors, including test blueprint, item information function, and/or passage-related factors, into consideration. The details related to these factors are discussed below.

6.3.1 Test Blueprint

The adaptive item selection algorithm is designed to cover a standards-based blueprint, which includes the content standards, DOKs, item types, and score-point constraints. The M-STEP CAT algorithm closely resembles a modified constrained CAT (MCCAT) design (Leung, Chang, & Hau, 2003). The general idea is that the CAT algorithm is configured with upper and lower bounds that specify the minimum and maximum numbers of items that will be administered to students at the total-test, claim, content-category, assessment-target, and/or item-type levels. For the set of items configured, further configurations can be set up so only items at the specified DOK level and/or score point will be selected. The configurations specified in the test blueprint can be prioritized to ensure that the blueprint is met for each administration.

6.3.2 Item Information Function

After a content rule, among the content rules with the same priority level, is selected randomly or by the highest need, the M-STEP CAT algorithm targets the top N-ranked items, which are configurable, with the higher information function at the theta estimate. In general, the most efficient way to run an M-STEP CAT is to select items with the highest information function, which contains the smallest standard error for any given number of items. However, the consequence is that the items with high discriminations tend to be used more frequently. At the beginning of the test, it may not be necessary to select an item with the highest information function because the theta estimate used for calculating the information function contains a large measurement error. To control the item exposure rate for the high-discriminating items in the bank, a randomization process is introduced. Instead of the item with the highest information function being selected, the item to be used next is randomly selected from the top N (e.g., N = 5) number of items ranked by the item information (given interim theta estimate) and content-related criteria (see the “Distinct Top-Ranked Passage #” column in Table 6-2). Table 6-3 provides the scaling constants (i.e., slope A and intercept B), LOSS, and HOSS. All are fixed for each grade and content, and all were finalized before the test administration. They are used to convert students’ estimated scores to the scale scores. More information about the scale transformations can be found in Chapter 8.

Table 6-2. Passage Selection Criteria

Content Area	Grade	Item Range	Passage Min #	Passage Max #	Passage Min %	Distinct Top-Ranked Passage #	Percentage of Items Used (% of Weight)	Items Delivered (% of Weight)	Max Information (% of Weight)
ELA	3–7	1–4	1	1	100	10	100	0	0
ELA	3–7	5–6	2	3	60	5	0	100	0
ELA	3–7	7	1	1	100	4	100	0	0
ELA	3–7	8–9	3	4	100	5	0	50	50
ELA	3–7	>=10	1	4	50	5	0	40	60
Mathematics	3–6	1–3	1	1	100	15	0	100	0
Mathematics	3–6	>=4	1	1	66	10	0	0	100
Mathematics	7	1–2	1	1	100	15	0	100	0
Mathematics	7	>=3	1	1	66	10	0	0	100

Table 6-3. Scoring Algorithm

Content	Grade	Slope A	Intercept B	LOSS	HOSS
ELA	3	26.0061	1322.5934	1203	1357
ELA	4	24.6036	1409.5875	1301	1454
ELA	5	25.8718	1501.3628	1409	1560
ELA	6	24.5491	1592.9699	1508	1655
ELA	7	23.8151	1687.3543	1618	1753
Mathematics	3	26.3725	1325.7407	1217	1361
Mathematics	4	25.2608	1409.0233	1310	1455
Mathematics	5	23.3374	1495.6493	1409	1550
Mathematics	6	20.4573	1589.9260	1518	1650
Mathematics	7	19.6292	1686.6036	1621	1752

6.3.3 Passage Related Concerns

Each passage in the ELA test has one or more associated items. The M-STEP CAT algorithm does not require that all items associated with a passage be administered; instead, it evaluates all possible combinations of items within a passage. Item sequencing within a passage is preserved when items are presented to the student. For example, if a six-item passage is selected and items 1 and 4 are not administered, then the items administered in order will be 2, 3, 5, and 6.

The configurable elements of a passage-based M-STEP CAT include the following:

Passage Minimum Percentage—This element defines the minimum percentage of the items associated with a passage to be used.

> For example, if the distinct passage minimum percentage is set at 80, then the selection routine will consider passage combinations such as 1 of 1 (100%), 4 of 5 (80%), 5 of 6 (83%), and 6 of 6 (100%). It will not consider combinations such as 1 of 2 (50%), 3 of 4 (75%), 3 of 5 (60%), etc. Near the end of a test, the passage minimum percentage constraint may need to be loosened by a configurable reduction factor to meet content constraints such as the number of items per assessment target.

Passage Minimum and Maximum Number—This element defines the minimum and maximum numbers of items in a passage combination.

> In the example above, 6 of 6 (100%) meets the passage minimum percentage (i.e., >=80%); however, this passage combination may not be selected if the maximum number of items in a passage is specified as 5.

Passage Evaluation Criteria—Multiple factors are considered when evaluating and ranking each passage combination to determine the best combination to administer to a student. Passage combinations with higher criteria rankings are more likely to be administered. The criteria used in M-STEP CAT were as follows:

- Percentage of items used—the percentage of items associated with the passage selected for consideration
- Items delivered—the total number of items associated with a passage relative to the number of items selected to be delivered per passage
- Max information of passage combination—the higher the item information, the higher the combination is ranked

Different weights may be assigned to each of the factors mentioned above. For example, if 100% of the weight is assigned to the number of items delivered, then the algorithm will select the passages with the highest number of associated items and administer all those items until the maximum number of items is reached. Based on the simulation results, the criteria shown in Table 6-2 provided a better result that balanced the psychometric and test blueprint specifications.

6.4 Test Navigation

Due to a variety of reasons, many versions of CAT engines do not allow students to skip items in a test or return to previously answered items to change answers. Currently, all mathematics tests do not allow students to skip items or return to items to change answers. However, in the ELA tests, students are allowed to skip items within a passage. For example, when presented with a passage and five associated items, a student does not have to answer questions 1–5 in that order. However, if the student tries to navigate to the next passage without answering all items associated with the previous passage, the test engine will prompt the student to answer all items and will not move to the next passage until all are answered.

6.5 Termination

The CAT algorithm allows for both a fixed- and a variable-length test. With a fixed-length test, the test ends when a student has taken a predefined fixed number of items. With a variable-length test, in some cases, the algorithm stops administering items when the threshold of SEM or the maximum number of items is reached. Following the criteria set by Smarter Balanced, which MDE adopted for M-STEP, the algorithm stops administering items when a student has taken a predefined minimum number of items and the test blueprint specifications have been met.

6.6 Forced Submission

Tests are considered "complete" when students respond to the minimum number of operational items specified in the blueprint. Otherwise, the tests are "incomplete." MLE is used to score the incomplete tests, counting unanswered items as incorrect.

When tests are adaptive, the specific unanswered items are unknown; thus, simulated items are used in place of administered items. Simulated items are generated with the following rules:

- Minimum operational test length is used to determine the test length of the incomplete tests.
- It is assumed that all unanswered operational items are dichotomously scored items. The item parameters of all unanswered operational items are equal to the average values of all the dichotomously scored operational items in the bank for discrimination and difficulty parameters.
- All unanswered operational items are scored as "incorrect."

Table 6-4 lists the average discrimination and difficulty item parameter estimates and the minimum number of items needed to calculate the scores of the forced submitted students.

Table 6-4. Key Values Used for the 2018–19 Forced Submission

Content	Grade	Mean Discrimination	Mean Difficulty	Minimum Number of Items
ELA	3	0.662	-0.52	44
ELA	4	0.586	0.043	44
ELA	5	0.599	0.415	44
ELA	6	0.545	0.91	44
ELA	7	0.524	1.185	44
Mathematics	3	0.842	-0.826	36
Mathematics	4	0.826	-0.098	36
Mathematics	5	0.762	0.495	36
Mathematics	6	0.690	1.135	36
Mathematics	7	0.712	1.858	36

6.7 Summary of Simulation Results Evaluating the CAT Algorithm

This section summarizes the CAT simulation results with regard to the evaluation of the operational 2018–19 CAT algorithm. It is described in two subsections: (1) adherence to the test blueprint and (2) control of item exposure. Overall, the results are as expected and meet the acceptable psychometric requirements given the available item pool. For comparisons, the Smarter Balanced 2017 Simulation Document[2] can be used.

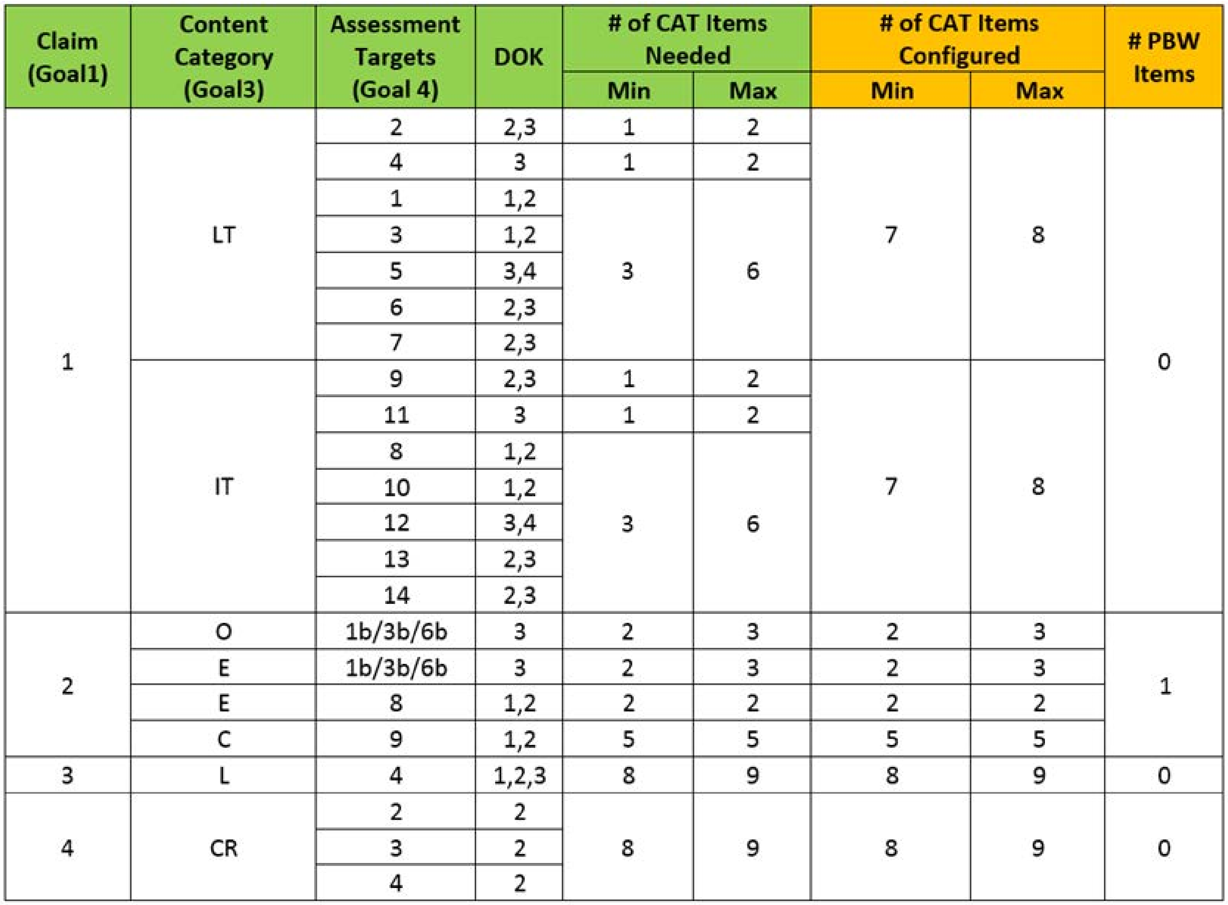
6.7.1 Adherence to the Test Blueprint

During the M-STEP CAT simulations, blueprint constraints were used to ensure that the test blueprint was adhered to for all grades and content areas (see Figures 6-1 to 6-6). Note that for all the ELA tests, given the available items in the item pool and the fact that all the items are passage based in Claim 1, the number of items in Claim 1 can be met only at the content-category level. The simulation results show that every student received the number of items configured.

Tables 6-5 and 6-7 summarize the minimum and maximum numbers of items and points by claim and total for ELA and mathematics. The minimum and maximum numbers of passages and the number of items per passage per claim and per content category in Claim 1 are also summarized for ELA in Table 6.6. The results indicate that the CAT engine offered students the expected number of items and points, the expected number of passages, and a reasonable number of items delivered per passage.

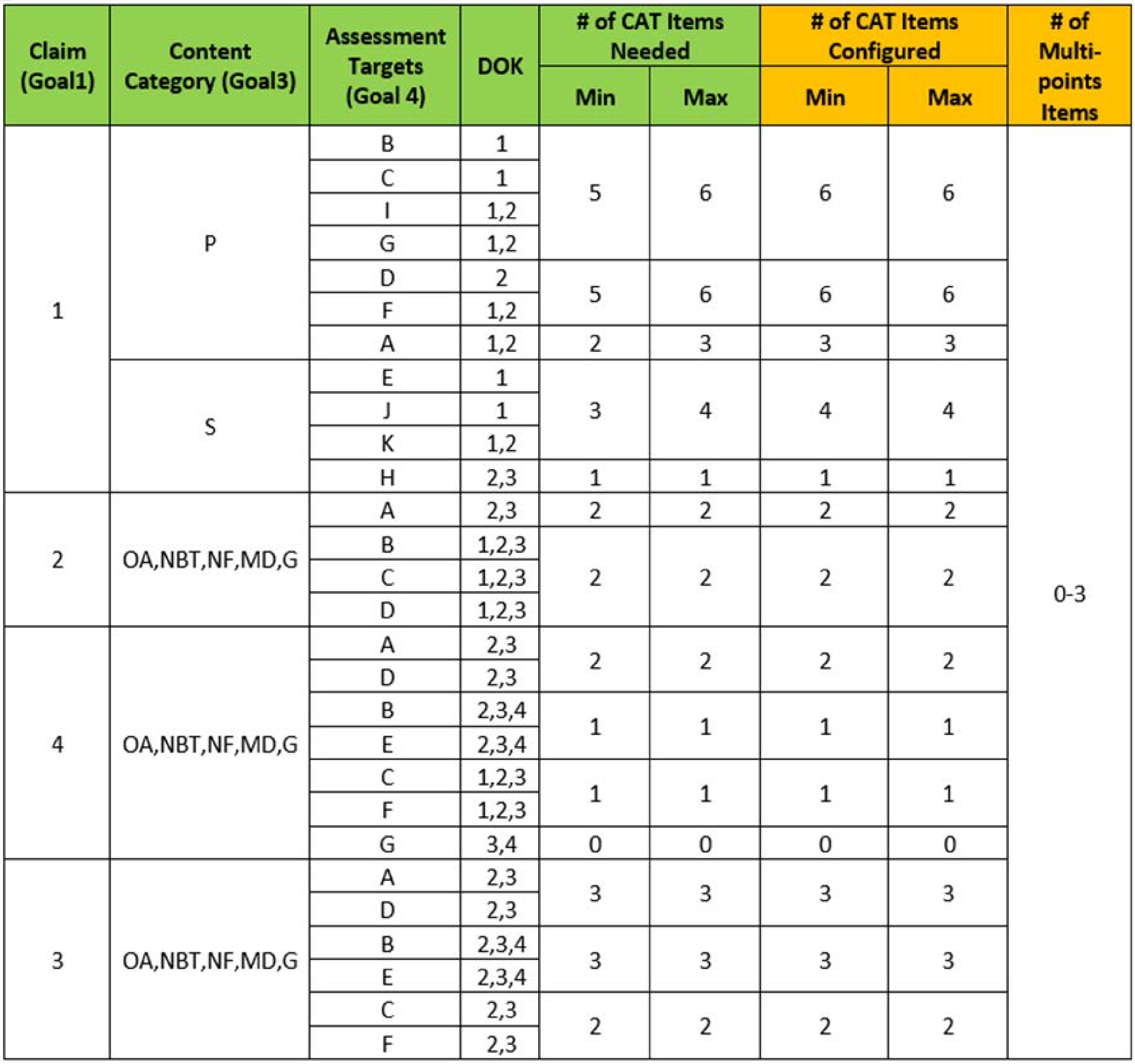
[2] https://portal.smarterbalanced.org/library/en/2016-17-summative-assessments-simulation-results.pdf

Figure 6-1. Blueprint Target Sampling, Grade 3 through Grade 7 ELA



Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# PBW Items
				Min	Max	Min	Max	
1	LT	2	2,3	1	2	7	8	0
		4	3	1	2			
		1	1,2	3	6			
		3	1,2					
		5	3,4					
		6	2,3					
		7	2,3					
	IT	9	2,3	1	2	7	8	
		11	3	1	2			
		8	1,2	3	6			
		10	1,2					
		12	3,4					
		13	2,3					
		14	2,3					
2	O	1b/3b/6b	3	2	3	2	3	1
	E	1b/3b/6b	3	2	3	2	3	
	E	8	1,2	2	2	2	2	
	C	9	1,2	5	5	5	5	
3	L	4	1,2,3	8	9	8	9	0
4	CR	2	2	8	9	8	9	0
		3	2					
		4	2					

Figure 6-2. Blueprint Target Sampling, Grade 3 Mathematics

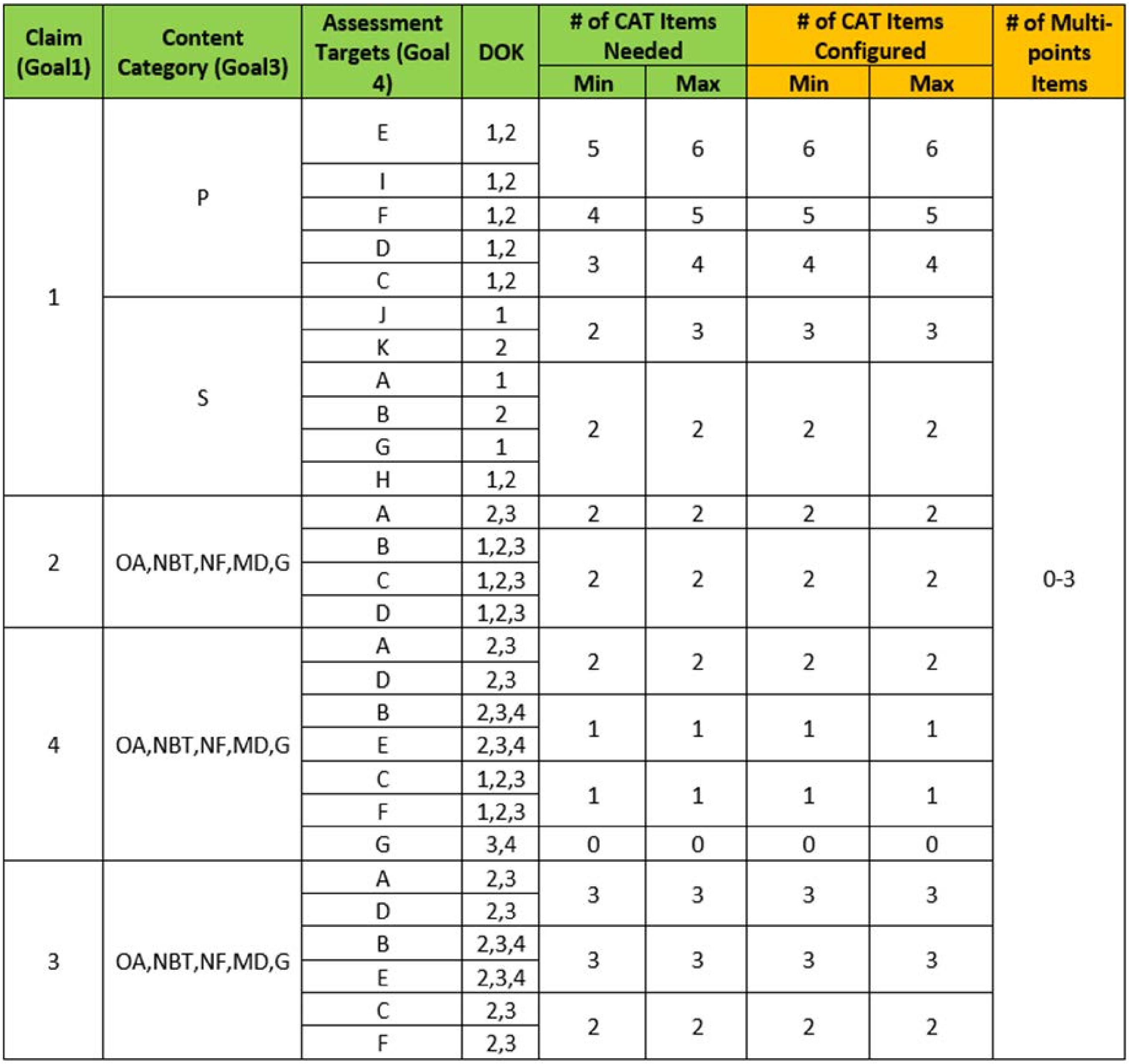


Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# of Multi-points Items
				Min	Max	Min	Max	
1	P	B	1	5	6	6	6	0-3
		C	1					
		I	1,2					
		G	1,2					
		D	2	5	6	6	6	
		F	1,2					
		A	1,2	2	3	3	3	
	S	E	1	3	4	4	4	
		J	1					
		K	1,2					
		H	2,3	1	1	1	1	
2	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		B	1,2,3	2	2	2	2	
		C	1,2,3					
		D	1,2,3					
4	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		D	2,3					
		B	2,3,4	1	1	1	1	
		E	2,3,4					
		C	1,2,3	1	1	1	1	
		F	1,2,3					
		G	3,4	0	0	0	0	
3	OA,NBT,NF,MD,G	A	2,3	3	3	3	3	
		D	2,3					
		B	2,3,4	3	3	3	3	
		E	2,3,4					
		C	2,3	2	2	2	2	
		F	2,3					

Figure 6-3. Blueprint Target Sampling, Grade 4 Mathematics

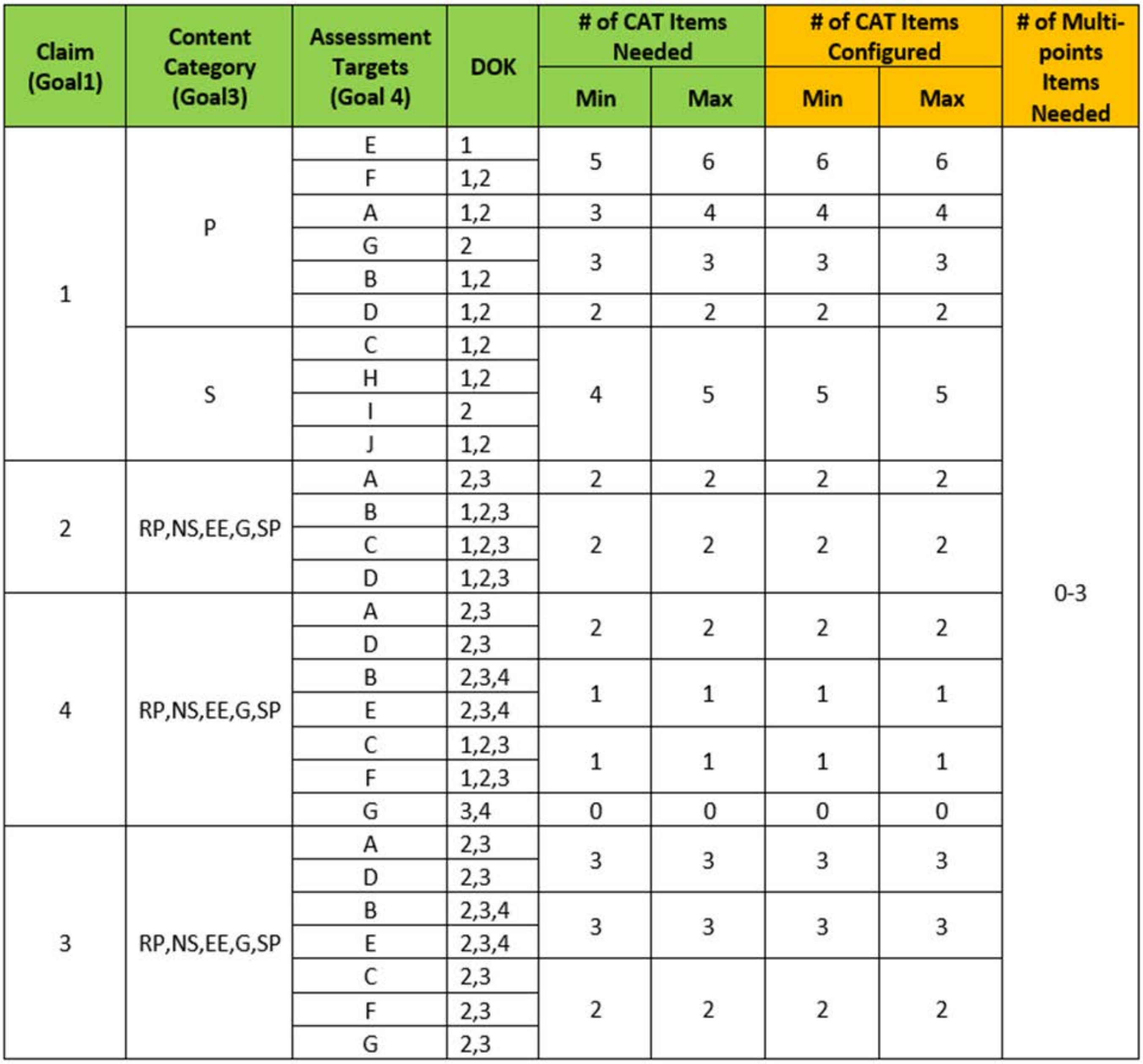
Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# of Multi-points Items
				Min	Max	Min	Max	
1	P	A	1,2	8	9	9	9	0-3
		E	1,2					
		F	1,2					
		G	1,2	2	3	3	3	
		D	1,2	1	2	2	2	
		H	1,2	1	1	1	1	
	S	I	1,2	2	3	3	3	
		K	1,2					
		B	1,2	1	1	1	1	
		C	2,3					
		J	1,2					
		L	1,2	1	1	1	1	
2	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		B	1,2,3	2	2	2	2	
		C	1,2,3					
		D	1,2,3					
4	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		D	2,3					
		B	2,3,4	1	1	1	1	
		E	2,3,4					
		C	1,2,3	1	1	1	1	
		F	1,2,3					
		G	3,4	0	0	0	0	
3	OA,NBT,NF,MD,G	A	2,3	3	3	3	3	
		D	2,3					
		B	2,3,4	3	3	3	3	
		E	2,3,4					
		C	2,3	2	2	2	2	
		F	2,3					

Figure 6-4. Blueprint Target Sampling, Grade 5 Mathematics



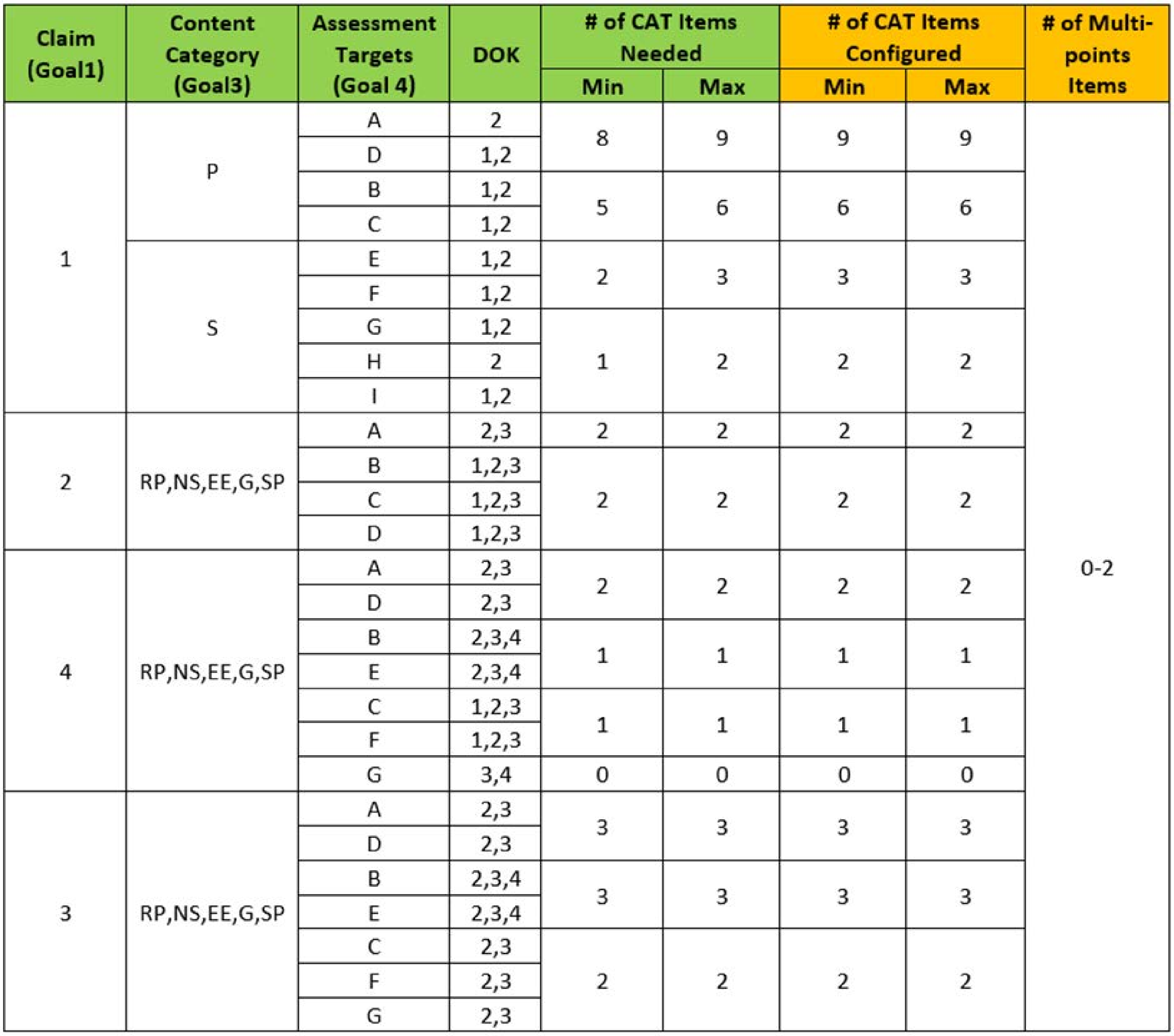
Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# of Multi-points Items
				Min	Max	Min	Max	
1	P	E	1,2	5	6	6	6	0-3
		I	1,2					
		F	1,2	4	5	5	5	
		D	1,2	3	4	4	4	
		C	1,2					
	S	J	1	2	3	3	3	
		K	2					
		A	1	2	2	2	2	
		B	2					
		G	1					
		H	1,2					
2	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		B	1,2,3	2	2	2	2	
		C	1,2,3					
		D	1,2,3					
4	OA,NBT,NF,MD,G	A	2,3	2	2	2	2	
		D	2,3					
		B	2,3,4	1	1	1	1	
		E	2,3,4					
		C	1,2,3	1	1	1	1	
		F	1,2,3					
		G	3,4	0	0	0	0	
3	OA,NBT,NF,MD,G	A	2,3	3	3	3	3	
		D	2,3					
		B	2,3,4	3	3	3	3	
		E	2,3,4					
		C	2,3	2	2	2	2	
		F	2,3					

Figure 6-5. Blueprint Target Sampling, Grade 6 Mathematics



Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# of Multi-points Items Needed
				Min	Max	Min	Max	
1	P	E	1	5	6	6	6	0-3
		F	1,2					
		A	1,2	3	4	4	4	
		G	2	3	3	3	3	
		B	1,2					
		D	1,2	2	2	2	2	
	S	C	1,2	4	5	5	5	
		H	1,2					
		I	2					
		J	1,2					
2	RP,NS,EE,G,SP	A	2,3	2	2	2	2	
		B	1,2,3	2	2	2	2	
		C	1,2,3					
		D	1,2,3					
4	RP,NS,EE,G,SP	A	2,3	2	2	2	2	
		D	2,3					
		B	2,3,4	1	1	1	1	
		E	2,3,4					
		C	1,2,3	1	1	1	1	
		F	1,2,3					
		G	3,4	0	0	0	0	
3	RP,NS,EE,G,SP	A	2,3	3	3	3	3	
		D	2,3					
		B	2,3,4	3	3	3	3	
		E	2,3,4					
		C	2,3	2	2	2	2	
		F	2,3					
		G	2,3					

Figure 6-6. Blueprint Target Sampling, Grade 7 Mathematics



Claim (Goal1)	Content Category (Goal3)	Assessment Targets (Goal 4)	DOK	# of CAT Items Needed		# of CAT Items Configured		# of Multi-points Items
				Min	Max	Min	Max	
1	P	A	2	8	9	9	9	0-2
		D	1,2					
		B	1,2	5	6	6	6	
		C	1,2					
	S	E	1,2	2	3	3	3	
		F	1,2					
		G	1,2	1	2	2	2	
		H	2					
		I	1,2					
2	RP,NS,EE,G,SP	A	2,3	2	2	2	2	
		B	1,2,3	2	2	2	2	
		C	1,2,3					
		D	1,2,3					
4	RP,NS,EE,G,SP	A	2,3	2	2	2	2	
		D	2,3					
		B	2,3,4	1	1	1	1	
		E	2,3,4					
		C	1,2,3	1	1	1	1	
		F	1,2,3					
		G	3,4	0	0	0	0	
3	RP,NS,EE,G,SP	A	2,3	3	3	3	3	
		D	2,3					
		B	2,3,4	3	3	3	3	
		E	2,3,4					
		C	2,3	2	2	2	2	
		F	2,3					
		G	2,3					

Table 6-5. Summary of Items and Points for ELA

Grade	Level	Min # of Items	Max # of Items	Min # of Points	Max # of Points
3	Total	45	46	48	49
3	Claim 1	16	16	16	16
3	Claim 2	13	13	16	16
3	Claim 3	8	9	8	9
3	Claim 4	8	8	8	8
3	Claim 1_LT	8	8	8	8
3	Claim 1_IT	8	8	8	8
4	Total	45	46	48	49
4	Claim 1	16	16	16	16
4	Claim 2	13	13	16	16
4	Claim 3	8	9	8	9
4	Claim 4	8	8	8	8
4	Claim 1_LT	8	8	8	8
4	Claim 1_IT	8	8	8	8
5	Total	44	46	47	49
5	Claim 1	15	16	15	16
5	Claim 2	13	13	16	16
5	Claim 3	8	9	8	9
5	Claim 4	8	8	8	8
5	Claim 1_LT	8	8	8	8
5	Claim 1_IT	7	8	7	8
6	Total	44	46	47	49
6	Claim 1	15	16	15	16
6	Claim 2	13	13	16	16
6	Claim 3	8	9	8	9
6	Claim 4	8	8	8	8
6	Claim 1_LT	7	8	7	8
6	Claim 1_IT	8	8	8	8
7	Total	45	46	48	49
7	Claim 1	16	16	16	16
7	Claim 2	13	13	16	16
7	Claim 3	8	9	8	9
7	Claim 4	8	8	8	8
7	Claim 1_LT	8	8	8	8
7	Claim 1_IT	8	8	8	8

Table 6-6. Summary of Passages and Items per Passage for ELA

Grade	Level	Min # of Passages	Max # of Passages	Min # of Items per Passage	Max # of Items per Passage
3	Total	7	8	2	4
3	Claim 1	4	4	4	4
3	Claim 3	3	4	2	3
3	Claim 1_LT	2	2	4	4
3	Claim 1_IT	2	2	4	4
4	Total	7	8	2	4
4	Claim 1	4	4	4	4
4	Claim 3	3	4	2	3
4	Claim 1_LT	2	2	4	4
4	Claim 1_IT	2	2	4	4
5	Total	7	8	2	4
5	Claim 1	4	4	3	4
5	Claim 3	3	4	2	3
5	Claim 1_LT	2	2	4	4
5	Claim 1_IT	2	2	3	4
6	Total	7	8	2	4
6	Claim 1	4	4	3	4
6	Claim 3	3	4	2	3
6	Claim 1_LT	2	2	3	4
6	Claim 1_IT	2	2	4	4
7	Total	7	8	2	4
7	Claim 1	4	4	4	4
7	Claim 3	3	4	2	3
7	Claim 1_LT	2	2	4	4
7	Claim 1_IT	2	2	4	4

Table 6-7. Summary of Items and Points for Mathematics

Grade	Level	Min # of Items	Max # of Items	Min # of Points	Max # of Points
3	Total	36	36	36	40
3	Claim 1	20	20	20	22
3	Claim 2	4	4	4	5
3	Claim 3	8	8	8	11
3	Claim 4	4	4	4	8
3	Claim 2 & 4	8	8	8	12
4	Total	36	36	36	40
4	Claim 1	20	20	20	23
4	Claim 2	4	4	4	4
4	Claim 3	8	8	8	10
4	Claim 4	4	4	4	6
4	Claim 2 & 4	8	8	8	10
5	Total	36	36	36	40
5	Claim 1	20	20	20	20
5	Claim 2	4	4	4	5
5	Claim 3	8	8	8	11
5	Claim 4	4	4	4	7
5	Claim 2 & 4	8	8	8	11
6	Total	36	36	36	40
6	Claim 1	20	20	20	22
6	Claim 2	4	4	4	6
6	Claim 3	8	8	8	11
6	Claim 4	4	4	4	7
6	Claim 2 & 4	8	8	8	11
7	Total	36	36	36	39
7	Claim 1	20	20	20	20
7	Claim 2	4	4	4	5
7	Claim 3	8	8	8	11
7	Claim 4	4	4	4	6
7	Claim 2 & 4	8	8	8	10

6.7.2 Controlling for Item Exposure

A common concern when implementing a CAT is the exposure rate of the items. It is important to control the item exposure rate while balancing the other constraints of the CAT. Tables 6-8 and 6-9 show the item exposure rates for ELA and mathematics, respectively. Each table provides the number and proportion of items for each of the six exposure rate categories, including no exposure. For example, an exposure rate of (0.0, 0.1] means that between 0% (excluding 0, as it forms its own category) and 10% of the students took that item. For grade 3 ELA, 606 items, or 70% of the items in the pool, had an exposure rate between 0% and 10%. For both ELA and mathematics, most items had a low exposure rate and were categorized with an exposure rate between 0% and 10%.

Table 6-8. Summary of Item Exposure Rate by Grade and Level for ELA

Grade	Level	Number Items	Proportion of Items
3	0	104	0.12
3	(0.0, 0.1]	606	0.70
3	(0.1, 0.2]	87	0.10
3	(0.2, 0.3]	32	0.04
3	(0.3, 0.4]	21	0.02
3	> 0.4	10	0.01
4	0	69	0.08
4	(0.0, 0.1]	604	0.74
4	(0.1, 0.2]	68	0.08
4	(0.2, 0.3]	52	0.06
4	(0.3, 0.4]	23	0.03
4	> 0.4	4	0.00
5	0	102	0.13
5	(0.0, 0.1]	516	0.67
5	(0.1, 0.2]	90	0.12
5	(0.2, 0.3]	38	0.05
5	(0.3, 0.4]	21	0.03
5	> 0.4	8	0.01
6	0	78	0.10
6	(0.0, 0.1]	525	0.70
6	(0.1, 0.2]	58	0.08
6	(0.2, 0.3]	47	0.06
6	(0.3, 0.4]	25	0.03
6	> 0.4	13	0.02
7	0	77	0.12
7	(0.0, 0.1]	446	0.67
7	(0.1, 0.2]	58	0.09

Grade	Level	Number Items	Proportion of Items
7	(0.2, 0.3]	35	0.05
7	(0.3, 0.4]	32	0.05
7	> 0.4	15	0.02

Table 6-9. Summary of Item Exposure Rate by Grade and Level for Mathematics

Grade	Level	Number Items	Proportion of Items
3	0	171	0.14
3	(0.0, 0.1]	921	0.76
3	(0.1, 0.2]	117	0.10
3	(0.2, 0.3]	2	0.00
3	(0.3, 0.4]	0	0.00
3	> 0.4	0	0.00
4	0	150	0.12
4	(0.0, 0.1]	1010	0.80
4	(0.1, 0.2]	106	0.08
4	(0.2, 0.3]	3	0.00
4	(0.3, 0.4]	0	0.00
4	> 0.4	0	0.00
5	0	149	0.12
5	(0.0, 0.1]	915	0.77
5	(0.1, 0.2]	129	0.11
5	(0.2, 0.3]	0	0.00
5	(0.3, 0.4]	0	0.00
5	> 0.4	0	0.00
6	0	126	0.12
6	(0.0, 0.1]	842	0.77
6	(0.1, 0.2]	118	0.11
6	(0.2, 0.3]	8	0.01
6	(0.3, 0.4]	0	0.00
6	> 0.4	0	0.00
7	0	69	0.07
7	(0.0, 0.1]	761	0.79
7	(0.1, 0.2]	112	0.12
7	(0.2, 0.3]	25	0.03
7	(0.3, 0.4]	0	0.00
7	> 0.4	0	0.00

6.8 Summary of Simulation Results for the Student Ability Estimates

For Smarter Balanced tests with an adaptive component, test reliability is estimated through simulations conducted using the operational summative item pool. For fixed-form tests, reliability and SEM are calculated using the items on the forms and their psychometric properties relative to the population. DRC conducted simulation studies for the 2017–18 tests using the 2016–17 M-STEP ability estimates, which had the means and SDs shown in Table 6-10. These results have remained stable with the consistent M-STEP design, with no difference shown when updating the ability distributions.

Table 6-10. Mean and Standard Deviation of the Sample Used in the Simulation Study

Content	Grade	Mean	SD
ELA	3	-1.07	1.01
ELA	4	-0.62	1.05
ELA	5	-0.28	1.08
ELA	6	0.06	1.09
ELA	7	0.30	1.12
Mathematics	3	-1.10	0.98
Mathematics	4	-0.63	1.02
Mathematics	5	-0.28	1.08
Mathematics	6	-0.10	1.23
Mathematics	7	0.08	1.35

6.8.1 Ability Estimates at the Extremes

The examinee ability in the simulation study was estimated using MLE. To provide a limit to the score range for extreme values, the test scoring algorithm used the HOSS and LOSS that were derived during the Smarter Balanced 2014 achievement level setting. Scores above HOSS or below LOSS are assigned HOSS and LOSS values respectively. Table 6-11 presents the LOSS and HOSS values that were used in the simulation and the percentage of the affected scores at those values.

Table 6-11. HOSS/LOSS and Percentages of Affected Scores from Simulation Results

Content	Grade	LOSS	HOSS	Percentage of Scores at LOSS	Percentage of Scores at HOSS
ELA	3	-4.59	1.34	0.09	0.93
ELA	4	-4.40	1.80	0.03	0.83
ELA	5	-3.58	2.25	0.10	0.63
ELA	6	-3.48	2.51	0.13	0.77
ELA	7	-2.91	2.75	0.33	1.03
Mathematics	3	-4.11	1.33	0.43	0.53
Mathematics	4	-3.92	1.82	0.07	0.67
Mathematics	5	-3.73	2.33	0.17	0.37
Mathematics	6	-3.53	2.95	0.70	0.37
Mathematics	7	-3.34	3.32	1.40	0.40

6.8.2 Standard Error of Measurement

The SEM, in the theta metric, is calculated for each reportable claim score and the total score. Note that for mathematics, the combined score for Claims 2 and 4 is reported, so the SEM for the combined score is calculated. Tables 6-12 and 6-13 provide statistical summaries (including the minimum, maximum, mean, median, and SD values) of the SEMs for claim scores and total scores. For all the tests, the average SEMs for claim scores are larger than the SEMs for the total scores. This is expected because the number of items in each claim is smaller than the number of items in the total test. As the grade increases, the average SEM increases. This is possibly due to the mismatch between the item difficulty distributions and student ability distributions in higher grades. The 3,000 simulated students' abilities or scores were randomly selected from the previous year's operational results on M-STEP. It was found that the items in the higher grades were relatively harder for the students. The SEMs are reasonable given the length of the total test and the claim level of the test.

Table 6-12. Summary of Standard Error of Measurement by Grade and Level for ELA

Grade	Level	Mean	SD	Min	Max	Median
3	Total	0.23	0.04	0.2	0.95	0.22
3	Claim 1	0.41	0.14	0.31	3.24	0.37
3	Claim 2	0.46	0.12	0.37	2.06	0.44
3	Claim 3	0.81	0.31	0.52	3.21	0.72
3	Claim 4	0.54	0.21	0.39	3.46	0.48
4	Total	0.26	0.03	0.23	0.75	0.25
4	Claim 1	0.45	0.17	0.36	3.14	0.42
4	Claim 2	0.50	0.12	0.36	2.23	0.49
4	Claim 3	0.81	0.29	0.55	3.49	0.73
4	Claim 4	0.63	0.22	0.45	3.33	0.56
5	Total	0.26	0.03	0.22	0.83	0.25
5	Claim 1	0.47	0.15	0.37	2.95	0.44
5	Claim 2	0.52	0.10	0.38	2.00	0.51
5	Claim 3	0.83	0.28	0.61	3.88	0.76
5	Claim 4	0.55	0.18	0.42	2.19	0.5
6	Total	0.28	0.05	0.24	1.62	0.26
6	Claim 1	0.50	0.16	0.40	2.79	0.46
6	Claim 2	0.56	0.17	0.44	3.28	0.52
6	Claim 3	0.82	0.33	0.54	3.59	0.72
6	Claim 4	0.64	0.25	0.46	2.85	0.56
7	Total	0.30	0.05	0.25	1.01	0.29
7	Claim 1	0.52	0.17	0.37	2.63	0.49
7	Claim 2	0.63	0.15	0.50	3.22	0.59
7	Claim 3	0.86	0.31	0.54	2.99	0.76
7	Claim 4	0.72	0.27	0.51	2.97	0.65

Table 6-13. Summary of Standard Error of Measurement by Grade and Level for Mathematics

Grade	Level	Mean	SD	Min	Max	Median
3	Total	0.21	0.05	0.18	1.41	0.20
3	Claim 1	0.27	0.05	0.23	1.33	0.26
3	Claim 3	0.59	0.30	0.34	3.09	0.50
3	Claim 2 & 4	0.56	0.30	0.33	2.65	0.45
4	Total	0.21	0.05	0.17	0.72	0.20
4	Claim 1	0.27	0.06	0.22	0.88	0.25
4	Claim 3	0.57	0.29	0.34	3.12	0.46
4	Claim 2 & 4	0.57	0.25	0.33	2.05	0.49
5	Total	0.25	0.09	0.17	1.47	0.22
5	Claim 1	0.32	0.11	0.22	1.49	0.28
5	Claim 3	0.63	0.33	0.37	2.99	0.51
5	Claim 2 & 4	0.71	0.42	0.33	3.28	0.55
6	Total	0.28	0.12	0.20	3.89	0.25
6	Claim 1	0.34	0.14	0.26	3.84	0.31
6	Claim 3	0.84	0.49	0.41	6.78	0.67
6	Claim 2 & 4	0.85	0.61	0.37	5.82	0.63
7	Total	0.32	0.17	0.19	2.31	0.28
7	Claim 1	0.39	0.24	0.24	5.44	0.33
7	Claim 3	1.04	0.68	0.41	6.34	0.79
7	Claim 2 & 4	1.03	0.72	0.35	5.77	0.76

6.8.3 Statistical Measures of Bias

This section presents the statistics calculated for the annual Michigan simulation investigation. Note that these statistics are the same as those reported in the *Smarter Balanced 2017–2018 Technical Report* (2018). Therefore, a direct quote from this Smarter Balanced report is used here for describing these statistics.

- Bias: [T]he statistical bias of the estimated theta parameter. This is a test of the assumption that error is randomly distributed around true ability. It is a measure of whether scores systematically underestimate or overestimate ability.
- Mean squared error (MSE): This is a measure of the magnitude of difference between true and estimated theta.
- Significance of bias [“Bias Sig” in Tables 6-14 and 6-15]: [A]n indicator of the statistical significance of bias.
- Average standard error of the estimated theta: This is the average of the simulated standard error of measurement [SEM] over all examinees. It is the marginal reliability for the simulated population.
- Standard error of estimates of theta at the 5th, 25th, 75th, and 95th percentiles
- Percentage of students’ estimated theta falling outside the 95% and 99% confidence intervals [Miss Rate]. (p. 2–3)

For detailed mathematical formulas in computing these statistics, please refer to pages 2–4 of the *Smarter Balanced 2017–2018 Technical Report* (2018).

Tables 6-14 and 6-15 present the bias of the estimated abilities for ELA and mathematics, respectively. As was found in the Smarter Balanced simulation study (Smarter Balanced, 2016), the bias in the overall scores is both small and insignificant. It should also be noted that claim scores do have some systematic bias. This is likely caused by the application of HOSS and LOSS values.

Table 6-14. Bias of the Estimated Theta from Simulation Results: ELA

Level	Grade	Mean Bias	SE of Mean Bias	Bias Sig	MSE	95% CI Miss Rate	99% CI Miss Rate
Overall	3	-0.01	0.01	0.52	0.06	4.81	0.92
Overall	4	-0.01	0.02	0.57	0.07	5.03	1.03
Overall	5	0.00	0.02	0.93	0.07	4.90	0.97
Overall	6	0.00	0.02	0.85	0.08	5.03	0.93
Overall	7	-0.01	0.02	0.59	0.10	5.23	1.07
Claim 1	3	-0.02	0.01	0.14	0.20	3.43	0.56
Claim 1	4	-0.04	0.02	0.07	0.24	3.77	0.43
Claim 1	5	-0.01	0.02	0.48	0.25	4.30	0.70
Claim 1	6	-0.02	0.02	0.26	0.29	3.40	0.53
Claim 1	7	-0.01	0.02	0.62	0.33	4.43	0.80
Claim 2	3	-0.01	0.01	0.64	0.24	3.23	0.24
Claim 2	4	0.00	0.02	0.93	0.28	3.43	0.23
Claim 2	5	0.01	0.02	0.58	0.29	3.47	0.23
Claim 2	6	-0.02	0.02	0.32	0.36	3.40	0.20
Claim 2	7	-0.05	0.02	0.02	0.43	2.83	0.40
Claim 3	3	-0.06	0.01	0.00	0.71	2.09	0.20
Claim 3	4	-0.04	0.02	0.05	0.77	2.00	0.17
Claim 3	5	-0.02	0.02	0.21	0.75	2.07	0.17
Claim 3	6	0.02	0.02	0.35	0.77	1.90	0.13
Claim 3	7	-0.02	0.02	0.41	0.80	1.80	0.17
Claim 4	3	-0.05	0.01	0.00	0.34	2.30	0.12
Claim 4	4	-0.06	0.02	0.00	0.43	2.40	0.13
Claim 4	5	-0.01	0.02	0.52	0.32	2.17	0.10
Claim 4	6	-0.06	0.02	0.00	0.44	2.13	0.20
Claim 4	7	-0.07	0.02	0.00	0.58	2.47	0.17

Table 6-15. Bias of the Estimated Theta from Simulation Results: Mathematics

Level	Grade	Mean Bias	SE of Mean Bias	Bias Sig	MSE	95% CI Miss Rate	99% CI Miss Rate
Overall	3	-0.01	0.02	0.57	0.05	4.77	1.20
Overall	4	-0.01	0.02	0.64	0.05	4.70	1.47
Overall	5	-0.02	0.02	0.33	0.07	4.57	0.70
Overall	6	-0.03	0.02	0.26	0.11	5.53	0.87
Overall	7	-0.04	0.02	0.15	0.14	4.83	0.83
Claim 1	3	-0.01	0.02	0.62	0.08	4.20	0.93
Claim 1	4	-0.01	0.02	0.64	0.07	4.50	0.80
Claim 1	5	-0.03	0.02	0.14	0.12	3.63	0.80
Claim 1	6	-0.03	0.02	0.25	0.14	5.03	1.10
Claim 1	7	-0.05	0.02	0.06	0.21	4.03	0.67
Claim 3	3	-0.08	0.02	0.00	0.37	2.47	0.43
Claim 3	4	-0.06	0.02	0.00	0.35	2.90	0.50
Claim 3	5	-0.08	0.02	0.00	0.44	2.43	0.23
Claim 3	6	-0.14	0.02	0.00	0.72	2.37	0.43
Claim 3	7	-0.22	0.03	0.00	1.09	2.77	0.50
Claims 2 & 4	3	-0.06	0.02	0.00	0.31	2.33	0.27
Claims 2 & 4	4	-0.07	0.02	0.00	0.33	2.60	0.47
Claims 2 & 4	5	-0.13	0.02	0.00	0.50	2.17	0.43
Claims 2 & 4	6	-0.12	0.02	0.00	0.58	2.90	0.57
Claims 2 & 4	7	-0.14	0.02	0.00	0.77	3.10	0.47

Tables 6-16 and 6-17 below present marginal reliability coefficients and precisions for the overall tests and for reported claims. As expected, estimated reliability coefficients for the overall tests are high and are in the acceptable range for a large-scale, high-stakes test. Reliability estimates at the claim level are lower, and corresponding errors are higher. Claims with smaller numbers of items and fewer points from the adaptive section of the test exhibit the lowest reliability. This shows the importance of incorporating error data in claim-level reports.

Table 6-16. Overall Score and Claim Score Precision/Reliability of Simulation Results: ELA

Level	Grade	Mean # Items	Mean SEM	Reliability	RMSE	SD theta
Overall	3	45.22	0.24	0.95	0.25	1.06
Overall	4	45.04	0.26	0.94	0.26	1.10
Overall	5	45.41	0.26	0.95	0.26	1.12
Overall	6	45.15	0.28	0.94	0.29	1.15
Overall	7	45.06	0.30	0.93	0.31	1.19
Claim 1	3	16.00	0.41	0.86	0.44	1.15
Claim 1	4	16.00	0.45	0.84	0.49	1.21
Claim 1	5	16.00	0.47	0.84	0.50	1.23
Claim 1	6	16.00	0.50	0.83	0.54	1.26
Claim 1	7	16.00	0.52	0.82	0.57	1.31
Claim 2	3	13.00	0.47	0.83	0.49	1.17
Claim 2	4	13.00	0.50	0.82	0.53	1.21
Claim 2	5	13.00	0.52	0.82	0.54	1.24
Claim 2	6	13.00	0.56	0.79	0.60	1.29
Claim 2	7	13.00	0.63	0.77	0.66	1.36
Claim 3	3	8.22	0.81	0.62	0.84	1.40
Claim 3	4	8.04	0.81	0.65	0.88	1.45
Claim 3	5	8.43	0.83	0.64	0.87	1.45
Claim 3	6	8.17	0.82	0.65	0.88	1.50
Claim 3	7	8.06	0.86	0.65	0.89	1.55
Claim 4	3	8.00	0.55	0.77	0.58	1.23
Claim 4	4	8.00	0.63	0.74	0.66	1.29
Claim 4	5	8.00	0.55	0.79	0.57	1.28
Claim 4	6	8.00	0.64	0.74	0.67	1.33
Claim 4	7	8.00	0.72	0.70	0.76	1.42

Table 6-17. Overall Score and Claim Score Precision/Reliability of Simulation Results: Mathematics

Level	Grade	Mean # Items	Mean SEM	Reliability	RMSE	SD theta
Overall	3	36	0.21	0.95	0.22	1.02
Overall	4	36	0.21	0.96	0.22	1.05
Overall	5	36	0.25	0.95	0.27	1.14
Overall	6	36	0.28	0.94	0.33	1.30
Overall	7	36	0.32	0.94	0.38	1.44
Claim 1	3	20	0.27	0.93	0.28	1.04
Claim 1	4	20	0.27	0.93	0.27	1.07
Claim 1	5	20	0.32	0.91	0.34	1.17
Claim 1	6	20	0.34	0.92	0.38	1.32
Claim 1	7	20	0.39	0.90	0.46	1.48
Claim 3	3	8	0.59	0.71	0.61	1.22
Claim 3	4	8	0.57	0.73	0.59	1.23
Claim 3	5	8	0.63	0.72	0.66	1.34
Claim 3	6	8	0.84	0.61	0.85	1.56
Claim 3	7	8	1.04	0.53	1.04	1.80
Claim 2 & 4	3	8	0.56	0.70	0.56	1.17
Claim 2 & 4	4	8	0.57	0.73	0.57	1.20
Claim 2 & 4	5	8	0.71	0.62	0.71	1.35
Claim 2 & 4	6	8	0.85	0.51	0.76	1.50
Claim 2 & 4	7	8	1.03	0.41	0.88	1.64

One of the advantages of adaptive tests is that SEM can be controlled for all ability levels. Ideally, the SEM should be similar throughout the ability distribution. Table 6-18 presents average error by decile of the true thetas, which were generated based on the Michigan population. For both ELA and mathematics, the results show that the error at the lower end of the test tends to be the highest, indicating that there is more error associated with the ability estimation at the lower end of the ability distribution, which is caused by the relative difficulty of the item pools.

Table 6-18. Average Standard Errors by Grade and by Deciles of True Proficiency Scores of Simulation Results

Subject	Grade	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
ELA	3	8.11	6.51	6.07	5.91	5.70	5.56	5.58	5.77	5.97	6.66
ELA	4	7.61	6.32	6.07	6.02	6.01	6.01	6.00	6.01	6.09	6.84
ELA	5	8.06	6.51	6.10	6.05	6.04	6.12	6.36	6.68	6.94	7.35
ELA	6	9.28	7.30	6.71	6.38	6.17	6.06	6.04	6.09	6.26	6.97
ELA	7	9.61	7.62	7.12	6.83	6.74	6.70	6.72	6.82	6.94	7.39
Math	3	8.09	5.99	5.63	5.27	5.12	5.03	5.01	5.00	5.02	5.53
Math	4	8.10	6.08	5.47	5.11	5.01	4.98	4.81	4.61	4.58	5.11
Math	5	9.94	7.74	6.79	6.05	5.34	4.90	4.46	4.15	4.22	4.78
Math	6	10.04	6.90	6.16	5.74	5.31	5.06	4.91	4.60	4.21	4.33
Math	7	13.28	8.66	7.09	6.38	5.78	5.24	4.75	4.21	4.04	4.06

6.9 Summary

In summary, Chapter 6 of this report demonstrates M-STEP’s adherence to AERA, APA, & NCME (2014) *Standards* regarding construct-related validity and reliability. The analyses described above are related to the following standards:

- Standard 2.0— Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.
- Standard 2.1—The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation.
- Standard 4.3—Test developers should document the rationale and supporting evidence for the administration, scoring, and reporting rules used in computer-adaptive, multistage-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or sets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and in controlling item exposure.

Chapter 7: Scoring

Chapter 7 shows how M-STEP scoring adhered to the AERA, APA, & NCME *Standards*. Standard 4.18 provides some general guidance for Chapter 7:

> Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays. (p. 91)

Chapter 7 explains the procedures used for autoscoring and handscoring, with the latter applicable to the Passage-Based Writing items. There was no AI scoring of student writing in 2019. The scoring criteria used for each item are not presented in this chapter to preserve the integrity of the items for future use.

7.1 Online Scoring

7.1.1 Autoscoring

All content areas of M-STEP contain items that required autoscoring. Autoscoring was used for Technology Enhanced items which could involve combining many components to form a single correct answer. Scoring rules for each item were set up prior to the start of testing. These rules listed all the different correct components per item. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring began. Quality checks were run against all autoscored items using the autoscoring simulator tool to ensure the item was scored as designed. The autoscoring simulator tool allowed specialists to respond to items with expected responses. In some cases, the simulator generated all possible responses and expected values. Once student responses were entered, the scoring engine was run against the student response and the expected value. The tool alerted the specialist of any mismatches, which could then be updated. Quality checks occurred before the start of the testing window. During the testing window, the autoscoring process ran continuously for CAT items and daily for fixed forms. All available, completed items were scored. After testing was complete, a secondary check was completed by the psychometrics team. Any items that did not perform as expected were communicated back to the autoscoring specialists, who reran the simulations to assure the autoscoring was set up as requested. If an autoscoring setup issue was found at this point, the items are updated and rescored. This occurred before reporting.

DRC provided MDE with complete item frequency reports, which includes the following information for each response pattern/combination: (1) the number/percentage of students gave that response pattern/combination, and (2) the score provided by the scoring system.

7.1.2 Multiple Choice Scoring

The online scoring process includes the scoring of multiple-choice items, in which students chose only one correct answer from choices A–D. The items were scored against a scoring key that was prepared and validated before the start of each testing window. Responses to multiple-choice items were captured during the online test administration, and items were scored as “right,” “wrong,” or “blank” (i.e., not answered). Additional answer key checks were conducted during the testing windows to ensure that the items were scored based on the provided key.

7.2 Handscoring

Measurement Incorporated performed all required scoring of paper/pencil and of online items needing handscoring. For M-STEP ELA, these were Passage-Based Writing (PBW) items for grades 3–7. For M-STEP mathematics, these included short-text and short-text fill-in table items for grades 3–7.

M-STEP items were scored by readers working in Taylor, Michigan; Grand Rapids, Michigan; and at other scoring centers (i.e., Durham, Greensboro, Wilmington, and Charlotte, North Carolina; Nashville, Tennessee; Tampa, Florida). Readers also scored remotely through Virtual Scoring Center (VSC Score) (i.e., distributive scoring).

AERA, APA, & NCME (2014) Standard 4.20 specifies the following:

> The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers’ responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters’ scoring. (p. 92)

Sections 7.2.1 through 7.2.5 explain how scorers are selected and trained for the M-STEP handscoring process. Sections 7.2.6 and 7.2.7 describes how the scorers are monitored throughout the M-STEP handscoring process.

7.2.1 Security

All Measurement Incorporated scoring rooms are designated secure areas with stringent security regulations that are vigorously enforced. Measurement Incorporated routinely implements a number of measures to help safeguard the security of student responses while they are in Measurement Incorporated’s possession and to maintain the confidentiality of student identity.

In the scoring rooms, the use of cellphones, tablets, MP3 players, laptops, or recording or photographic equipment is prohibited. The copying of materials for anything other than the training purposes that are expressly permitted by MDE is prohibited.

All buildings that house student responses, including Measurement Incorporated headquarters, scoring centers, and warehouses, utilize an electronic security system during nonbusiness hours.

All readers scoring remotely are required to work from a private, password-protected environment. No free or public Wi-Fi can be used. Readers can access a project website only from a secure, password-protected network. Readers cannot access any project website from a public computer or a public network, such as a wireless network at a hotel or restaurant. While in VSC Score, readers are unable to take screenshots or to access e-mail or other applications. Maintaining a secure workstation is a condition for employment for all remote employees.

Before receiving any training materials, all scoring project staff are required to sign a confidentiality and proprietary agreement, which indicates that no participant in training and/or scoring may reveal any specific information about the test or about the criteria and methods for scoring to any person as part of his or her contractual obligation to score student responses.

At scoring centers, all training materials remain on the premises during a project and are collected at the end of each workday to be secured. All materials are collected and accounted for at the end of the scoring project.

Readers who score remotely access training materials from an online resource library. The software does not allow readers to print or download data.

No identifying student information is provided on the images sent to readers via VSC Score software.

Readers do not have the ability to access training materials or student responses unless they and their team leader are logged on to the system.

Violation of any portion of the Measurement Incorporated security policy results in termination.

7.2.2 Measurement Incorporated Reader and Team Leader Hiring

Measurement Incorporated recruits, interviews, and hires a pool of readers to ensure sufficient staff for scoring projects.

All readers must have a minimum of a bachelor’s degree. MDE has the right to review the names, demographics, educational backgrounds, and experience (including scoring experience) of all readers. Reader degrees are verified before the applicants are interviewed. Applicants must provide either an official transcript with a seal (no copies accepted), an official letter from a registrar’s office (which would be mailed to the Site Manager), or access to a third-party company such as Parchment or Student Clearing House. Reader applicants can also bring their original diploma with a seal when they come for an interview.

Team leaders are selected and recruited from experienced reader staff. Each team leader supervises a group of 10–12 readers during live scoring.

7.2.3 Preparation of Training Materials for M-STEP

Three sets of student responses were used in training readers and team leaders:

- Anchor sets consisted of typical student responses at each score point, with examples of what would barely earn that point, a median answer for that point, and a high response within that point without quite reaching the next point. These sets were used to show readers and team leaders how the rubric was applied to each response.
- Training sets consisted of atypical student responses and were used to further demonstrate application of the rubric to actual student responses.
- Qualifying sets consisted of student responses similar to those in the anchor and training sets. These sets were used for readers to demonstrate their understanding of the application of the rubric to student responses.

Measurement Incorporated scoring directors used MDE-approved training materials. Anchor sets consisted of three responses at each score point. Each response was annotated to explain how the rubric criteria were applied. Training sets contained 5–10 papers. There was a training set for each trait for analytic scoring and a training set that combined the traits. The responses in each of these sets were arranged in random score-point order, and all score points were represented.

7.2.4 Training and Qualifying Readers and Team Leaders

AERA, APA, & NCME (2014) Standard 6.9 specifies the following:

> Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (p. 118)

Readers and team leaders were trained by the scoring director on the scoring criteria approved by MDE and were required to achieve qualifying standards set by MDE.

Readers were divided into teams consisting of one team leader and 10–15 readers.

For brief write, research, and mathematics scoring, the scoring director presented the item and anchor set and then discussed each score point as readers and team leaders took notes.

Following the presentation of these anchor sets, readers and team leaders scored a training set and then one or two qualifying sets.

For full write scoring, the scoring director introduced the readers and team leaders to the three analytic traits (i.e., Organization/Purpose, Evidence/Elaboration, and Conventions) using a unique anchor set for each trait so that readers and team leaders were fluent in the individual traits before they scored the traits simultaneously.

Following the presentation of each trait anchor set, a training set was scored for the trait and discussed; readers and team leaders then prepared to score all traits concurrently. Readers and team leaders took two qualifying sets, in which scores were assigned for all three analytic traits on each student response.

Readers and team leaders were provided a copy of anchor sets, training sets, and qualifying sets. Readers and team leaders were required to refer to the anchor sets and their notes when taking training sets and qualifying sets.

Readers and team leaders scored the qualifying set and submitted their scores. The percentage of correct scores was recorded. After the set was completed, the scoring director discussed the set with the group.

If a particular response or type of response generated numerous questions across teams, the scoring director discussed the problem with the group or posted a note to chat to ensure that everyone heard the same explanation.

Once the group had finished discussing the first qualifying set, the readers and team leaders scored the next set. Training continued until all training sets and qualifying sets were scored and discussed.

Readers were required to demonstrate their ability to score accurately by attaining the qualifying agreement percentage approved by MDE before they gained access to actual student responses.

Any reader or team leader unable to meet the qualifying standards set by MDE was released.

Reference Tables 7-1 and 7-2 for additional information.

Table 7-1. Qualifying Sets

Content	Number of Qualifying Sets per Item
Math	1 or 2
Research	1
Brief Write	1
Full Write	2 for each trait

Table 7-2. Qualifying Standards

Score Point Range	Qualifying Standard (Exact Agreement)
0–1	90%; no nonadjacent scores
0–2	80%; no nonadjacent scores
0–3	70%; no nonadjacent scores

7.2.5 Virtual Scoring Center

Measurement Incorporated used its VSC Score system for the image-based scoring of paper/pencil responses and for the scoring of online responses transferred to Measurement Incorporated from DRC.

Readers and team leaders accessed the VSC Score system through a secure web-based interface with the use of a unique user ID and password. Each team leader and reader was assigned a unique number for easy identification of his or her scoring work throughout the scoring session. VSC Score enabled readers and team leaders to score only those items that they were trained and qualified to score.

Each PBW response was randomly assigned to be read by one reader. A random sample of all student responses (i.e., 10% of responses) was then randomly assigned to a second reader. VSC Score managed readers' individual workloads and allowed readers to review and submit their scores.

Readers were trained on how to use the VSC Score performance assessment scoring system—how to assign scores, how to adjust the image for legibility, how to "flag" responses that were atypical from the anchor sets, training sets, and qualifying sets for review by the team lead and scoring director, etc.

Readers logged in and "checked out" a scoring set of student responses. This scoring set was generated by randomly selecting student responses from the pool of unscored student responses. The reader evaluated the first response, entered the score by clicking the appropriate value on the scoring toolbar, and clicked the "Submit" button. The next response in the scoring set then appeared for the reader to score and submit. This process continued until all responses in the set had been scored. After scoring all responses in a set, the reader had the option to review any of the responses and modify the scores before submitting them to the system.

Once the scores had been submitted, the set was "checked in" and responses were routed to other qualified readers as necessary. The requirements for subsequent readings were defined in the system during setup, and student responses were not marked as complete until the requisite number of independent readers had scored the response.

When a reader had a question about a response, he or she could transfer the image (along with the question and/or comments) from the current scoring set to a review set, which was assigned to a team leader. The team leader could forward the question to the scoring director, submit the appropriate score, or return the response to the reader with comments. This procedure was used whenever a reader had scoring concerns or encountered apparent non-scorable responses. Readers could mark completely blank responses as non-scorable, but otherwise only scoring directors or the project director could assign a non-scorable condition code to a student response.

7.2.6 Quality Control and Reliability of Scoring

AERA, APA, & NCME (2014) Standard 6.8 states the following:

> Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (p. 118)

Section 7.2.6 explains the monitoring procedures that Measurement Incorporated uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored. Detailed scoring rubrics are available for all PBW prompts, which specify the criteria for scoring those PBW prompts. These rubrics will not be presented in this report in order to preserve the integrity of the items for use in future assessments.

MDE reader production and reliability statistics, including reader training results, were available to MDE via a suite of VSC reports, which could be accessed online using secure credentials supplied to MDE staff.

Detailed Reader Status Reports were generated for each scoring project, utilizing a comprehensive system for collecting and analyzing score data. Daily analyses of the Reader Status Reports alerted management personnel to individual or group retraining needs.

After the readers' scores were submitted in the VSC Score system, the data was uploaded into the primary Scoring Resource Center servers. The scores were then validated and processed.

Updated real-time reports that showed both daily and cumulative data (i.e., project-to-date data) were available 24 hours a day via a secure website. Reports included data on the number of responses scored by each reader, the percentage of responses scored that day in exact agreement or adjacent agreement with a second reader, and the total number of responses scored at each score point.

For M-STEP performance assessment scoring, a random sample of 10% of all student responses are scored a second time to generate agreement data.

Readers were required to demonstrate the ability to assign scores consistently according to the rubric and anchor papers that were introduced during training. Their scoring accuracy is under scrutiny using validity responses that were included daily with the actual student responses (for details, see Section 7.2.7).

If questionable reader reliability indications were found, the affected responses were scored again.

The monitoring and retraining process was sustained throughout the project to promote strict adherence to MDE-approved scoring criteria and consistency throughout the scoring effort.

Scoring directors and team leaders provided consistent monitoring of the scoring patterns of each reader throughout the project, responded to questions, spot-checked (i.e., read behind) reader scoring, provided feedback, and counseled readers who were having difficulty with the criteria.

Scoring directors continued to look for atypical types of responses that were not covered in the initial training and presented further instruction about handling these types of responses when necessary.

The inter-rater reliability information for the handscored ELA and mathematics items is presented in Table 7-3.

Table 7-3. Human-to-Human Inter-rater Reliability

Content	Grade	Item ID	Maturity	% Perfect Plus	N Perfect	% Perfect	N Adjacent	% Adjacent	N Nonadjacent	% Nonadjacent
ELA	3	945969	OP	97.6	901	61.7	525	35.9	35	2.4
ELA	3	945975	OP	98.0	994	63.4	541	34.5	32	2.0
ELA	3	945976	OP	97.6	793	58.4	534	39.3	32	2.4
ELA	4	945962	OP	96.0	1383	61.9	763	34.1	89	4.0
ELA	4	945980	OP	96.7	1587	64.3	800	32.4	81	3.3
ELA	4	945981	OP	95.5	1449	64	714	31.5	101	4.5
ELA	5	945968	OP	98.2	1912	67.1	888	31.2	50	1.8
ELA	5	945974	OP	97.9	1832	69.1	763	28.8	57	2.1
ELA	5	945983	OP	97.9	1780	66.9	825	31.0	56	2.1
ELA	6	945965	OP	97.8	2102	74.1	670	23.6	63	2.2
ELA	6	945972	OP	98.1	2217	73.5	741	24.6	57	1.9
ELA	6	945984	OP	98.6	2286	75.4	704	23.2	41	1.4
ELA	7	945966	OP	99.6	2338	76.5	707	23.1	13	0.4
ELA	7	945970	OP	97.1	1245	59.2	797	37.9	61	2.9
ELA	7	945985	OP	99.2	2326	73.3	824	26.0	25	0.8
Math	3	10773	OP	100	48	76.2	15	23.8	-	0.0
Math	3	3030	OP	100	60	98.4	1	1.6	-	0.0
Math	3	6265	OP	100	61	96.8	2	3.2	-	0.0
Math	3	76503	OP	100	63	100	-	0.0	-	0.0
Math	3	78864	OP	100	62	98.4	1	1.6	-	0.0
Math	4	2668	OP	100	65	100	-	0.0	-	0.0
Math	4	3395	OP	100	65	98.5	1	1.5	-	0.0
Math	5	5523	OP	100	97	94.2	6	5.8	-	0.0
Math	5	78706	OP	100	103	98.1	2	1.9	-	0.0
Math	6	79766	OP	100	57	100	-	0.0	-	0.0
Math	7	12069	OP	100	34	100	-	0.0	-	0.0
Math	7	7180	OP	100	34	100	-	0.0	-	0.0

7.2.7 Validity

Measurement Incorporated used validity responses, similar to the student responses found in the qualifying sets, during live scoring to monitor readers' accuracy in scoring. Preselected validity responses were approved by MDE. Scoring directors also had the ability to select live responses as validity responses, which were also subject to MDE approval. The true scores for these responses were entered into a validity database.

Validity responses were randomly incorporated into readers' sets each day of the project. Team leaders reviewed the validity results and provided feedback to the readers.

A validity report was generated that included the response identification number, the scores assigned by the readers, and the "true" scores. Measurement Incorporated provided MDE with daily and project-to-date summaries of what percentages of papers scored by readers matched the validity checks or were high or low at each score point. Five percent of the responses that a reader scored were validity papers. These responses appeared to the reader daily throughout the entire scoring project. The validity standards can be found in Table 7-4.

Table 7-4. Validity Standards

Score Point Range	Validity Standard (Exact Agreement)
0–1	90%
0–2	80%
0–3	80%
0–4	70%

7.2.8 Alerts

Measurement Incorporated implemented a formal process for notifying MDE when student responses reflected a possibly dangerous situation for the student, which may include responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

Measurement Incorporated also alerted MDE if there appeared to be possible instances of teacher or proctor interference or student collusion with other students.

Measurement Incorporated always takes immediate action following a scoring alert.

7.3 Summary

The information presented in this chapter summarizes the scoring procedures for different types of items and the steps taken by DRC and Measurement Incorporated to ensure accuracy in the technology-enhanced item scoring and handscoring processes. The reliability statistics presented in Sections 7.2.6 and 7.2.7 demonstrate that the items are scored reliably. These efforts follow multiple best practices of the testing industry, particularly AERA, APA, & NCME (2014) *Standards* 4.18 4.20, 6.8, and 6.9:

- Standard 4.18—Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
- Standard 4.20—The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.
- Standard 6.8—Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.
- Standard 6.9—Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.

Chapter 8: Operational Data Analyses

This chapter describes the analyses conducted with the operational (OP) data. Item/test analyses from both the Classical Test Theory (CTT) and the item response theory (IRT) frameworks are used (when appropriate) and reported here.

This chapter demonstrates adherence of M-STEP to AERA, APA, & NCME (2014) *Standards* 1.8, 5.2, 5.13, and 5.15. Each standard will be explicated within the appropriate section of this chapter. Standard 7.2 provides general guidance that is relevant to this chapter:

> The population for whom a test is intended and specifications for the test should be documented. (p. 126)

Chapter 3 presents the test specifications. Information regarding reported data is discussed in detail in Chapter 9.

8.1 Operational Analysis of ELA and Mathematics

The *Smarter Balanced 2016–2017 Technical Report* (2017) states that part of the Smarter Balanced Theory of Action is to leverage appropriate technology and innovation. The use of CAT methodologies helps ensure that students across the range of proficiency levels have an assessment experience with items well targeted to their skill level. Adaptive testing allows average-, low-, and high-performing students to stay engaged in the assessment because they respond to items specifically targeted to their skill level. CATs are also efficient because they provide a higher level of score precision than fixed-form tests with the same number of items. For the CAT component, there are both content constraints (e.g., a long reading passage in ELA must be administered) and psychometric criteria that must be optimized for each student.

8.1.1 CAT Item Pool Characteristics

8.1.1.1 CAT Item Types

This section presents different item types used by Smarter Balanced (*Smarter Balanced 2017–2018 Technical Report*, 2018, p. 4–28) to compose the summative item pools. These different item types are listed in Table 8-1.

Table 8-1. Item Types Found in the Summative Item Pools

Item Types	ELA	Mathematics
Multiple Choice (MC)	X	X
Multi-Select (MS)	X	X
Evidence-Based Selected Response (EBSR)	X	
Match Interaction (MI)	X	X
Hot Text (HTQ)	X	
Passage Based Writing (PBW)	X	
Equation Response (EQ)		X
Grid-Item Response (GI)		X
Table Interaction (TI)		X
Table Interaction (TI)		X
Constructed Response (CR)		X

The Smarter Balanced item/task type characteristics are defined as sufficient to ensure that the content measured the intent of the Common Core State *Standards* (CCSS) and that there was consistency across item/task writers and editors. This included item types such as selected-response, constructed-response, and technology-enhanced.

As shown in Table 8-1, the common item types for both ELA and mathematics are MC, MS, and MI. In addition, ELA also included the following item types: EBSR, HTQ, and PBW. Mathematics also included the following item types: EQ, GI, and TI.

For ELA, PBW prompts are included in the pool. For more information on PBW prompts, please see Chapter 3. Additionally, it should be noted that the following sections provide information about the ELA and mathematics item pools that were administered in Michigan.

8.1.1.2 CAT Item Pool Specification

“An item pool refers to a collection of test questions (known as items) that support the test blueprint for a particular content area and grade. The Consortium takes multiple steps to ensure the quality of the items in the Smarter Balanced item pool. Building on the continuing process of developing item/task specifications and test blueprints, the Consortium uses an iterative process for creating and revising each item as well as the collection of items” (*Smarter Balanced 2017–2018 Technical Report*, 2018, p. 4–17).

8.1.1.3 CAT Distribution of Item Types

The M-STEP distribution of item types is shown in Tables 8-2 and 8-3.

Table 8-2. Distribution of ELA Item Types by Grade and Claim

Grade	Claim	MC	MS	EBSR	MI	HTQ	PBW	Total
3	1	167	48	47		54		316
3	2	119	57			53	3	232
3	3	78	39	47	20			184
3	4	64	40		4	21		129
3	Total	428	184	94	24	128	3	861
4	1	106	49	48		51		254
4	2	130	48			55	3	236
4	3	89	38	47	21			195
4	4	64	49		2	20		135
4	Total	389	184	95	23	126	3	820
5	1	115	63	54		46		278
5	2	115	62			44	3	224
5	3	65	27	37	16			145
5	4	56	45			27		128
5	Total	351	197	91	16	117	3	775
6	1	78	50	38		59		225
6	2	91	68			56	3	218
6	3	77	24	41	18			160
6	4	67	58			18		143
6	Total	313	200	79	18	133	3	746
7	1	85	45	36		44		210
7	2	86	62			51	3	202
7	3	68	31	43	14			156
7	4	30	28		3	34		95
7	Total	269	166	79	17	129	3	663

Table 8-3. Distribution of Mathematics Item Types by Grade and Claim

Grade	Claim	MC	MS	MI	EQ	GI	TI	Total
3	1	124	4	77	499	68	33	805
3	2	14	5	6	54	16		95
3	3	85	33	22	12	47		199
3	4	31	12	8	47	16		114
3	Total	254	54	113	612	147	33	1213
4	1	109		191	452	58	14	824
4	2	30	4	7	58	6	2	107
4	3	65	33	19	22	74		213
4	4	60	10	4	32	13	6	125
4	Total	264	47	221	564	151	22	1269
5	1	233	1	88	436	45		803
5	2	8	2	2	63	11	2	88
5	3	85	24	19	15	54	2	199
5	4	27	6	6	32	29	4	104
5	Total	353	33	115	546	139	8	1194
6	1	71	132	102	362	66	14	747
6	2	7	11	3	49	12	1	83
6	3	46	41	31	18	46		182
6	4	9	12	3	48	12	4	88
6	Total	133	196	139	477	136	19	1100
7	1	56	129	72	370	32		659
7	2	7	11	6	53	7	2	86
7	3	34	32	17	15	40		138
7	4	14	10	2	46	15	1	88
7	Total	111	182	97	484	94	3	971

8.1.1.4 Item Pool Calibration and Model Fit Evaluation

Item parameters contained in ELA and mathematics tests were estimated using a marginal maximum-likelihood procedure with either the 2-parameter logistic (2PL) model for MC items or the generalized partial credit model (Muraki, 1992) for technology-enhanced (TE) items administered after the 2013–14 Smarter Balanced field-test administration. Additionally, for model fit, the evaluation of goodness-of-fit used the likelihood ratio test in PARSCALE (Muraki & Bock, 2003).

For details on item calibration and model fit for ELA and mathematics, please refer to Chapter 9 of the Smarter Balanced 2013–2014 Technical Report (2015), which was published on the Smarter Balanced website.[1]

With the exception of the PBW prompts, Smarter Balanced ELA and mathematics operational item parameters were used to score Michigan students who took ELA and mathematics assessments.

To place the PBW prompts on the same scale as the Smarter Balanced item pool that is used for M-STEP ELA, DRC used a common-item, non-equivalent groups design to link the PBW prompts to the established scale. After the initial IRT item calibration, item parameters were linked to the existing M-STEP scale using the Stocking & Lord (1983) equating procedure, where all other items in the pool were used as anchor items.

8.1.2 Item Pool IRT Statistics

The distributions of item parameters by grade and claim are shown in Tables 8-4 and 8-5. Item difficulty is represented by the b-parameter, and discrimination is represented by the a-parameter. Note that there is a wide range of difficulty in each category.

[1] http://www.smarterbalanced.org/wp-content/uploads/2015/08/Chapter-9-Field-Test-IRT.pdf

Table 8-4. Distribution of Item Difficulty (b-parameter) and Discrimination (a-parameter) for ELA

Grade	Claim	*N* Items	Difficulty Mean	Difficulty Min	Difficulty Max	Discrimination Mean
3	1	316	-0.54	-2.60	4.69	0.70
3	2	232	-0.82	-2.90	4.12	0.69
3	3	184	-0.17	-2.92	3.82	0.54
3	4	129	-0.38	-2.22	1.70	0.68
3	Total	861	-0.51	-2.92	4.69	0.66
4	1	254	0.29	-2.53	6.23	0.62
4	2	236	-0.39	-3.25	2.94	0.59
4	3	195	0.03	-2.82	4.25	0.55
4	4	135	0.41	-2.00	3.73	0.56
4	Total	820	0.05	-3.25	6.23	0.59
5	1	278	0.67	-1.78	5.65	0.61
5	2	224	0.02	-2.28	3.29	0.60
5	3	145	0.53	-2.40	3.48	0.52
5	4	128	0.45	-1.49	3.83	0.67
5	Total	775	0.42	-2.40	5.65	0.60
6	1	225	1.00	-1.64	4.78	0.58
6	2	218	0.86	-2.72	5.54	0.54
6	3	160	0.83	-1.50	7.38	0.50
6	4	143	0.96	-1.30	3.61	0.56
6	Total	746	0.91	-2.72	7.38	0.55
7	1	210	1.23	-1.84	6.63	0.55
7	2	202	1.13	-2.02	5.31	0.51
7	3	156	0.89	-1.71	5.88	0.50
7	4	95	1.71	-0.82	5.61	0.52
7	Total	663	1.19	-2.02	6.63	0.52

Table 8-5. Distribution of Item Difficulty (b-parameter) and Discrimination (a-parameter) for Mathematics

Grade	Claim	*N* Items	Difficulty Mean	Difficulty Min	Difficulty Max	Discrimination Mean
3	1	805	-1.14	-4.34	4.16	0.84
3	2	95	-0.31	-2.54	1.38	1.00
3	3	199	-0.12	-2.42	5.12	0.72
3	4	114	-0.10	-2.68	3.20	0.80
3	Total	1213	-0.81	-4.34	5.12	0.83
4	1	824	-0.30	-3.26	4.48	0.85
4	2	107	0.21	-2.25	2.57	0.89
4	3	213	0.25	-2.08	5.18	0.74
4	4	125	0.25	-2.15	3.28	0.69
4	Total	1269	-0.11	-3.26	5.18	0.82
5	1	803	0.30	-2.79	3.61	0.77
5	2	88	1.07	-1.27	3.94	0.95
5	3	199	0.77	-1.90	5.28	0.66
5	4	104	1.15	-1.23	4.63	0.68
5	Total	1194	0.51	-2.79	5.28	0.76
6	1	747	0.83	-2.85	9.16	0.69
6	2	83	1.55	-2.98	5.50	0.78
6	3	182	1.86	-2.16	8.75	0.58
6	4	88	1.85	-0.71	6.44	0.78
6	Total	1100	1.13	-2.98	9.16	0.69
7	1	659	1.72	-1.79	7.80	0.73
7	2	86	2.05	-1.09	5.07	0.82
7	3	138	2.25	-1.65	6.59	0.56
7	4	88	2.18	-0.79	4.78	0.71
7	Total	971	1.87	-1.79	7.80	0.71

It is also beneficial to examine the distribution of item difficulty compared to the distribution of abilities across the student population. This can be used to ensure that the item pool is deep enough to measure the abilities of the student population without item exposure rates being too high. Figures 8-1 and 8-2 show the comparison of item difficulty, student scores, and cut scores for ELA and mathematics, respectively. For most grades, the item pool has good alignment with the student ability distribution. However, in grades 6 to 8 for mathematics, the item pool appears to be more difficult when compared to the corresponding student ability distribution.

Figure 8-1. ELA Item Pool Difficulty in Comparison to the Student Ability Distribution

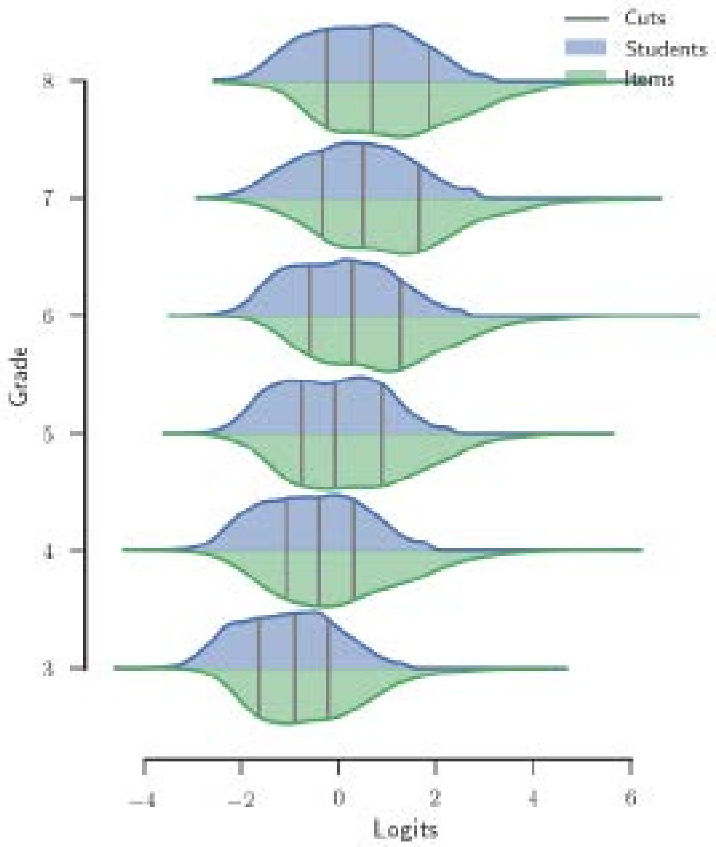
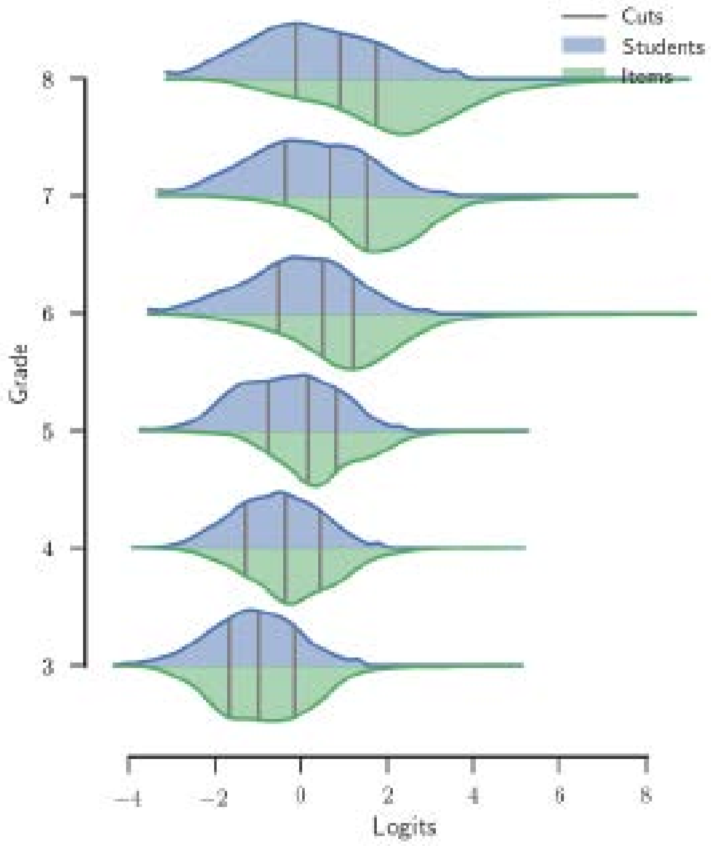


Figure 8-2. Mathematics Item Pool Difficulty in Comparison to the Student Ability Distribution



8.2 Operational CAT ELA and Mathematics Implementation

8.2.1 The Scale

The scales on which M-STEP ELA and mathematics scale scores are reported were established by Smarter Balanced after the 2014 field test. The underlying scales are not unique to Michigan but have been adapted by several states that were members of the Smarter Balanced Assessment Consortium. Michigan has used the underlying scale to create state-specific M-STEP scales used solely by Michigan.

The Smarter Balanced ELA and mathematics scores are reported on vertical scales, sometimes referred to as growth scales, showing student progress from grade to grade. For details on ELA and mathematics vertical scale development, refer to Chapter 9 of the *Smarter Balanced 2013–2014 Technical Report* (2015), which is posted on the Smarter Balanced website.[2] However, the scale scores reported by Michigan should not be considered vertical scale scores.

Additional information regarding M-STEP scale scores can be found in Chapter 10.

8.2.2 Lowest and Highest Obtainable Scale Scores (LOSS and HOSS)

A maximum likelihood procedure cannot produce scale-score estimates for students with perfect scores or scores below the level expected by guessing. In addition, although maximum likelihood estimates are available for students with extreme scores other than zero or perfect, occasionally these estimates have standard errors of measurement that are very large and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure. These values, which are set separately by grade, are called the lowest obtainable theta (LOT) and the highest obtainable theta (HOT). For reporting purposes, the LOT and HOT are transformed, using a linear transformation, to the LOSS and the HOSS. For more information on the LOSS and HOSS, see Chapter 10 and Table 10-1.

8.2.3 Item-Pattern Scoring

M-STEP scale scores are derived using item-pattern scoring; thus, these scale scores are based on the student’s responses to all items on a given test and account for the characteristics of the test items, such as item difficulty. A scale score can be interpreted as a highly probable estimate of a student’s ability in a given content area.

Using item-pattern scoring, a student’s scale score is based on the student’s response to each item (i.e., his or her item-response vector). Each item uses optimal item weights in terms of item information, meaning that items do not contribute equally to the overall scale score. Students with the same raw score may be assigned to different scale scores, depending on which items they answered correctly.

[2] http://www.smarterbalanced.org/wp-content/uploads/2015/08/Chapter-9-Field-Test-IRT.pdf

8.2.4 Blueprint Fidelity Summary

The *Smarter Balanced 2017–2018 Technical Report* notes that “A key design document of the summative assessments is the test blueprint, which specifies the number and nature of items to be administered” (p. 6–9).Chapter 6 of that same report states that “The analyses showed that the operational tests delivered in the 2016–2017 administration fulfilled the blueprint requirements very well” (p. 6–9).

For M-STEP ELA and mathematics implementation, a review of the blueprint fulfillment was completed for both the simulation (see Chapter 6 of this report) and the OP tests.

8.3 Operational Analysis of Social Studies

This section describes analyses conducted for social studies and reports corresponding results. As mentioned above, item/test analyses from the CTT and IRT frameworks have been carried out. They are reported below separately. If the IRT models fit the empirical item-response data for the population for which generalizations are made (i.e., Michigan students), then it is likely that the scores are valid indicators of an underlying ability.

8.3.1 CTT Statistics Social Studies

This section presents test-level summary statistics for each form and grade of social studies. This is followed by item-level statistics for each form and grade of social studies. These statistics were produced using census data.

8.3.1.1 Test-Level Analysis

This section presents the test-level summary statistics for social studies. In addition to the number of students taking the form (N), Table 8-6 provides the following raw score descriptive statistics for a given grade and form: mean, standard deviation (SD), minimum (Min), and maximum (Max).

Table 8-6. Test-Level Descriptive Statistics by Form: Social Studies Raw Score Distribution

Grade	N OP Items	Form	N	Mean	SD	Min	Max
5	45	1	34846	22.10	8.09	2.00	45.00
5	45	2	34862	22.23	8.03	1.00	45.00
5	45	3	34848	22.07	8.01	4.00	45.00
5	45	4	767	19.38	7.54	4.00	43.00
8	44	1	35945	21.86	8.59	2.00	44.00
8	44	2	35897	21.86	8.62	1.00	44.00
8	44	3	35830	21.67	8.54	2.00	44.00
8	44	4	394	17.87	7.47	4.00	41.00
11	38	1	34266	20.71	8.18	1.00	38.00
11	38	2	34220	20.72	8.17	0.00	38.00
11	38	3	34240	20.67	8.23	0.00	38.00
11	38	4	621	19.23	8.17	4.00	38.00

8.3.1.2 Item-Level Analysis

This section presents various item-level statistics for all OP items[3] on the spring 2019 M-STEP social studies tests. Specifically, item difficulty and adjusted item-total correlations defined by the CTT are reported here.

Since all items on the spring 2019 M-STEP social studies tests are dichotomously scored, the *p*-value is computed as an indicator for item difficulty. The *p*-value equals the proportion of students who answer an item correctly. A high *p*-value means that an item is easy, and a low *p*-value means that an item is difficult.

The adjusted item-total correlation is an index of the association between students' performance on an item and their performance on the test as a whole; however, the item of interest is excluded from the total raw score. A high adjusted item-total correlation is desired, as high correlations indicate that students with high scores on all other test items (i.e., students with high ability) tend to get a correct answer on the item under consideration and that the students with low scores on all other test items (i.e., students with low ability) tend to get this specific item incorrect. Since all items are dichotomously scored, the adjusted point biserial correlation is computed.

The item-level descriptive statistics (by grade and form) for all OP items are presented in Tables 8-7 and 8-8 for social studies. For each grade, forms 1 through 3 are online forms and form 4 is the paper/pencil form.[4] All online forms for social studies have the same set of OP

[3] All statistics for field-test items are excluded from this report.

[4] One emergency form and one braille form per grade were also created. However, responses from these forms are excluded from any analysis due to negligible occurrences (for braille forms, which are exactly the same as the corresponding paper/pencil forms) and a different calibration approach (for emergency forms, banked values from the item pool would be activated for scale-score computation). In 2019, no emergency forms were used at all.

items; therefore, forms 1 through 3 are reported together per grade in Tables 8-7 and 8-8. Each paper/pencil form for social studies per grade has a few different OP items from its online counterpart because the TE items could not be presented on paper/pencil forms. As shown in Tables 8-7 and 8-8, for both item difficulty (*p*-value) and adjusted item-total correlation (adjusted point biserial), the following descriptive statistics are reported: number of OP items (N OP Items), mean, SD, minimum (Min), and maximum (Max), for a given grade by form.

Table 8-7. Item-Level Descriptive Statistics by Form: Social Studies *P*-Value

Grade	N OP Items	Form	Mean	SD	Min	Max
5	45	1–3	0.49	0.12	0.29	0.85
5	45	4	0.43	0.11	0.25	0.77
8	44	1–3	0.50	0.10	0.32	0.76
8	44	4	0.41	0.08	0.25	0.58
11	38	1–3	0.54	0.12	0.21	0.75
11	38	4	0.51	0.10	0.30	0.68

Table 8-8. Item-Level Descriptive Statistics by Form: Social Studies Adjusted Point Biserial

Grade	N OP Items	Form	Mean	SD	Min	Max
5	45	1–3	0.31	0.06	0.19	0.44
5	45	4	0.29	0.06	0.10	0.42
8	44	1–3	0.35	0.09	0.21	0.54
8	44	4	0.29	0.09	0.09	0.50
11	38	1–3	0.40	0.08	0.17	0.53
11	38	4	0.39	0.09	0.17	0.51

8.3.2 IRT Statistics: Social Studies

The unidimensional 2PL model is used for M-STEP social studies at each grade level, as all items are dichotomously scored. For this model, the probability that person j answers item i correctly is defined as follows (adapted from Embretson & Reise, 2000, p. 70):

$$P\left(X_{ij}=1\middle|\theta_j,b_i,a_i\right)=\frac{\exp\left(a_i\left(\theta_j-b_i\right)\right)}{1+\exp\left(a_i\left(\theta_j-b_i\right)\right)}, \quad (8.1)$$

where θ_j, b_i, and a_i represent person j's ability, item i's difficulty, and item i's discrimination, respectively. Note that $P\left(X_{ij}=1\middle|\theta_j,b_i,a_i\right)$ is referred to as for simplicity $P_i(\theta)$ hereafter.

8.3.3 Item Calibration for Social Studies

The common-item nonequivalent groups design (Kolen & Brennan, 2004) and the fixed item parameter calibration approach were used to put all items onto the base scale. The IRT software used was flexMIRT (Cai, 2017). An outline of the annual calibration, equating, and scaling procedures for M-STEP social studies is presented below:

- A free run was carried out using flexMIRT (i.e., a 2PL model) with all online OP items for each grade.
- After each free run, the obtained item parameters for anchor items were compared with their banked values (all banked values have already been put onto the base scale). In addition to checking the scatterplots of item difficulty parameters and item-discrimination parameters, a simple linear regression was carried out using free run results as outcomes (i.e., one regression for item difficulty parameters and one regression for item-discrimination parameters) and the corresponding banked values as the predictor. Model diagnostics analyses focusing on finding unusual points were then carried out, which included checking leverage points, outliers, and influential observations. The OEAA psychometrician then made the anchor item inclusion/exclusion decisions and shared with National Center for Research on Evaluation, *Standards*, and Student Testing (CRESST)[5], which functions as an independent party validating the psychometric work done by the OEAA psychometrician on M-STEP social studies.
- CRESST psychometricians conducted their independent anchor item stability check (with their own methods) and compared their conclusion with what the OEAA psychometrician obtained.
- After the OEAA psychometrician and CRESST agreed on the anchor item inclusion/exclusion decisions, the OEAA psychometrician carried out a mean/mean method to transform the parameters from free run to the base scale for all anchor items. The constants A and B are computed as follows (adapted from formulas presented in Kolen & Brennan, 2004, p. 163):

$$A = \frac{\bar{a}_{free}}{\bar{a}_{base}},$$

$$B = \bar{b}_{base} - A * \bar{b}_{free} \quad (8.2)$$

where $\bar{a}_{free} =$ mean of anchor set item discrimination parameters from the free run,

$\bar{a}_{base} =$ mean of anchor set item discrimination parameters from the bank,

$\bar{b}_{free} =$ mean of anchor set item difficulty parameters from the free run,

$\bar{b}_{base} =$ mean of anchor set item difficulty parameters from the bank.

[5] For details of CRESST psychometricians' verification work on M-STEP Social Studies, please see appendix G

After obtaining the constants A and B as mentioned above, the following formulas are used to transform all anchor item parameters onto the base scale (adapted from formulas presented in Kolen & Brennan, 2004, p. 162):

$a_{equated} = \frac{a_{free}}{A}$,

$b_{equated} = Ab_{free} + B$, (8.3)

where $\bar{a}_{free} =$ item discrimination parameter from the free run for an anchor item,

$a_{equated} =$ transformed item discrimination parameter for that anchor item,

$\bar{b}_{free} =$ item difficulty parameter from the free run for an anchor item,

$b_{equated} =$ transformed item difficulty parameter for that anchor item.

- A validation check is then carried out by CRESST psychometricians to confirm the transformed item difficulty and item-discrimination parameters. After the anchor item values are verified per grade, a fixed item parameter calibration was used to put all OP items onto the base scale.
- Summed score to Expected *A Posteriori* (EAP) conversion tables were then created using flexMIRT. For social studies, one conversion table was used for all online forms as all OP items are the same across all forms at each grade level. Note that in each year, when conversion tables are made, paper/pencil data are not available for equating and calibration; thus, the online form conversion tables are applied to the paper/pencil forms.
- When the final data (both online and paper/pencil data) are available later in the year, a fixed item parameter calibration (where all online OP items are fixed at the values found during the conversion table creation stage) is carried out once again using the final data for all OP items. Then the obtained OP items' parameters are used in a fixed item parameter calibration to put all field-test items onto the same scale.

8.3.4 Anchor Item Evaluation for Social Studies

There are various methods for evaluating anchor item stability. As mentioned above, model diagnostic analyses were used by the OEAA psychometrician in checking the stability of anchor items at the conversion table creation stage. In this section, an ad hoc approach is reported, which evaluates the anchor quality using anchor item response patterns. This method uses all possible information about student performance that is shared between the 2018 and 2019 online[6] administrations of M-STEP social studies tests. The annually used evaluation steps are as follows:

- Obtain the item response patterns in both the 2018 and the 2019 online administrations for the anchor items used in 2019. Note that only the same response patterns appearing in both years are used for this evaluation.

[6] This analysis limits its scale to online responses only because the equating procedures are carried out with online data only.

- Aggregate these item response patterns to obtain the number of unique item response patterns per grade as well as the mean scale score for each specific item response pattern in 2018 and 2019.
- Plot the mean scale score in 2019 against the mean scale score in 2018 by grade for each anchor item response pattern.
- Plot a 45-degree line on that scatterplot. The observations plotted should cluster relatively tightly and be randomly distributed around the 45-degree line.
- Plot the “proficient” cut score on both the vertical and horizontal axes to divide the graph into four quadrants (i.e., item response patterns that are scored proficient in both years, those that are scored proficient in 2018 but not 2019 and vice versa, and those that are scored not proficient in both years).

The final steps in the analysis are to evaluate the degree to which the scatterplot for each grade deviates from expectations for good equating (i.e., deviation from tight clustering and random distribution around the 45-degree line) and to evaluate the distribution of item score patterns in the four quadrants by grade.

Table 8-9 presents the anchor points (same as the number of anchor items) per grade on each form. The results of the anchor quality evaluation are presented in Figures 8-3 to 8-5 and in Table 8-10. As shown in Figures 8-3 to 8-5, the points plotted on the scatterplot for each grade tend to lie along the 45-degree line, indicating that the majority of students who shared the same item response patterns on the anchor set also obtained similar mean scale scores (per item response pattern) across the two years. Therefore, these anchor items are considered to be stable across these two years.

Table 8-9. Number of Anchor Items by Content and Grade for Each Form

Content Area	Grade	Total Points	Anchor Points	Percentage of Anchor Points
Social Studies	5	45	12	27
Social Studies	8	44	11	25
Social Studies	11	38	10	26

Figure 8-3. Social Studies Grade 5

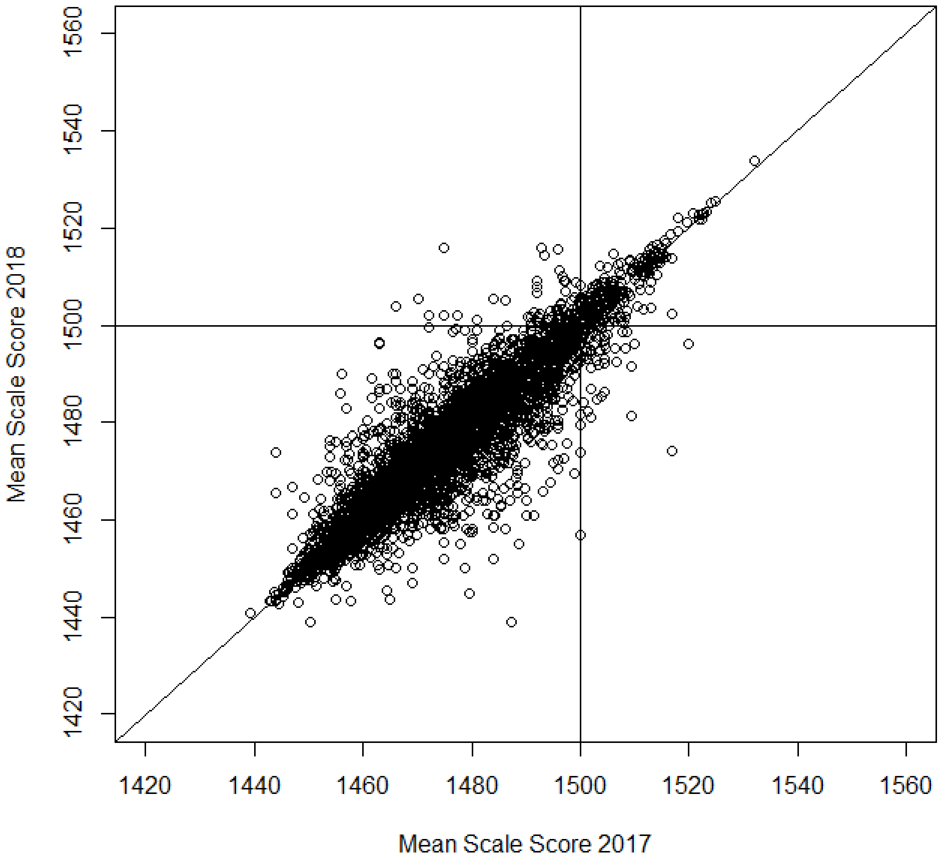


Figure 8-4. Social Studies Grade 8

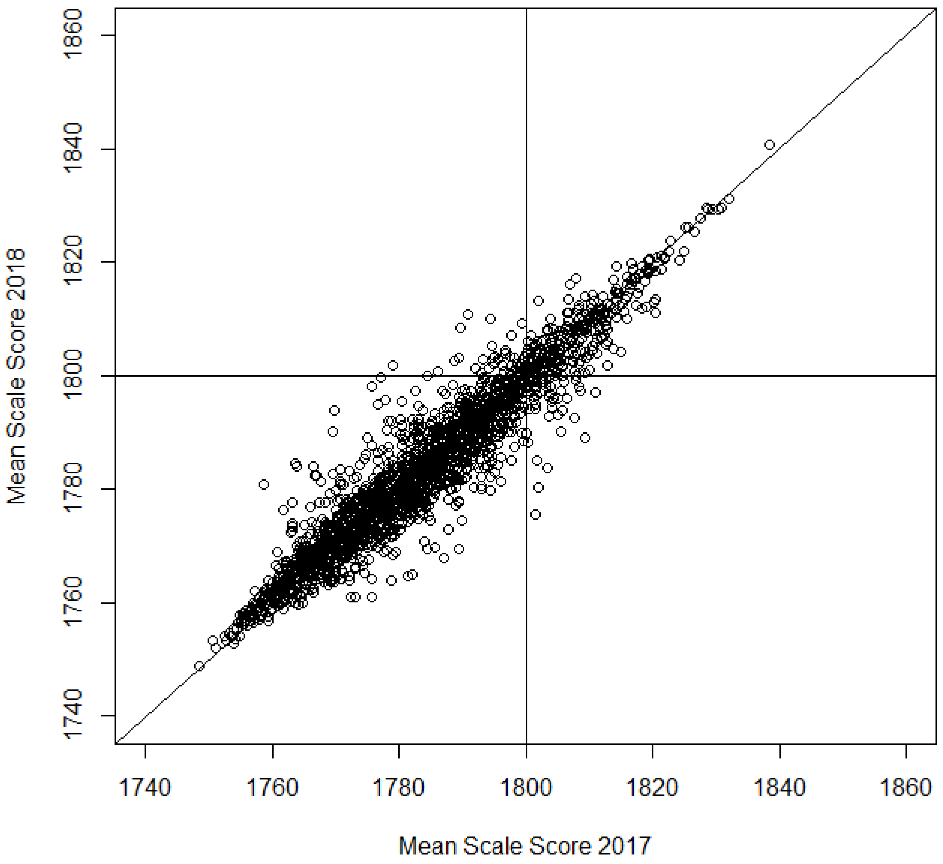
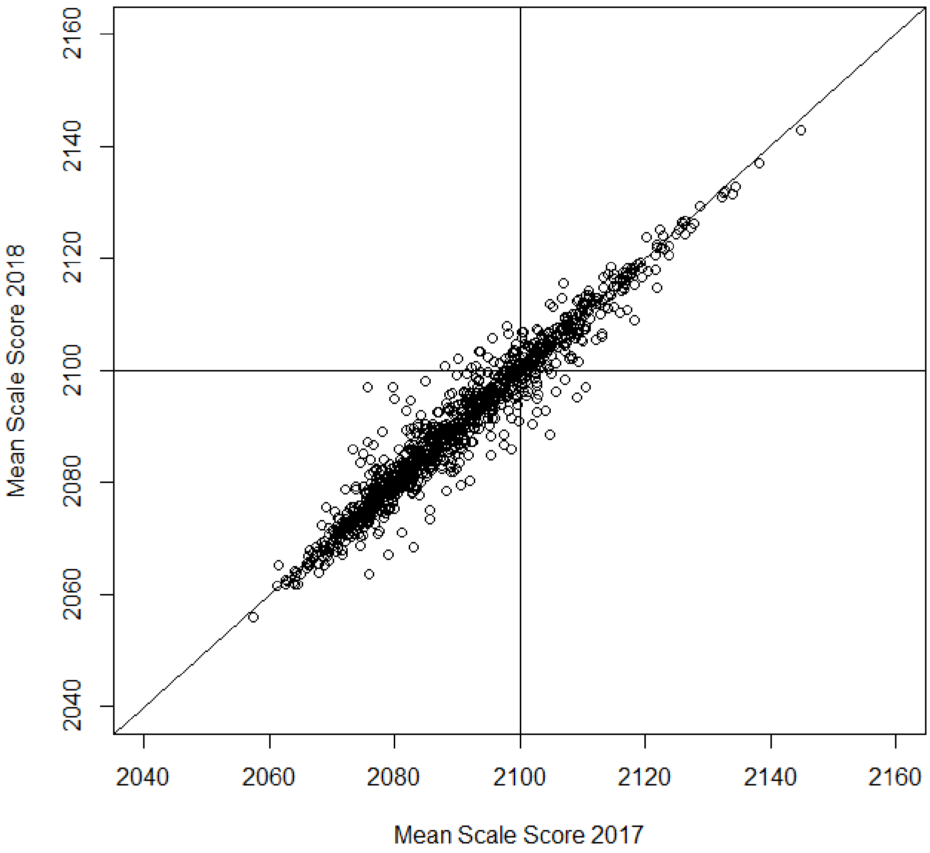


Figure 8-5. Social Studies Grade 11



The number and percentage of these anchor item response patterns that fall into each of the four quadrants by grade are summarized in Table 8-10. The percentage of response patterns that are associated with consistent performance categorization (based on the mean scale score for each item response pattern) across the two administrations ranged from 94.48% to 96.72%. According to this table, grade 5 had the highest consistency rate (96.72%), while grade 8 had the lowest consistency rate of about 94.48%.

Table 8-10. Evaluation of Equating Quality Using Linking Item Response Patterns

Content Area	Grade	Item Response Pattern	Proficient in Both Years	Not Proficient in Both Years	Proficient in 2018 Only	Proficient in 2019 Only	Consistent Classification	Inconsistent Classification
Social Studies	5	Count	241	3650	61	71	3891	132
Social Studies	5	Percentage	5.99	90.73	1.52	1.76	96.72	3.28
Social Studies	8	Count	267	1665	62	51	1932	113
Social Studies	8	Percentage	13.06	81.42	3.03	2.49	94.48	5.52
Social Studies	11	Count	237	734	29	22	971	51
Social Studies	11	Percentage	23.19	71.82	2.84	2.15	95.01	4.99

Note. Some rows may have percentages that sum to more than 100 due to rounding.

8.3.5 Evidence of Model Fit for Social Studies

Due to sparse contingency tables, the limited-information fit statistics M_2 (Cai & Hansen, 2013) of the fitted model were requested for each fixed item parameter calibration run in flexMIRT. Due to the large sample size (>34,000 per online form), the model selection index tends to prefer more complex models (Cudeck & Henly, 1991). Taking model parsimony into considerations, the RMSEA values are considered rather than the M_2 statistics. The RMSEA values are 0.01 for social studies at grade 5, and 0.02 for social studies at grades 8 and 11. The fact that the RMSEA values are small in magnitude (i.e., close to 0) is evidence to support the use of the 2PL fixed item parameter calibration.

8.3.6 Test Characteristic Curves (TCCs) and Conversion Tables

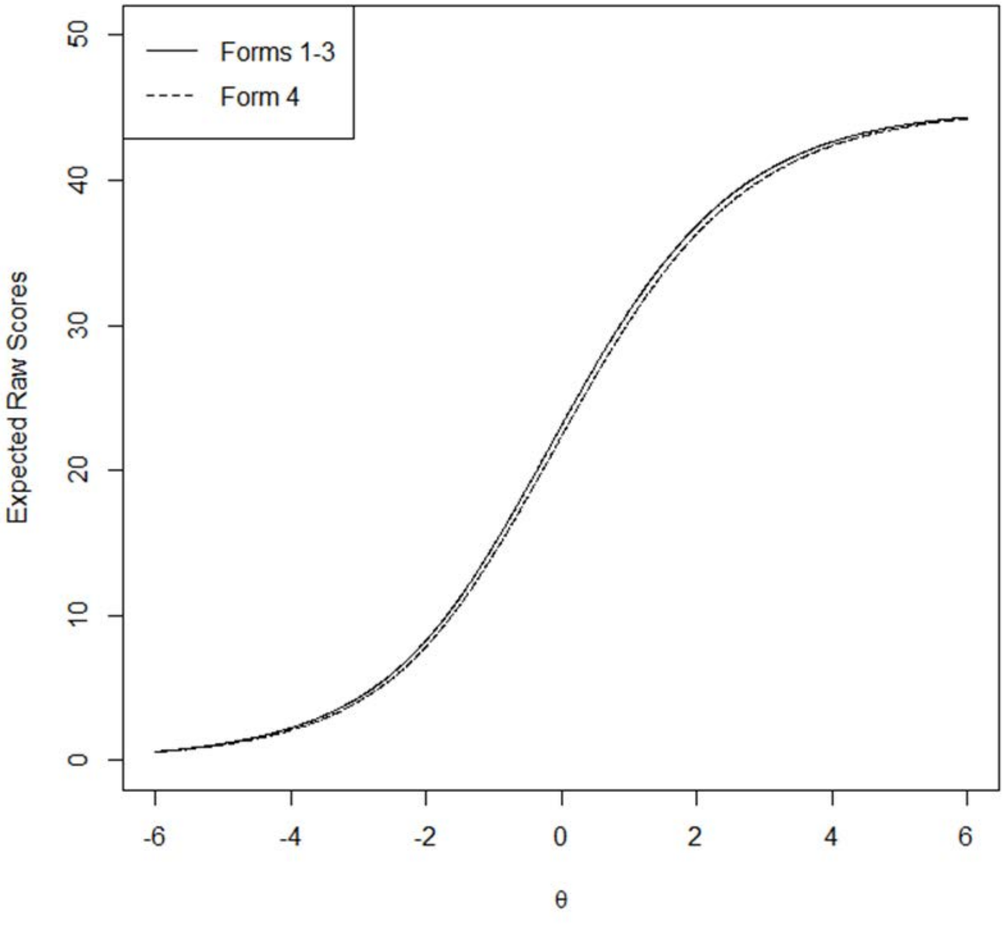
The TCC is the graphical representation of the test characteristic function (TCF), and TCF is the expected raw total score given θ. Since all items are dichotomously scored, the expression of TCF is as follows (adapted from Yen & Fitzpatrick, 2006, p. 125):

$$E(X.|\theta) = \sum_{i=1}^{n} E(X_i|\theta) = \sum_{i=1}^{n} P_i(\theta) \quad (8.4)$$

Figures 8-6 to 8-8 display the TCCs for the spring 2019 M-STEP social studies tests by grade. These graphs were made using the item parameter estimates obtained from the post-administration calibration in 2018 (based on unidimensional 2PL models). Two TCCs are shown for social studies at each grade level (one for online forms 1–3 and one for paper/pencil form 4) (see Figures 8-6 to 8-8). Note that these curves were created for OP items per form based on the item parameter estimates obtained from the last step mentioned in Section 8.3.3. Due to item differences between online forms and paper/pencil forms (i.e., TE items that appear on online forms cannot appear on paper/pencil forms), slight differences in TCCs can be seen. In general, for each grade, the TCCs across all forms are very close to each other.

Table 8-11 presents the summed scores to EAP conversion tables by grade for social studies. These are used for operational reporting. Note that when conversion tables were made, no paper/pencil data were available for calibration; thus, an operational decision was made to apply the conversion tables from the online forms to the corresponding paper/pencil forms.[7] Note that these tables present very similar results as those shown in the corresponding TCC graphs.

Figure 8-6. TCC for Social Studies Grade 5 by Form



[7] There are very few students taking social studies paper/pencil forms at each grade level (i.e., <0.8%) in 2019, and the number is expected to keep decreasing in the future years to come. Since all past investigations concluded such decision (i.e., apply the conversion tables from the online forms to the paper/pencil forms at each grade) as reasonable (e.g., see mode comparison studies in the 2017 and 2018 M-STEP technical reports), mode comparison related investigations will be skipped starting from the 2019 administration.

Figure 8-7. TCC for Social Studies Grade 8 by Form

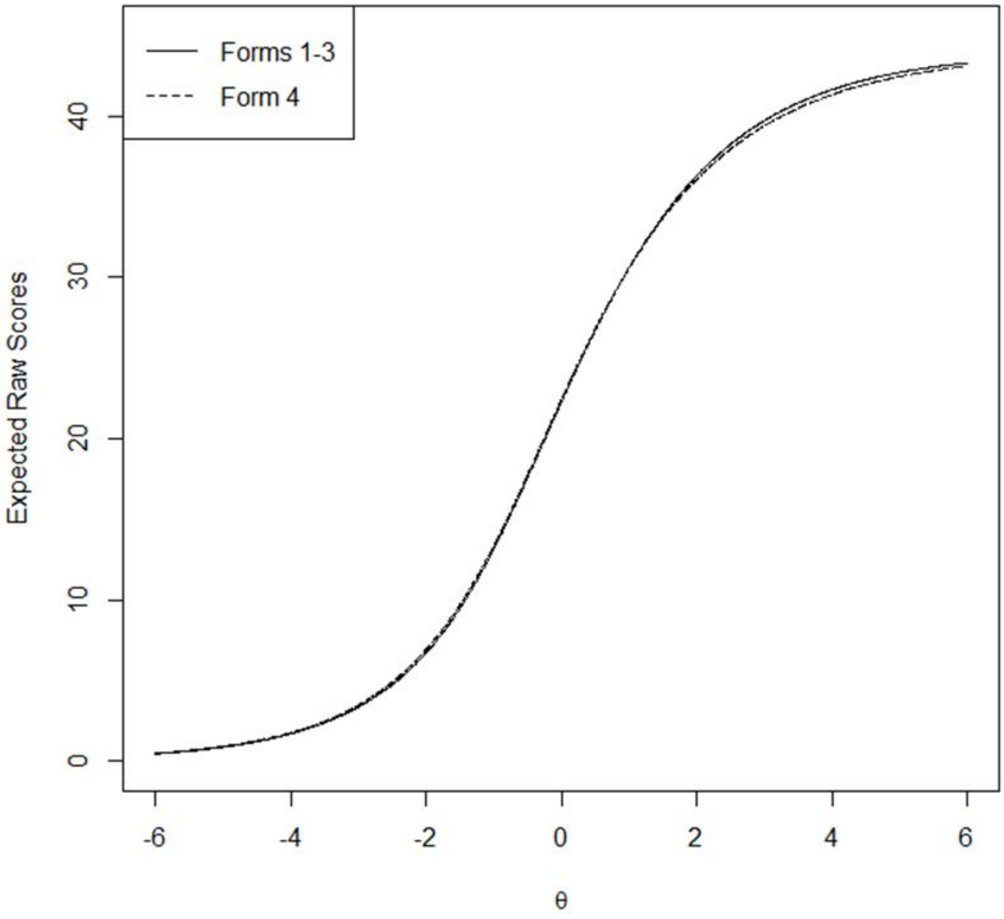


Figure 8-8. TCC for Social Studies Grade 11 by Form

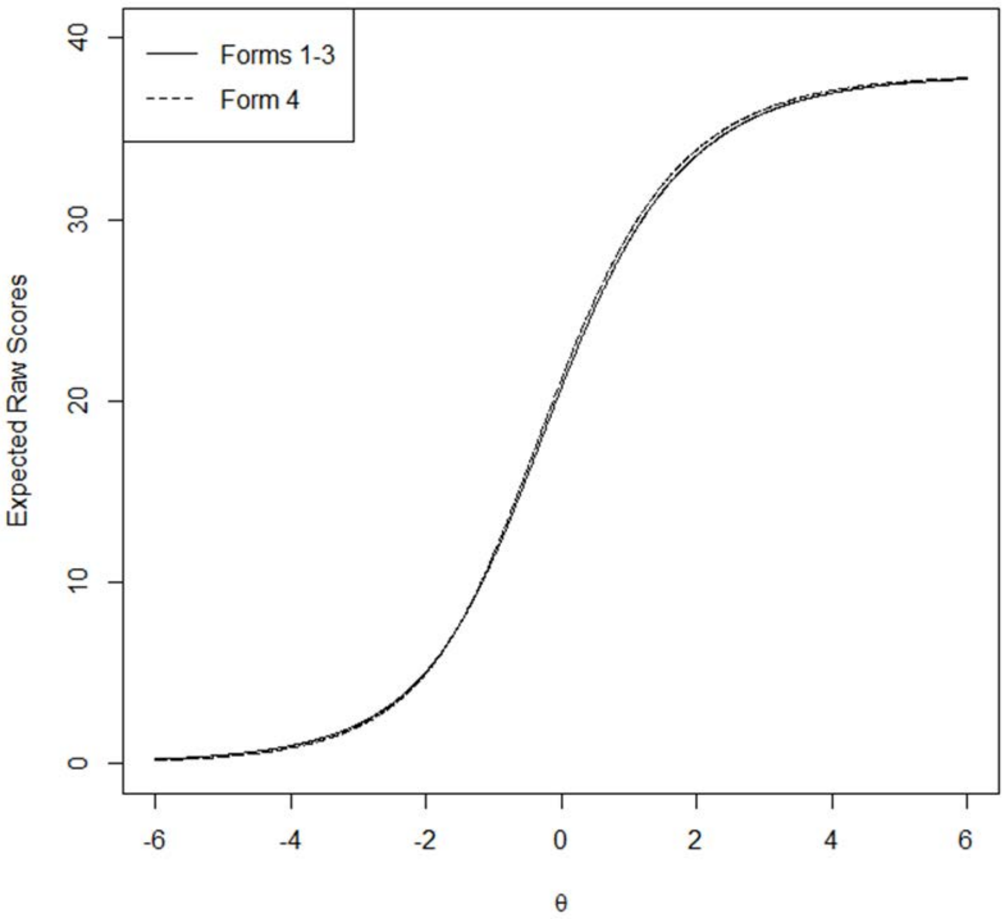


Table 8-11. Social Studies Summed Score to EAP Conversion Tables by Grade

Raw Score	Grade 5 Theta	Grade 5 SE	Grade 8 Theta	Grade 8 SE	Grade 11 Theta	Grade 11 SE
0	-3.0560	0.4890	-2.8990	0.5120	-2.7310	0.5110
1	-2.8750	0.4990	-2.7000	0.5120	-2.5020	0.4950
2	-2.6880	0.4970	-2.5030	0.5000	-2.2840	0.4690
3	-2.5030	0.4860	-2.3130	0.4810	-2.0820	0.4410
4	-2.3250	0.4700	-2.1350	0.4600	-1.8990	0.4140
5	-2.1570	0.4540	-1.9680	0.4400	-1.7320	0.3920
6	-1.9970	0.4390	-1.8120	0.4210	-1.5790	0.3720
7	-1.8450	0.4250	-1.6660	0.4050	-1.4380	0.3560
8	-1.7010	0.4130	-1.5290	0.3910	-1.3070	0.3430
9	-1.5640	0.4030	-1.3980	0.3780	-1.1830	0.3310
10	-1.4320	0.3940	-1.2740	0.3670	-1.0660	0.3220
11	-1.3050	0.3870	-1.1550	0.3580	-0.9540	0.3140

Raw Score	Grade 5 Theta	Grade 5 SE	Grade 8 Theta	Grade 8 SE	Grade 11 Theta	Grade 11 SE
12	-1.1830	0.3800	-1.0400	0.3500	-0.8470	0.3080
13	-1.0650	0.3740	-0.9300	0.3430	-0.7420	0.3030
14	-0.9490	0.3700	-0.8220	0.3380	-0.6410	0.3000
15	-0.8370	0.3660	-0.7180	0.3340	-0.5420	0.2970
16	-0.7270	0.3630	-0.6160	0.3310	-0.4440	0.2960
17	-0.6180	0.3610	-0.5150	0.3280	-0.3470	0.2950
18	-0.5110	0.3590	-0.4160	0.3270	-0.2500	0.2950
19	-0.4060	0.3580	-0.3170	0.3270	-0.1530	0.2960
20	-0.3010	0.3570	-0.2190	0.3270	-0.0550	0.2980
21	-0.1960	0.3570	-0.1210	0.3280	0.0430	0.3010
22	-0.0920	0.3570	-0.0220	0.3300	0.1440	0.3050
23	0.0120	0.3580	0.0760	0.3330	0.2460	0.3090
24	0.1170	0.3600	0.1760	0.3360	0.3510	0.3150
25	0.2220	0.3610	0.2770	0.3400	0.4590	0.3220
26	0.3290	0.3640	0.3800	0.3440	0.5720	0.3290
27	0.4360	0.3670	0.4850	0.3500	0.6890	0.3390
28	0.5450	0.3700	0.5920	0.3560	0.8110	0.3490
29	0.6560	0.3750	0.7020	0.3620	0.9400	0.3610
30	0.7690	0.3790	0.8150	0.3700	1.0770	0.3750
31	0.8850	0.3850	0.9320	0.3780	1.2220	0.3910
32	1.0030	0.3910	1.0520	0.3870	1.3780	0.4090
33	1.1260	0.3980	1.1780	0.3980	1.5460	0.4290
34	1.2520	0.4060	1.3080	0.4090	1.7290	0.4520
35	1.3830	0.4150	1.4440	0.4210	1.9270	0.4780
36	1.5190	0.4250	1.5860	0.4350	2.1440	0.5050
37	1.6610	0.4360	1.7360	0.4490	2.3780	0.5300
38	1.8090	0.4490	1.8930	0.4650	2.6250	0.5460
39	1.9660	0.4630	2.0590	0.4820		
40	2.1310	0.4770	2.2340	0.4990		
41	2.3040	0.4930	2.4170	0.5140		
42	2.4870	0.5060	2.6070	0.5250		
43	2.6760	0.5150	2.7990	0.5270		
44	2.8650	0.5150	2.9840	0.5170		
45	3.0470	0.5020				

Note. The possible maximum total raw score is 45 for grade 5, 44 for grade 8, and 38 for grade 11.

8.3.7 IRT Statistics

As discussed above, the 2PL model was used to calibrate the spring 2019 social studies items at each grade level. A summary (i.e., minimum, maximum, and mean values) of the item difficulty (b-parameter) and item discrimination (a-parameter) estimates for all OP items per form for each grade is presented in Table 8-12.

Table 8-12. Item Difficulty (b-Parameter) and Item Discrimination (a-Parameter) for Social Studies by Grade and Form

Grade	Form	Difficulty Minimum	Difficulty Maximum	Difficulty Mean	Discrimination Minimum	Discrimination Maximum	Discrimination Mean
5	1–3	-1.8253	1.4880	-0.0054	0.4660	1.9992	0.8351
5	4	-1.3384	4.1468	0.1938	0.2476	1.5188	0.8255
8	1–3	-1.0009	1.4471	0.0726	0.4679	2.1936	0.9299
8	4	-0.8551	1.7199	0.2289	0.3400	1.8956	0.8352
11	1–3	-0.9544	1.6046	-0.1116	0.3998	1.6736	1.1167
11	4	-0.9544	1.8814	-0.1045	0.3901	1.6994	1.0484

8.4 Summary

In summary, the overall purpose of the OP data analysis is to ensure that the test items, as well as the overall test, are functioning appropriately. It also helps maintain the test scale across years so that test results may be appropriately compared across years. The data analyses undertaken by Smarter Balanced, DRC, and the Michigan Department of Education are in alignment with multiple best practices of the assessment industry, particularly the following *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014):

- Standard 5.2—The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.
- Standard 5.13—When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

Chapter 9: Test Results

This chapter of the Technical Report contains information on the results of the spring 2019 administration of M-STEP along with descriptions of the score reports, data structure, and interpretive guide. The AERA, APA, and NCME (2014) *Standards* addressed in Chapter 9 include 5.1, 6.10, and 7.0. Each standard will be presented in the pertinent section of this chapter.

9.1 Test Completion

The spring 2019 M-STEP was administered to Michigan students in three content areas: ELA, mathematics, and social studies. For the purposes of this technical report, “percent valid” is the percentage of students who received a valid scale score given the total number of students who took the online or paper/pencil test. These test completion rates are summarized in Tables 9-1a through 9-3g.

Test completion information is reported for all students and the following demographic subgroups:

- Gender: Female and Male
- Race/Ethnicity: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, Two or More Races, and White
- Economically Disadvantaged: Yes, No
- English Language Learners: Yes, No
- Students with Disabilities: Yes, No
- Students Used Standard Accommodations: Yes, No

9.2 Current Administration Data Scale Score Summaries

Summaries of the scale-score (SS) descriptive statistics for the spring 2019 administration of the ELA, mathematics, and social studies assessments are reported in Tables 9-4a through 9-6g, by grade, content area, and demographic subgroup.

Additionally, Tables 9-7a through 9-9b present the scale-score descriptive statistics and the performance level percentages by grade for the 2019 M-STEP ELA, mathematics, and social studies tests. These tables provide the scale-score descriptive statistics (i.e., Mean, SD, Min, Max values) and the percentages of students in each performance level: Not Proficient, Partially Proficient, Proficient, and Advanced.

9.3 Description of Reports

Score reports are the primary means of communicating test scores to relevant district and school administrators, teachers, parents, and students. AERA, APA, and NCME (2014) Standard 6.10 states the following:

> When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (p. 119)

Standard 5.1 is also addressed:

> Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (p. 102)

Interpretations related to the test scores are provided in the *Spring 2019 Interpretive Guide to M-STEP Reports* (M-STEP IGTR) for grades 3 through 7, the *Spring 2019 Michigan Grade 8 Testing Interpretive Guide to Reports* (Grade 8 IGTR), and in the *Spring 2019 Interpretive Guide to MME Reports* (MME IGTR) for grade 11. The interpretive guides are provided in Appendix B.1 through Appendix B.3 respectively.

In addition to providing interpretation, it is important that the information is understandable by the target audience. Standard 7.0 of the AERA, APA, & NCME (2014) *Standards* states the following:

> Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (p. 125)

In support of Standard 7.0, the three Interpretive Guides to Reports (presented in Appendices B.1 through B.3) are accessible to parents, teachers, and laypeople alike.

M-STEP score reports comprise student-level reports and data files and aggregate reports and data files. Depending on the audience, reports and data files are available in several systems.

1. The OEAA Secure Site allows authorized school and district users to download student-level and state, district, and building level aggregate data files. It also provides access to the online Dynamic Score Reporting Site.

2. The Dynamic Score Reporting Site (DSRS) provides authorized school and district users access to interactive online student-level and aggregate reports. Reports are available by grade and content area and can be filtered by reporting group (gender, ethnicity, economically disadvantaged, English Learner, former English Learner, foster care, homeless, migrant, military connected, students with disabilities, homeschooled).

3. The Michigan Linked Educational Assessment Reporting Network (MiLearn) provides teachers, parents, and students direct access to student level data through the district's Student Information System. The reports presented are a subset of those available in the DSRS.

4. MiSchoolData is Michigan's public portal to education-related data. It contains assessment data for public consumption.

This technical report will address the data files and reports available through the OEAA Secure Site and the Dynamic Score Reporting Site.

Brief descriptions of the reports and data files are provided below. More extensive descriptions with samples are included in the M-STEP IGTR for grades 3 through 7, the Grade 8 IGTR, and in the MME IGTR for grade 11. The interpretive guides also include information on how to use the data, limitations of the data, and the online functionality associated with each report.

9.3.1 Student-Level Data Reports and Data Files

- The Student Record Labels provide a summary of student performance levels for individual students. The labels include district and school information, student demographic information, M-STEP administration cycle information, and overall student performance level for tested content areas.
  Student Record Labels are provided for inclusion in the student's Cumulative Student Record (CA60) folder. They are printed and shipped to the school in which the student tested in late summer and are available through the Secure Site if the school needs to print additional copies.
  Additional information can be found starting on page 20 of the M-STEP IGTR, page 20 of the Grade 8 IGTR, and page 21 of the MME IGTR.

- The Individual Student Report (ISR) provides information about student performance by content area. Each student will have a separate ISR for each content area assessed. The report is divided into three main sections:
  - Student demographic information
  - Overall content performance and growth reporting for ELA, mathematics, and social studies
  - Detailed claim data for ELA and mathematics and discipline and content expectation data for social studies

  Additional information can be found starting on page 21 of the M-STEP IGTR, page 21 of the Grade 8 IGTR, and page 22 of the MME IGTR. Individual Student Reports are also available to educators, parents, and students in MI-Learn where they are referred to as the M-STEP Student Detail Report.

- Parent Reports are printed and shipped to schools for distribution to parents. The parent report provides information for parents about student performance in tested content areas. This report includes five main sections:
  - Superintendent letter
  - Overall performance level and scale score
  - Claim data for ELA and mathematics and discipline data for social studies

- Definitions for parents
- Performance-level descriptors

Additional information can be found starting on page 24 of the M-STEP IGTR, page 25 of the Grade 8 IGTR, and page 28 of the MME IGTR. Parent Reports are also available in the Dynamic Score Reporting Site for schools to access and print copies.

- Student Roster allows users to view student scale scores, growth, and claim performance data for ELA and mathematics or discipline data for social studies by content area and grade. The data provided are:
  - Overall proficiency summary of the rostered students
  - An alphabetical listing of the selected students
  - Overall content performance, including growth data when available and
  - Subscore data: claim data for ELA and mathematics, discipline data for social studies
  - Passage-based writing raw score data for ELA

  Additional information can be found starting on page 26 of the M-STEP IGTR, page 28 of the Grade 8 IGTR, and page 31 of the MME IGTR. The Student Roster report is also available to educators in MiLearn.

- The Student Overview provides summary information about student performance in all tested content areas in the selected grade. For each selected student, the following data are displayed for each tested content area in both graphical and table format:
  - scale score
  - margin of error
  - performance level
  - claim or discipline performance

  Additional information can be found starting on page 31 of the M-STEP IGTR, page 32 of the Grade 8 IGTR, and page 35 of the MME IGTR. The Student Overview is also available to educators in MiLearn.

- The Student Growth and Proficiency Report provides information about student growth by content area. Each student in grades containing reportable growth data has a separate Student Growth and Proficiency report for each content area test taken. The report is divided into three main sections:
  - scale score
  - margin of error
  - performance level
  - claim or discipline performance

  Additional information can be found starting on page 34 of the M-STEP IGTR, page 34 of the Grade 8 IGTR, and page 37 of the MME IGTR. The Student Overview is also available to educators in MiLearn.

- The Student Data File (SDF) contains student level demographics, test scores, and performance data for each tested content area. The Downloadable SDF contains detailed individual student data in an Excel file. This data includes school information, student demographic data, test administration data, and student performance data. The SDF is provided for schools to use as a data resource for school- or district-level data reviews. Schools or districts can use the SDF to manipulate and evaluate data in ways that support school improvement goals or for other data-based decision-making purposes.

  Additional information can be found starting on page 53 of the M-STEP IGTR, page 48 of the Grade 8 IGTR, and page 49 of the MME IGTR. The SDF can be downloaded from the OEAA Secure Site. The file layout is in Appendix B.4.

9.3.2 Aggregate Data Reports and Data Files

- The Target Analysis Report is available at the school, district, and state levels for ELA and mathematics. The report is intended to provide an overview of relative strengths and weaknesses in ELA and mathematics by assessment target as compared to student performance on the test as a whole.
  Additional information can be found starting on page 39 of the M-STEP IGTR, page 39 of the Grade 8 IGTR, and page 41 of the MME IGTR.

- The Expectation Analysis Report provides the percentage of points earned by grade, the content area expectations in each social studies discipline, and the number of students scoring in each of the four quartiles. The report is intended to provide an overview of performance by content expectation. The report displays the number of students assessed in each expectation (not all students were assessed on every expectation), the average percentage of points earned, and the number of students scoring in one of four bands or quartiles: 0%–25%, 26%–50%, 51%–75%, and 76%–100% of all possible points.

  Additional information can be found starting on page 41 of the M-STEP IGTR, page 39 of the Grade 8 IGTR, and page 41 of the MME IGTR. The Expectation Analysis Report is also available in MiLearn for educators.

- The Demographic Report provides a comparison of students by grade and content area, aggregated across selected demographic groups, showing the percentages proficient at each level (i.e., advanced, proficient, partially proficient, and not proficient). The demographic report is available at the school, district, and state levels. Users can select different populations of students to be displayed. The following student populations may be selected:
  - All Students—this is the default
  - All Except Students with Disabilities—students who are not marked Special Education in the Michigan Student Data System (MSDS) at the time of testing
  - Students with Disabilities—students who are marked Special Education in MSDS at the time of testing

Additional information can be found starting on page 43 of the M-STEP IGTR, page 41 of the Grade 8 IGTR, and page 43 of the MME IGTR.

- The Comprehensive Report provides a comparison of students by grade and content area, aggregated across schools and within a district, showing the percentages proficient
  at each level (i.e., advanced, proficient, partially proficient, and not proficient). The Comprehensive Report is available at the and district level. After the user selects a grade to view, all tested content areas for that grade are displayed sequentially in alphabetical order.
  Additional information can be found starting on page 49 of the M-STEP IGTR, page 44 of the Grade 8 IGTR, and page 46 of the MME IGTR.

- The Aggregate Data File contains student performance data aggregated for all students and by select demographic subgroups across buildings, districts, and the state. This data includes school information, student population, demographic group, and student performance data.

  The Aggregate Data File is provided for schools and districts to use as a data resource for school- or district-level data reviews. Schools or districts can use the Aggregate Data Files to evaluate data in ways that support school improvement goals or other data-based decision-making purposes.

  Additional information can be found starting on page 54 of the M-STEP IGTR, page 49 of the Grade 8 IGTR, and page 50 of the MME IGTR. The Aggregated Data Files can be downloaded from the OEAA Secure Site. The file layout is in Appendix B.4.

9.4 Interpretive Guides

For the spring 2019 M-STEP, MDE produced individual and aggregate reports for students, schools, districts, and the state. The information provided in these reports can be interpreted and used in a variety of ways. In addition to providing interpretation, it is important that the information is understandable by the target audience. Standard 7.0 of the AERA, APA, and NCME (2014) *Standards* states the following:

> Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (p. 125)

To aid in interpretation, MDE prepared the *Spring 2019 Interpretive Guide to M-STEP Reports* for grades 3 through 7, the *Spring 2019 Michigan Grade 8 Testing Interpretive Guide to Reports*, and the *Spring 2019 Interpretive Guide to MME Reports* for grade 11.

The *Spring 2019 Interpretive Guide to M-STEP Reports* can be found in Appendix B.1 of this technical report. The *Spring 2019 Michigan Grade 8 Testing Interpretive Guide to Reports* can be found in Appendix B.2. The *Spring 2019 Interpretive Guide to MME Reports* can be found in Appendix B.3.

9.5 Summary

In summary, the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information developed by MDE are in alignment with multiple best practices of the testing industry, particularly the following *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014):

- Standard 5.1—Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.
- Standard 6.10—When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.
- Standard 7.0—Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.

Table 9-1a. M-STEP Test Completion Rates by Grade: English Language Arts—All Students

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
All Students	Total Tested	100,879	102,400	105,150	109,026	109,052
All Students	Number Valid	100,793	102,327	105,078	108,948	108,975
All Students	Percent Valid	99.91%	99.93%	99.93%	99.93%	99.93%

Table 9-1b. M-STEP Test Completion Rates by Grade: English Language Arts—Gender

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Female	Total Tested	49,474	50,246	51,548	53,383	53,748
Female	Number Valid	49,431	50,208	51,517	53,349	53,716
Female	Percent Valid	99.91%	99.92%	99.94%	99.94%	99.94%
Male	Total Tested	51,405	52,154	53,602	55,643	55,304
Male	Number Valid	51,362	52,119	53,561	55,599	55,259
Male	Percent Valid	99.92%	99.93%	99.92%	999.92%	99.92%

Table 9-1c. M-STEP Test Completion Rates by Grade: English Language Arts—Race/Ethnicity

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
American Indian or Alaska Native	Total Tested	550	593	626	719	689	736
American Indian or Alaska Native	Number Valid	549	593	625	718	688	736
American Indian or Alaska Native	Percent Valid	99.82%	100%	99.84%	99.86%	99.85%	100.00%
Asian	Total Tested	3,525	3,515	3,520	3,651	3,669	3,798
Asian	Number Valid	3,519	3,513	3,519	3,649	3,669	3,798
Asian	Percent Valid	99.83%	99.94%	99.97%	99.95%	100%	100.00%
Black or African American	Total Tested	18,994	18,999	18,736	18,982	18,882	
Black or African American	Number Valid	18,970	18,976	18,718	18,955	18,860	18,313
Black or African American	Percent Valid	99.87%	99.88%	99.90%	99.86%	99.88%	99.85%
Hispanic or Latino	Total Tested	8,426	8,355	8,751	9,069	8,835	8,344
Hispanic or Latino	Number Valid	8,420	8,344	8,747	9,065	8,823	8,338
Hispanic or Latino	Percent Valid	99.93%	99.87%	99.95%	99.96%	99.86%	99.93%
Native Hawaiian or Other Pacific Islander	Total Tested	97	87	78	83	88	3,722
Native Hawaiian or Other Pacific Islander	Number Valid	97	85	78	82	88	3,617
Native Hawaiian or Other Pacific Islander	Percent Valid	100%	97.70%	100%	98.80%	100%	100.00%
Two or More Races	Total Tested	4,900	4,717	4,807	4,617	4,578	3,918
Two or More Races	Number Valid	4,897	4,713	4,803	4,614	4,574	3,914
Two or More Races	Percent Valid	99.94%	99.92%	99.92%	99.94%	99.91%	99.90%
White	Total Tested	64,387	66,134	68,632	71,905	72,311	75,216
White	Number Valid	64,341	66,103	68,588	71,865	72,273	75,180
White	Percent Valid	99.93%	99.95%	99.94%	99.94%	99.95%	99.95%

Table 9-1d. M-STEP Test Completion Rates by Grade: English Language Arts—Economically Disadvantaged

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	56,620	56,637	56,948	57,821	56,048
Yes	Number Valid	56,600	56,621	56,931	57,790	56,019
Yes	Percent Valid	99.96%	99.97%	99.97%	99.95%	99.95%
No	Total Tested	44,259	45,763	48,202	51,205	53,004
No	Number Valid	44,193	45,706	48,147	51,158	52,956
No	Percent Valid	99.85%	99.88%	99.89%	99.91%	99.91%

Table 9-1e M-STEP Test Completion Rates by Grade: English Language Arts—English Language Learners

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	9,670	9,031	7,786	6,520	6,609
Yes	Number Valid	9,668	9,031	7,784	6,519	6,607
Yes	Percent Valid	99.98%	100%	99.97%	99.98%	99.97%
No	Total Tested	91,209	93,369	97,364	102,506	102,443
No	Number Valid	91,125	93,296	97,294	102,429	102,368
No	Percent Valid	99.91%	99.92%	99.93%	99.92%	99.93%

Table 9-1f. M-STEP Test Completion Rates by Grade: English Language Arts—Students with Disabilities

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	11,930	12,148	12,586	12,373	11,946
Yes	Number Valid	11,922	12,138	12,574	12,355	11,933
Yes	Percent Valid	99.93%	99.92%	99.90%	99.85%	99.89%
No	Total Tested	88,949	90,252	92,564	96,653	97,106
No	Number Valid	88,871	90,189	92,504	96,593	97,042
No	Percent Valid	99.91%	99.93%	99.94%	99.94%	99.93%

Table 9-1g. M-STEP Test Completion Rates by Grade: English Language Arts—Students Used Standard Accommodations

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	315	316	324	4,881	4,841
Yes	Number Valid	313	315	321	4,875	4,838
Yes	Percent Valid	99.37%	99.68%	99.07%	99.88%	99.94%
No	Total Tested	100,564	102,084	104,826	104,145	104,211
No	Number Valid	100,480	102,012	104,757	104,073	104,137
No	Percent Valid	99.92%	99.93%	99.93%	99.93%	99.93%

Table 9-2a. M-STEP Test Completion Rates by Grade: Mathematics—All Students

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
All Students	Total Tested	101,114	102,684	105,344	109,187	109,157
All Students	Number Valid	101,019	102,602	105,272	109,108	109,072
All Students	Percent Valid	99.91%	99.92%	99.93%	99.93%	99.92%

Table 9-2b. M-STEP Test Completion Rates by Grade: Mathematics—Gender

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Female	Total Tested	49,578	50,358	51,656	53,443	53,812
Female	Number Valid	49,531	50,313	51,622	53,410	53,778
Female	Percent Valid	99.91%	99.91%	99.93%	99.94%	99.94%
Male	Total Tested	51,536	52,326	53,688	55,744	55,345
Male	Number Valid	51,488	52,289	53,650	55,698	55,294
Male	Percent Valid	99.91%	99.93%	99.93%	99.92%	99.91%

Table 9-2c. M-STEP Test Completion Rates by Grade: Mathematics—Race/Ethnicity

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
American Indian or Alaska Native	Total Tested	551	594	628	720	688	736
American Indian or Alaska Native	Number Valid	550	594	627	719	688	736
American Indian or Alaska Native	Percent Valid	99.82%	100%	99.84%	99.86%	100%	100.00%
Asian	Total Tested	3,637	3,617	3,618	3,720	3,722	
Asian	Number Valid	3,623	3,610	3,614	3,715	3,722	3,832
Asian	Percent Valid	99.62%	99.81%	99.89%	99.87%	100%	100.00%
Black or African American	Total Tested	19,008	19,008	18,733	18,983	18,860	18,285
Black or African American	Number Valid	18,980	18,988	18,718	18,955	18,841	18,994
Black or African American	Percent Valid	99.85%	99.89%	99.92%	99.85%	99.90%	99.84%
Hispanic or Latino	Total Tested	8,472	8,422	8,806	9,118	8,889	8,372
Hispanic or Latino	Number Valid	8,465	8,409	8,800	9,111	8,873	8,368
Hispanic or Latino	Percent Valid	99.92%	99.85%	99.93%	99.92%	99.82%	99.95%
Native Hawaiian or Other Pacific Islander	Total Tested	97	87	78	83	88	93
Native Hawaiian or Other Pacific Islander	Number Valid	97	85	78	82	88	93
Native Hawaiian or Other Pacific Islander	Percent Valid	100%	97.70%	100%	98.80%	100%	100.00%
Two or More Races	Total Tested	4,905	4,720	4,799	4,613	4,575	3,910
Two or More Races	Number Valid	4,899	4,716	4,795	4,610	4,572	3,908
Two or More Races	Percent Valid	99.88%	99.92%	99.92%	99.93%	99.93%	99.95%
White	Total Tested	64,444	66,236	68,682	71,950	72,335	3,832
White	Number Valid	64,405	66,200	68,640	71,916	72,288	75,191
White	Percent Valid	99.94%	99.95%	99.94%	99.95%	99.94%	99.97%

Table 9-2d. M-STEP Test Completion Rates by Grade: Mathematics—Economically Disadvantaged

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	56,705	56,748	57,014	57,887	56,068
Yes	Number Valid	56,682	56,731	57,002	57,864	56,038
Yes	Percent Valid	99.96%	99.97%	99.98%	99.96%	99.95%
No	Total Tested	44,409	45,936	48,330	51,300	53,089
No	Number Valid	44,337	45,871	48,270	51,244	53,034
No	Percent Valid	99.84%	99.86%	99.88%	99.89%	99.90%

Table 9-2e. M-STEP Test Completion Rates by Grade: Mathematics—English Language Learners

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Yes	Total Tested	9,873	9,287	7,992	6,682	6,764	6,268
Yes	Number Valid	9,870	9,285	7,989	6,680	6,763	6,261
Yes	Percent Valid	99.97%	99.98%	99.96%	99.97%	99.99%	99.93%
No	Total Tested	91,241	93,397	97,352	102,505	102,393	104,149
No	Number Valid	91,149	93,317	97,283	102,428	102,309	104,082
No	Percent Valid	99.90%	99.91%	99.93%	99.92%	99.92%	99.94%

Table 9- 2f. M-STEP Test Completion Rates by Grade: Mathematics—Students with Disabilities

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	11,972	12,193	12,589	12,361	11,910
Yes	Number Valid	11,968	12,185	12,582	12,351	11,899
Yes	Percent Valid	99.97%	99.93%	99.94%	99.92%	99.91%
No	Total Tested	89,142	90,491	92,755	96,826	97,247
No	Number Valid	89,051	90,417	92,690	96,757	97,173
No	Percent Valid	99.90%	99.92%	99.93%	99.93%	99.92%

Table 9-2g. M-STEP Test Completion Rates by Grade: Mathematics—Students Used Standard Accommodations

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	Total Tested	143	1,867	2,652	2,815	2,522
Yes	Number Valid	143	1,864	2,650	2,814	2,521
Yes	Percent Valid	100%	99.84%	99.92%	99.96%	99.96%
No	Total Tested	100,971	100,817	102,692	106,372	106,635
No	Number Valid	100,876	100,738	102,622	106,294	106,551
No	Percent Valid	99.91%	99.92%	99.93%	99.93%	99.92%

Table 9-3a. M-STEP Test Completion Rates by Grade: Social Studies—All Students

		Grade 5	Grade 8	Grade 11
All Students	Total Tested	105,324	108,037	102,522
All Students	Number Valid	105,116	107,811	102,297
All Students	Percent Valid	99.8%	99.8%	99.8%

Table 9-3b. M-STEP Test Completion Rates by Grade: Social Studies—Gender

		Grade 5	Grade 8	Grade 11
Female	Total Tested	51,648	53,028	51,276
Female	Number Valid	51,562	52,938	51,175
Female	Percent Valid	99.8%	99.8%	99.8%
Male	Total Tested	53,676	55,009	51,246
Male	Number Valid	53,554	54,873	51,122
Male	Percent Valid	99.8%	99.8%	99.8%

Table 9-3c. M-STEP Test Completion Rates by Grade: Social Studies—Race/Ethnicity

		Grade 5	Grade 8	Grade 11
American Indian or Alaska Native	Total Tested	627	738	677
American Indian or Alaska Native	Number Valid	626	735	671
American Indian or Alaska Native	Percent Valid	99.8%	99.6%	99.1%
Asian	Total Tested	3,617	3,757	3,962
Asian	Number Valid	3,611	3,756	3,960
Asian	Percent Valid	99.83%	99.97%	99.95%
Black or African American	Total Tested	18,721	18,193	15,388
Black or African American	Number Valid	18,630	18,097	15,308
Black or African American	Percent Valid	99.51%	99.47%	99.48%
Hispanic or Latino	Total Tested	8,803	9,003	7,144
Hispanic or Latino	Number Valid	8,788	8,991	7,118
Hispanic or Latino	Percent Valid	99.83%	99.87%	99.64%
Native Hawaiian or Other Pacific Islander	Total Tested	79	68	90
Native Hawaiian or Other Pacific Islander	Number Valid	79	68	90
Native Hawaiian or Other Pacific Islander	Percent Valid	100%	100%	100%
Two or More Races	Total Tested	4,800	4,027	3,078
Two or More Races	Number Valid	4,792	4,020	3,071
Two or More Races	Percent Valid	99.83%	99.83%	99.77%
White	Total Tested	68,677	72,251	72,183
White	Number Valid	68,590	72,144	72,079
White	Percent Valid	99.87%	99.85%	99.86%

Table 9-3d. M-STEP Test Completion Rates by Grade: Social Studies—Economically Disadvantaged

		Grade 5	Grade 8	Grade 11
Yes	Total Tested	57,021	53,584	43,033
Yes	Number Valid	56,893	53,419	42,883
Yes	Percent Valid	99.78%	99.69%	99.65%
No	Total Tested	48,303	54,453	59,489
No	Number Valid	48,223	54,392	59,414
No	Percent Valid	99.83%	99.89%	99.87%

Table 9-3e. M-STEP Test Completion Rates by Grade: Social Studies—English Language Learners

		Grade 5	Grade 8	Grade 11
Yes	Total Tested	7,990	6,339	5,025
Yes	Number Valid	7,973	6,323	5,007
Yes	Percent Valid	99.79%	99.75%	99.64%
No	Total Tested	97,334	101,698	97,497
No	Number Valid	97,143	101,488	97,290
No	Percent Valid	99.80%	99.79%	99.79%

Table 9-3f. M-STEP Test Completion Rates by Grade: Social Studies—Students with Disabilities

		Grade 5	Grade 8	Grade 11
Yes	Total Tested	12,621	11,846	9,431
Yes	Number Valid	12,589	11,779	9,396
Yes	Percent Valid	99.75%	99.43%	99.63%
No	Total Tested	92,703	96,191	93,091
No	Number Valid	92,527	96,032	92,901
No	Percent Valid	99.81%	99.83%	99.80%

Table 9-3g. M-STEP Test Completion Rates by Grade: Social Studies—Students Used Standard Accommodations

		Grade 5	Grade 8	Grade 11
Yes	Total Tested	208	652	237
Yes	Number Valid	206	651	235
Yes	Percent Valid	99.04%	99.85%	99.16%
No	Total Tested	105,116	107,385	102,285
No	Number Valid	104,910	107,160	102,062
No	Percent Valid	99.80%	99.79%	99.78%

Table 9-4a. Scale-Score Descriptive Statistics by Grade: English Language Arts—All Students

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
All Students	N	100,793	102,327	105,078	108,948	108,975
All Students	Mean SS	1,295.3	1,395.8	1,496.0	1,592.9	1,693.9
All Students	SD SS	26.2	26.3	27.2	26.3	26.4

Table 9-4b. Scale-Score Descriptive Statistics by Grade: English Language Arts—Gender

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Female	N	49,431	50,208	51,517	53,349	53,716
Female	Mean SS	1,297.3	1,397.9	1,498.8	1,596.0	1,697.5
Female	SD SS	26.2	25.8	26.9	26.0	25.5
Male	N	51,362	52,119	53,561	55,599	55,259
Male	Mean SS	1,293.3	1,393.8	1,493.3	1,590.0	1,690.4
Male	SD SS	26.1	26.5	27.2	26.3	26.9

Table 9-4c. Scale-Score Descriptive Statistics by Grade: English Language Arts—Race/Ethnicity

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
American Indian or Alaska Native	N	549	593	625	718	688
American Indian or Alaska Native	Mean SS	1,289.6	1,389.4	1,489.8	1,587.1	1,689.4
American Indian or Alaska Native	SD SS	23.0	23.9	25.3	24.6	25.1
Asian	N	3,519	3,513	3,519	3,649	3,669
Asian	Mean SS	1,308.2	1,410.0	1,512.6	1,609.5	1,710.7
Asian	SD SS	26.0	26.0	27.0	26.4	26.4
Black or African American	N	18,970	18,976	18,718	18,955	18,860
Black or African American	Mean SS	1,278.6	1,379.6	1,479.4	1,577.4	1,678.6
Black or African American	SD SS	23.3	23.8	24.2	22.9	23.7
Hispanic or Latino	N	8,420	8,344	8,747	9,065	8,823
Hispanic or Latino	Mean SS	1,289.5	1,389.8	1,489.6	1,586.5	1,688.0
Hispanic or Latino	SD SS	24.1	24.1	24.7	24.4	24.3
Native Hawaiian or Other Pacific Islander	N	97	85	78	82	88
Native Hawaiian or Other Pacific Islander	Mean SS	1,300.1	1,396.2	1,496.0	1,597.9	1,695.4
Native Hawaiian or Other Pacific Islander	SD SS	24.9	24.4	27.2	26.8	27.7
Two or More Races	N	4,897	4,713	4,803	4,614	4,574
Two or More Races	Mean SS	1,294.8	1,394.5	1,494.8	1,591.1	1,692.4
Two or More Races	SD SS	25.9	26.0	26.8	26.2	26.0
White	N	64,341	66,103	68,588	71,865	72,273
White	Mean SS	1,300.3	1,400.7	1,500.6	1,597.1	1,697.9
White	SD SS	25.0	25.1	26.2	25.4	25.6

Table 9-4d. Scale-Score Descriptive Statistics by Grade: English Language Arts—Economically Disadvantaged

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	56,600	56,621	56,931	57,790	56,019
Yes	Mean SS	1,286.8	1,387.1	1,486.7	1,583.9	1,684.6
Yes	SD SS	24.5	24.5	25.1	24.4	24.7
No	N	44,193	45,706	48,147	51,158	52,956
No	Mean SS	1,306.1	1,406.7	1,507.0	1,603.0	1,703.7
No	SD SS	24.2	24.2	25.3	24.6	24.6

Table 9-4e. Scale-Score Descriptive Statistics by Grade: English Language Arts—English Language Learners

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	9,668	9,031	7,784	6,519	6,607
Yes	Mean SS	1,288.8	1,386.6	1,481.9	1,575.0	1,675.2
Yes	SD SS	23.9	23.2	22.1	20.4	21.0
No	N	91,125	93,296	97,294	102,429	102,368
No	Mean SS	1,296.0	1,396.7	1,497.1	1,594.0	1,695.1
No	SD SS	26.3	26.4	27.2	26.2	26.3

Table 9-4f. Scale-Score Descriptive Statistics by Grade: English Language Arts—Students with Disabilities

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	11,922	12,138	12,574	12,355	11,933
Yes	Mean SS	1,279.5	1,377.8	1,475.0	1,571.4	1,671.2
Yes	SD SS	22.9	23.2	22.9	21.2	21.9
No	N	88,871	90,189	92,504	96,593	97,042
No	Mean SS	1,297.4	1,398.3	1,498.8	1,595.6	1,696.7
No	SD SS	25.9	25.7	26.5	25.6	25.6

Table 9-4g. Scale-Score Descriptive Statistics by Grade: English Language Arts—Students Used Standard Accommodations

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	313	315	321	4,875	4,838
Yes	Mean SS	1,273.8	1,377.1	1,475.5	1,569.2	1,669.0
Yes	SD SS	20.0	24.5	24.8	18.6	19.1
No	N	100,480	102,012	104,757	104,073	104,137
No	Mean SS	1,295.3	1,395.9	1,496.1	1,594.0	1,695.1
No	SD SS	26.2	26.3	27.2	26.1	26.1

Table 9-5a. Scale-Score Descriptive Statistics by Grade: Mathematics—All Students

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
All Students	N	101,019	102,602	105,272	109,108	109,072
All Students	Mean SS	1,296.6	1,393.6	1,487.6	1,587.9	1,688.3
All Students	SD SS	27.4	25.7	26.4	26.0	26.6

Table 9-5b. Scale-Score Descriptive Statistics by Grade: Mathematics—Gender

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Female	N	49,531	50,313	51,622	53,410	53,778
Female	Mean SS	1,295.2	1,392.3	1,486.4	1,587.6	1,688.3
Female	SD SS	26.6	24.6	25.2	25.1	25.4
Male	N	51,488	52,289	53,650	55,698	55,294
Female	Mean SS	1,298.0	1,394.9	1,488.7	1,588.2	1,688.3
Female	SD SS	28.0	26.7	27.5	26.8	27.8

Table 9-5c. Scale-Score Descriptive Statistics by Grade: Mathematics—Race/Ethnicity

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
American Indian or Alaska Native	N	550	594	627	719	688
American Indian or Alaska Native	Mean SS	1,291.2	1,387.9	1,481.8	1,581.8	1,683.9
American Indian or Alaska Native	SD SS	23.1	22.3	24.3	23.7	24.5
Asian	N	3,623	3,610	3,614	3,715	3,722
Asian	Mean SS	1,315.9	1,412.3	1,509.2	1,609.6	1,710.0
Asian	SD SS	26.5	25.5	26.2	26.5	28.0
Black or African American	N	18,980	18,988	18,718	18,955	18,841
Black or African American	Mean SS	1,278.6	1,376.1	1,469.0	1,569.1	1,669.6
Black or African American	SD SS	25.2	23.3	23.1	23.2	23.2
Hispanic or Latino	N	8,465	8,409	8,800	9,111	8,873
Hispanic or Latino	Mean SS	1,290.4	1,387.3	1,480.8	1,580.8	1,681.1
Hispanic or Latino	SD SS	24.9	23.4	23.8	23.8	24.4
Native Hawaiian or Other Pacific Islander	N	97	85	78	82	88
	Mean SS	1,298.6	1,389.6	1,487.3	1,591.4	1,685.8
	SD SS	27.6	25.0	23.7	25.5	26.6
Two or More Races	N	4,899	4,716	4,795	4,610	4,572
Two or More Races	Mean SS	1,295.2	1,391.4	1,485.3	1,585.1	1,685.3
Two or More Races	SD SS	26.8	25.8	26.0	26.1	26.3
White	N	64,405	66,200	68,640	71,916	72,288
White	Mean SS	1,301.8	1,398.7	1,492.6	1,592.8	1,693.1
White	SD SS	25.6	23.9	24.8	24.0	24.9

Table 9-5d. Scale-Score Descriptive Statistics by Grade: Mathematics—Economically Disadvantaged

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	56,682	56,731	57,002	57,864	56,038
Yes	Mean SS	1,287.6	1,384.6	1,477.9	1,578.2	1,678.0
Yes	SD SS	25.9	24.1	24.4	24.4	24.6
No	N	44,337	45,871	48,270	51,244	53,034
No	Mean SS	1,308.2	1,404.7	1,499.0	1,598.8	1,699.1
No	SD SS	24.7	23.2	24.1	23.2	24.3

Table 9-5e. Scale-Score Descriptive Statistics by Grade: Mathematics—English Language Learners

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	9,870	9,285	7,989	6,680	6,763
Yes	Mean SS	1,293.8	1,387.4	1,477.8	1,573.4	1,672.8
Yes	SD SS	25.8	24.0	23.6	23.5	23.9
No	N	91,149	93,317	97,283	102,428	102,309
No	Mean SS	1,296.9	1,394.3	1,488.4	1,588.8	1,689.3
No	SD SS	27.5	25.8	26.5	25.9	26.5

Table 9-5f. Scale-Score Descriptive Statistics by Grade: Mathematics—Students with Disabilities

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	11,968	12,185	12,582	12,351	11,899
Yes	Mean SS	1,276.8	1,373.9	1,465.7	1,562.7	1,662.9
Yes	SD SS	28.1	25.5	24.5	24.6	23.2
No	N	89,051	90,417	92,690	96,757	97,173
No	Mean SS	1,299.3	1,396.3	1,490.5	1,591.1	1,691.4
No	SD SS	26.1	24.6	25.3	24.4	25.3

Table 9-5g. Scale-Score Descriptive Statistics by Grade: Mathematics—Students Used Standard Accommodations

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Yes	N	143	1,864	2,650	2,814	2,521
Yes	Mean SS	1,278.3	1,363.9	1,457.5	1,554.7	1,656.1
Yes	SD SS	26.9	20.0	19.2	20.1	18.8
No	N	100,876	100,738	102,622	106,294	106,551
No	Mean SS	1,296.7	1,394.2	1,488.3	1,588.8	1,689.0
No	SD SS	27.4	25.5	26.1	25.6	26.3

Table 9-6a. Scale-Score Descriptive Statistics by Grade: Social Studies—All Students

		Grade 5	Grade 8	Grade 11
All Students	N	105,116	107,811	102,297
All Students	Mean SS	1,476.5	1,785.5	2,097.8
All Students	SD SS	24.7	25.2	25.1

Table 9-6b. Scale-Score Descriptive Statistics by Grade: Social Studies—Gender

		Grade 5	Grade 8	Grade 11
Female	N	51,562	52,938	51,175
Female	Mean SS	1,475.3	1,784.4	2,096.3
Female	SD SS	23.7	23.8	23.0
Male	N	53,554	54,873	51,122
Male	Mean SS	1,477.8	1,786.6	2,099.4
Male	SD SS	25.6	26.5	27.0

Table 9-6c. Scale-Score Descriptive Statistics by Grade: Social Studies—Race/Ethnicity

		Grade 5	Grade 8	Grade 11
American Indian or Alaska Native	N	626	735	671
American Indian or Alaska Native	Mean SS	1,472.8	1,781.9	2,094.0
American Indian or Alaska Native	SD SS	23.8	23.5	23.3
Asian	N	3,611	3,756	3,960
Asian	Mean SS	1,489.0	1,800.6	2,109.0
Asian	SD SS	26.9	26.1	26.7
Black or African American	N	18,630	18,097	15,308
Black or African American	Mean SS	1,461.4	1,769.0	2,082.4
Black or African American	SD SS	19.4	19.2	20.0
Hispanic or Latino	N	8,788	8,991	7,118
Hispanic or Latino	Mean SS	1,469.2	1,778.7	2,090.3
Hispanic or Latino	SD SS	21.5	22.5	22.6
Native Hawaiian or Other Pacific Islander	N	79	68	90
Native Hawaiian or Other Pacific Islander	Mean SS	1,474.4	1,784.8	2,096.3
Native Hawaiian or Other Pacific Islander	SD SS	21.8	26.7	24.2
Two or More Races	N	4,792	4,020	3,071
Two or More Races	Mean SS	1,474.6	1,783.9	2,097.6
Two or More Races	SD SS	23.7	25.2	25.4
White	N	68,590	72,144	72,079
White	Mean SS	1,481.1	1,789.8	2,101.3
White	SD SS	24.3	24.8	24.7

Table 9-6d. Scale-Score Descriptive Statistics by Grade: Social Studies—Economically Disadvantaged

		Grade 5	Grade 8	Grade 11
Yes	N	56,893	53,419	42,883
Yes	Mean SS	1,467.9	1,776.0	2,088.7
Yes	SD SS	21.4	22.0	22.4
No	N	48,223	54,392	59,414
No	Mean SS	1,486.7	1,794.8	2,104.5
No	SD SS	24.4	24.7	24.8

Table 9-6e. Scale-Score Descriptive Statistics by Grade: Social Studies—English Language Learners

		Grade 5	Grade 8	Grade 11
Yes	N	7,973	6,323	5,007
Yes	Mean SS	1,463.2	1,767.9	2,076.3
Yes	SD SS	18.9	17.2	16.9
No	N	97,143	101,488	97,290
No	Mean SS	1,477.6	1,786.6	2,098.9
No	SD SS	24.8	25.2	24.9

Table 9-6f. Scale-Score Descriptive Statistics by Grade: Social Studies—Students with Disabilities

		Grade 5	Grade 8	Grade 11
Yes	N	12,589	11,779	9,396
Yes	Mean SS	1,461.1	1,767.1	2,079.4
Yes	SD SS	20.3	19.5	20.4
No	N	92,527	96,032	92,901
No	Mean SS	1,478.7	1,787.8	2,099.7
No	SD SS	24.5	24.9	24.8

Table 9-6g. Scale-Score Descriptive Statistics by Grade: Social Studies—Students Used Standard Accommodations

		Grade 5	Grade 8	Grade 11
Yes	N	206	651	235
Yes	Mean SS	1,454.9	1,762.7	2,077.0
Yes	SD SS	15.9	16.6	18.5
No	N	104,910	107,160	102,062
No	Mean SS	1,476.6	1,785.6	2,097.9
No	SD SS	24.7	25.2	25.1

Table 9-7a. Scale-Score Descriptive Statistics: English Language Arts

Grade	*N*	Mean	SD	Min	Max
3	100,793	1295.3	26.2	1,203	1,357
4	102,327	1395.8	26.3	1,301	1,454
5	105,078	1496	27.2	1,409	1,560
6	108,948	1592.9	26.3	1,508	1,655
7	108,975	1693.9	26.4	1,618	1,753

Table 9-7b. Performance-Level Percentages: English Language Arts

Grade	*N*	Not Proficient	Partially Proficient	Proficient	Advanced
3	100,793	30.40%	24.50%	22.42%	22.68%
4	102,327	33.41%	20.76%	21.55%	24.28%
5	105,078	32.31%	21.50%	28.49%	17.69%
6	108,948	31.67%	26.59%	28.21%	13.52%
7	108,975	29.69%	27.56%	30.24%	12.51%

Table 9-8a. Scale-Score Descriptive Statistics: Mathematics

Grade	*N*	Mean	SD	Min	Max
3	101,019	1296.6	27.4	1,217	1,361
4	102,602	1393.6	25.7	1,310	1,455
5	105,272	1487.6	26.4	1,409	1,550
6	109,108	1587.9	26.0	1,518	1,650
7	109,072	1688.3	26.6	1,621	1,752

Table 9-8b. Performance-Level Percentages: Mathematics

Grade	*N*	Not Proficient	Partially Proficient	Proficient	Advanced
3	101,019	27.51%	25.76%	27.24%	19.50%
4	102,602	24.65%	33.54%	25.22%	16.59%
5	105,272	36.46%	28.73%	17.96%	16.85%
6	109,108	34.25%	30.62%	18.95%	16.17%
7	109,072	35.91%	28.34%	19.32%	16.43%

Table 9-9a. Scale-Score Descriptive Statistics: Social Studies

Grade	*N*	Mean	SD	Min	Max
5	105,116	1476.5	24.7	1,400	1,561
8	107,811	1785.5	25.2	1,713	1,866
11	102,297	2097.8	25.1	2,023	2,166

Table 9-9b. Performance-Level Percentages: Social Studies

Grade	*N*	Not Proficient	Partially Proficient	Proficient	Advanced
5	105,116	24.09%	58.53%	14.75%	2.63%
8	107,811	32.48%	39.49%	23.15%	4.88%
11	102,297	12.84%	40.60%	35.28%	11.27%

Chapter 10: Performance-Level Setting

This chapter briefly describes the M-STEP performance level setting and presents the cut scores established and the performance level descriptors derived from the performance level setting.

M-STEP is administered to assess Michigan students' mastery of the Michigan state standards. The assessments began as an implementation of the Smarter Balanced ELA and mathematics tests. The current cut scores for the tests are taken from the SBAC tests. A brief overview of the Smarter Balanced standard-setting procedures during which the cut scores for ELA and mathematics were derived can be found in the report about performance level setting, *Smarter Balanced Assessment Consortium: Achievement Level Setting Final Report* (2015d), which is posted on the Smarter Balanced library web page.[1]

Over the course of several years, important changes have been made to the assessments to make them more meaningful to Michigan educators. These include the alignment of the test items to the Michigan Academic *Standards*, the implementation of a Michigan-specific test blueprint, and a reduction and then elimination of performance tasks to reduce overall test time. These changes were made cautiously and deliberately with the active involvement of Michigan educators and stakeholders.

In school year 2017–18, the tests in grades 3–8 were shortened to a legislatively mandated median time of three hours to reduce the time burden on students and schools. To do so, all performance tasks in ELA were replaced with Passage-based Writing (PBW) items, a form of constructed response (CR) item. The ELA test blueprints were adjusted to accommodate the new item type and the reduction in test length. In grades 3–8 mathematics, the test was also shortened to reduce overall testing time, but this change did not involve adding new item types or significantly altering the test blueprint. Content-rich technology enhanced items were substituted to fulfill the blueprint.

As a result of these changes, MDE partnered with DRC and Michigan educators to evaluate the validity of the cut scores derived by Smarter Balanced for ELA and mathematics with a cut score validation meeting in July 2018. There have been no significant changes to the test blueprints since then.

For social studies, a statistical articulation was used to establish cut scores.

The AERA, APA, & NCME (2014) *Standards* addressed in this chapter are 5.21 and 5.22, which will be presented in the pertinent sections of the chapter.

The AERA, APA, and NCME Standard 5.21 states that:

> When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (p. 107)

[1] https://portal.smarterbalanced.org/library/en/achievement-level-setting-final-report-with-appendix.pdf

To evaluate the validity of M-STEP score interpretations, it is essential to understand that descriptors and cut scores are established in a collaborative and participatory process. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores.

10.1 Cut Score Validation for English Language Arts and Mathematics

Cut scores were validated after the spring 2018 M-STEP administration. The purpose of the standards validation was to determine whether the existing M-STEP cut scores for grades 3–8 (now 3-7) ELA and mathematics were still valid for continued use, given the 2018 updates to the tests.

A total of 54 Michigan educators engaged in a modification of the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) to validate the cut scores. This method has been used on large-scale assessments in Michigan and across the nation, including by Smarter Balanced.

Participants studied the existing Michigan performance level descriptors (PLDs) and Michigan state standards to review the knowledge, skills, and abilities expected of students in each performance level. The four performance levels on M-STEP are Not Proficient, Partially Proficient, Proficient, and Advanced. Each performance level is associated with a level of mastery of the Michigan Learning *Standards*. Participants then discussed the content-based expectations for students at the threshold of each performance level (e.g., a student who is "just" Proficient). To support their discussions of these threshold students, participants were provided with the Smarter Balanced achievement level descriptors (ALDs). These Smarter Balanced ALDs were used at the original standard setting where the cut scores were established.

Participants studied collections of test items that were ordered in terms of difficulty. The existing cut scores were presented as benchmarks for participants' consideration: participants were asked to consider the knowledge and skills that students would need to demonstrate on the updated ELA and mathematics tests, as based on the benchmarked (existing) cut scores. Then, participants compared these expectations against the content-based expectations for students at the thresholds of each performance level. Participants were instructed to recommend retaining the existing cut scores if there was good correspondence between the benchmarks and these content-based expectations or to recommend alternative cut scores that reflect better correspondence. Participants engaged in two rounds of individual judgments and group discussion. (The grade 5 mathematics committee engaged in three rounds of judgments to accommodate additional discussion.) The committees' median judgments were taken as their final recommendations.

The available validity evidence suggests that there were no significant differences between the updated ELA and mathematics assessments and the content assessed by the prior assessments and that the differences between the judgments made at the 2018 standards validation workshop and the existing cut scores were not statistically different. That is, the recommendations made by Michigan educators during the standards validation were consistent with the existing cut scores, and the validity evidence collected during this process supports the continued use of the cut scores. More information can be found in the full *M-STEP Standards Validation 2018 Technical Report* (2019) in Appendix E.

10.2 Statistical Articulation for Social Studies

MDE partnered with DRC to conduct a standard-setting workshop for M-STEP social studies in grades 5, 8, and 11 in June 2015. During the workshop, participants considered the test items, performance level descriptors, and test data. Following the workshop, MDE considered participants' recommendations and discussed with the state superintendent. MDE found that the participants' recommended proficiency cuts were much lower than in the past and thus determined that such recommendations were not consistent with the high expectations of career and college readiness. As a result, in consultation with members of Michigan Technical Advisory Committee, MDE used statistically articulated cut scores and considered such approach to be more appropriate.

10.3 Scale Scores

This section describes the slopes and intercepts for transforming thetas to scale scores, as well as the LOSS and the HOSS for various M-STEP content areas. The values for ELA and mathematics were derived by MDE and DRC using the work done by Smarter Balanced.

For a detailed description of the methods used in calibration and scaling *Smarter Balanced 2017–2018 Technical Report* (Smarter Balanced, 2017). After calibration, results were in the theta metric. MDE transformed the theta metric results onto a four-digit scale, which is more meaningful for stakeholders. The equation for this linear transformation is

$$\textit{Scale score} = (\textit{theta} * \textit{slope}) + \textit{intercept}$$

Table 10-1 presents the information of slopes and intercepts for all four content areas, along with the LOSS and HOSS values which give the effective range of M-STEP scales for each grade and content area.

Table 10-1. Scale Transformation Slopes and Intercepts for M-STEP Summative Assessments with LOSS and HOSS Values

Content Area	Grade	Slope A	Intercept B	LOSS	HOSS
ELA	3	26.0061	1322.5934	1203	1357
ELA	4	24.6036	1409.5875	1301	1454
ELA	5	25.8718	1501.3628	1409	1560
ELA	6	24.5491	1592.9699	1508	1655
ELA	7	23.8151	1687.3543	1618	1753
Math	3	26.3725	1325.7407	1217	1361
Math	4	25.2608	1409.0233	1310	1455
Math	5	23.3374	1495.6493	1409	1550
Math	6	20.4573	1589.9260	1518	1650
Math	7	19.6292	1686.6036	1621	1752
Social Studies	5	27.2005	1478.3212	1395	1568
Social Studies	8	26.9339	1785.9405	1703	1868
Social Studies	11	26.8528	2095.9989	2016	2166

10.4 Cut Scores

This section presents the cut scores for each grade/content area of M-STEP. Table 10-2 shows the cut scores for ELA and mathematics in grades 3–7 and for social studies in grades 5, 8, and 11. It should be noted that for ELA and mathematics, the Smarter Balanced established cut scores on the theta matric were transformed to the (Michigan specific) M-STEP scales using a linear transformation.

Table 10-2. Cut Scores for M-STEP Summative Assessments

Content Area	Grade	SS Cut between Levels 1 and 2	SS Cut between Levels 2 and 3	SS Cut between Levels 3 and 4
ELA	3	1280	1299.5	1317
ELA	4	1383	1399.5	1417
ELA	5	1481	1499.5	1524
ELA	6	1578	1599.5	1624
ELA	7	1679	1699.5	1726
Math	3	1281	1299.5	1321
Math	4	1376	1399.5	1420
Math	5	1478	1499.5	1515
Math	6	1579	1599.5	1614
Math	7	1679	1699.5	1716
Social Studies	5	1458	1500	1530
Social Studies	8	1771	1800	1831
Social Studies	11	2069	2100	2131

10.5 Claim Cut Scores

As stated in Section 2.3, student performance on ELA and mathematics claims was classified into one of the three performance levels: *Adequate progress, Attention may be needed*, and *Most at risk of falling behind*. Detailed rules for calculating performance levels for ELA and mathematics claims can be found in Appendix D of the *Smarter Balanced Scoring Specifications, 2014–2015 Administration Summative and Interim Assessments: ELA/Literacy Grades 3–8, 11* and *Mathematics Grades 3–8, 11, V.7* (AIR, 2016).

10.6 Performance Level Descriptors

The performance level descriptors that were adopted by MDE for reporting purposes can be found in Tables 10-3 and 10-4.

Table 10-3. Performance-Level Descriptors for M-STEP, Grades 3–7

Performance Level	Descriptor
Advanced—PL 4	The student’s performance exceeds grade-level content standards and indicates substantial understanding and application of key concepts defined for Michigan students. The student needs support to continue to excel.
Proficient—PL 3	The student’s performance indicates understanding and application of key grade-level content standards defined for Michigan students. The student needs continued support to maintain and improve proficiency.
Partially Proficient—PL 2	The student needs assistance to improve achievement. The student’s performance is not yet proficient, indicating a partial understanding and application of the grade-level content standards defined for Michigan students.
Not Proficient—PL 1	The student needs intensive intervention and support to improve achievement. The student’s performance is not yet proficient and indicates minimal understanding and application of the grade level content standards defined for Michigan students.

Table 10-4. Performance-Level Descriptors for M-STEP, Grade 11

Performance Level	Descriptor
Advanced—PL 4	The student’s performance exceeds the high school content standards and indicates substantial understanding and application of key concepts defined for Michigan students. The student needs support to continue to excel and to be college- and career-ready.
Proficient—PL 3	The student’s performance indicates understanding and application of key high school content standards defined for Michigan students. The student needs continued support to maintain and improve proficiency and to be college- and career-ready.
Partially Proficient—PL 2	The student needs assistance to improve achievement and to become career and college ready. The student’s performance is not yet proficient, indicating a partial understanding and application of the high school content standards defined for Michigan students.
Not Proficient—PL 1	The student needs intensive intervention and support to improve achievement and to become career and college ready. The student’s performance is not yet proficient and indicates minimal understanding and application of the high school content standards defined for Michigan students.

10.7 Summary

This chapter presented a brief overview of the process for performance level setting used by Smarter Balanced for derivation of the ELA and mathematics cut scores. It also presents an overview of the procedure used for social studies. These procedures are addressed in more detail in Sections 10.1 and 10.2.

The standard settings undertaken by Smarter Balanced support the following *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014):

- Standard 5.21—When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.[2]
- Standard 5.22—When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.

Chapter 11: Fairness

As noted in the *Standards* (AERA, APA, & NCME, 2014), there are varying definitions of fairness. This chapter examines fairness as it relates to minimizing bias on a test and looks at test performance among varying subgroups assessed by M-STEP. It should be noted that differences in test performance among subgroups do not mean that a test is unfair—they simply mean that groups perform differently on the test. Even when a test is carefully and properly constructed, differences may exist among subgroups as a result of differences in curriculum or learning by the students in the subgroup.

This chapter is particularly relevant to AERA, APA, & NCME (2014) *Standards* 3.1 through 3.6. These standards are from Chapter 3 of the AERA, APA, & NCME (2014) *Standards*, “Fairness in Testing.” Each of these standards will be presented below. Standard 3.6 states the following:

> **Standard 3.6** Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws. (p. 65)

There is no specific research on M-STEP showing that the test scores of examinee subgroups differ in meaning; however, this is an ongoing concern in any large-scale testing program. To lessen the possibility of differences in test score meaning, DRC, MDE, and Smarter Balanced follow several steps in the item development and selection processes as explained in Section 11.1 of this chapter. In addition, DRC, MDE, and Smarter Balanced have conducted content and bias reviews on items, as explained in Chapter 3. These practices adhere to Standard 3.3:

> **Standard 3.3** Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test. (p. 64)

MDE and DRC have conducted differential item functioning (DIF) studies following the operational administration of M-STEP. Typically, items are evaluated for possible DIF in the field-test phase of test development, and items flagged for DIF are further examined for possible bias. During test development, Smarter Balanced follows procedures to minimize the inclusion of items that may potentially favor one demographic group over another. MDE and DRC staff do the same for social studies. Section 11.2 of this chapter explains the steps taken to evaluate M-STEP items through the use of DIF to adhere with this standard.

In addition, standardized test administration and training of test administrators for M-STEP comply with *Standards* 3.4 and 3.5:

> **Standard 3.4** Test takers should receive comparable treatment during the test administration and scoring process. (p. 65)
>
> **Standard 3.5** Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population. (p. 65)

Section 11.1 of this chapter is also directly relevant to *Standards* 3.1 and 3.2:

> **Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)
>
> **Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (p. 64)

Section 11.1 explains the steps taken by MDE and DRC to minimize words, phrases, and content that may be regarded as offensive by members of particular demographic subgroups. Chapter 3 discusses item content considerations during item development and item bias reviews for items included in M-STEP. These reviews are also critical in fulfilling *Standards* 3.1 and 3.2.

11.1 Minimizing Bias through Careful Test Development

The development of a test that is fair for all examinees begins in the early stages of planning and development. The item and test development processes that are used to minimize bias are summarized below.

First, careful attention is paid to content validity during the item development and item selection processes. Bias can occur only if the test is measuring different things for different groups. By eliminating irrelevant skills or knowledge from the items, the possibility of bias is reduced. Second, item writers and test developers follow several published guidelines for reducing or eliminating bias.

11.1.1 ELA and Mathematics

Smarter Balanced developed *Bias and Sensitivity Guidelines* (ETS, 2012) to help ensure that the assessments are fair for all groups of test takers, despite differences in characteristics that include, but are not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Unnecessary barriers can be reduced by following some fundamental rules:

- Measuring only knowledge or skills that are relevant to the intended construct
- Not angering, offending, upsetting, or otherwise distracting test takers
- Treating all groups of people with appropriate respect in test materials

These rules help ensure that the test content is fair for test takers as well as acceptable to the many stakeholders and constituent groups within Smarter Balanced member organizations. The more typical view is that bias and sensitivity guidelines apply primarily to the review of test items. However, fairness must be considered in all phases of test development and use. Smarter Balanced strongly rely on the *Bias and Sensitivity Guidelines* (ETS, 2012) in the development of the Smarter Balanced assessments, particularly in item writing and review. Items must comply with the Bias and Sensitivity Guidelines in order to be included in the Smarter Balanced assessments.

Smarter Balanced assessments are developed using the principles of evidence-centered design (ECD). ECD requires a chain of evidence-based reasoning that links test performance to the claims made about test takers. Fair assessments are essential to the implementation of ECD. If test items are not fair, then the evidence they provide means different things for different groups of students. Under those circumstances, the claims cannot be equally supported for all test takers, which is a threat to validity. As part of the validation process, all items are reviewed for issues of bias and sensitivity using the *Bias and Sensitivity Guidelines* (ETS, 2012) prior to being presented to students. This helps ensure that item responses reflect only knowledge of the intended content domain, are free of offensive or distracting material, and portray all groups in a respectful manner. When the guidelines are followed, item responses provide evidence that supports assessment claims.

11.1.2 Social Studies

MDE and DRC item writers and test developers follow documented bias and sensitivity guidelines to ensure that the items are fair for all groups of test takers, despite differences in characteristics that include, but are not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Test developers review all items included in M-STEP and other testing materials with these guidelines in mind.

Careful attention is given to item statistics (if available) throughout the test development process. As part of the test assembly process, attempts are made to avoid using or reusing items with poor statistics. Additional steps to reduce bias, including the use of content and bias committees comprised of Michigan educators, are described in more detail in Chapter 3 of this report.

The goal of fairness in assessment is to ensure that test materials are as free as possible from unnecessary barriers to the success of diverse groups of students.

11.2 Evaluating Bias through Differential Item Functioning (DIF)

An empirical approach known as DIF is used to examine items after they have been administered. The DIF statistics indicate the degree to which members of a particular subgroup perform better or worse than expected on each item as compared to the members of the reference group. Therefore, DIF flags do not necessarily indicate that an item is biased; rather, DIF flags indicate that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). The DIF procedures and results are described in this section. Note that items are not necessarily suppressed from operational scoring if they are flagged for DIF.

The position of DRC concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test. Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting the development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test developers' task is to create assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements into the performances on which the measurement is based. If these tests require that students have culturally specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

To lessen such biases, MDE and DRC strive to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above and in Chapter 3 of this report, careful attention is given during the test development and form construction processes to lessen the influence of these elements for large numbers of students (including the use of content and bias review committees). Unfortunately, in some cases, extraneous elements may continue to play a substantial role. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted during field testing and after each operational test administration. DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. For M-STEP, DIF is conducted for ELA, mathematics, and social studies using very similar procedures. Details in Sections 11.3.1 and 11.3.2 provide DIF results for the following subgroups:

- ***Gender:*** The focal group is female; the reference group is male.
- ***Race/Ethnicity:*** The focal groups are students whose race/ethnicity is reported as African American or Black, Hispanic or Latino, or Asian; the reference group is students whose race/ethnicity is reported as White.

- ***Disability status:*** The focal group is students who are identified as students with disabilities (SWD); the reference group is all others.
- ***English Proficiency status:*** The focal group is students who are identified as Limited English Proficiency (LEP); the reference group is all others.
- ***Socio-economic status:*** The focal group is students who are identified as economically disadvantaged (EconDis); the reference group is all others.

11.3 DIF Statistics

Two commonly used DIF statistics were applied to M-STEP items and are described here. They are (1) the Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) for dichotomously scored items and an extension of the $MH\chi^2$ (Mantel, 1963) for polytomously scored items, and (2) the standardized mean difference (SMD) effect size (ES) for polytomously scored items (Dorans & Schmitt, 1991).

For dichotomously scored items (e.g., MC items), the MH statistic is computed as follows (Camilli & Shepard, 1994):

$$MH\chi^2 = \frac{\left\{\left|\sum_{j=1}^{S}[A_j - E(A_j)]\right| - 1/2\right\}^2}{\sum_{j=1}^{S} VAR(A_j)} \quad (11.1)$$

where $VAR(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$ and $E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$.

In Equation 11.1, $A_j - E(A_j)$ represents the difference between the observed number and the expected number of correct responses on the item by the reference group members who have the jth score on the matching variable;[1] n_{Rj} and n_{Fj} represent the number of examinees in the reference and focal groups, respectively, for the jth score on the matching variable; m_{1j} represents the total number of examinees (both reference and focal) with the jth score on the matching variable and with a correct response on the current item; m_{0j} represents the total number of examinees with the jth score on the matching variable and with an incorrect response on the current item. The $MH\chi^2$ is evaluated against the standard χ^2 critical with one degree of freedom.

The $MH\chi^2$ does not indicate the strength of association of the relationship between item performance and group membership. The MH odds ratio can be computed to estimate the strength of this association. The resulting estimate represents the relative likelihood of success on a particular item for members of two different groups of examinees (Camilli, 2006). This odds ratio thus provides an estimate of effect size (ES) with a value of 1.0, indicating no DIF. A value greater than 1.0 indicates that, on average, the reference group members performed better than comparable focal group members did. A value less than 1.0 indicates that, on average, the reference group members performed worse than comparable focal group members did.

[1] Total observed score is used as the matching variable for DIF analysis here.

The odds of a correct response (i.e., proportion passing divided by proportion failing) is *P/Q* (i.e., *P/[1-P]*). The MH odds ratio is simply the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. The formula for its estimation is as follows (Camilli & Shepard, 1994, p. 116):

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j} \quad (11.2)$$

where $S = K - 1$ and represents the actual number of 2×2 contingency tables (assuming the tables have at least 1 person in each cell); K represents the number of items on the test; j signifies the jth score on the matching variable and runs from 0 to K.[2] For jth score category, A_j represents the number of reference group members with a correct response, B_j represents the number of reference group members with an incorrect response, C_j represents the number of focal group members with a correct response, and D_j represents the number of focal group members with an incorrect response. T_j represents the total number of examinees who have the jth score on the matching variable.

The corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups (i.e., the odds ratio is equal to 1):

$$H_0 : \alpha_{MH} = 1 \quad (11.3)$$

To make the odds ratio symmetrical around zero with its range located in the interval $-\infty$ to $+\infty$, the odds ratio is transformed into a log-odds ratio as follows (Camilli & Shepard, 1994, p.116):

$$\hat{\lambda}_{MH} = \log(\alpha_{MH}) \quad (11.4)$$

The natural logarithm transformation of this odds ratio is symmetrical around zero (where 0 indicates no DIF). This DIF measure is a signed index, where a positive value represents DIF in favor of the reference group and a negative value indicates DIF in favor of the focal group.

The variance of the log-odds ratio estimate (V_λ) is computed as follows (Camilli & Shepard, 1994, p. 121):

$$V_\lambda = \frac{\sum_{j=1}^{S} T_j^{-2}(A_j D_j + \alpha_{MH} B_j C_j)[A_j + D_j + \alpha_{MH}(B_j + C_j)]}{2(\sum_{j=1}^{S} A_j D_j / T_j)^2}. \quad (11.5)$$

The terms included in Equation 11.5 correspond to those presented for Equation 11.2. In practice, a standardized MH log-odds ratio is computed by dividing the estimate $\hat{\lambda}_{MH}$ by the estimated standard error. According to Penfield (2007, p.16), “A value greater than 2.0 or less than -2.0 may be considered evidence of the presence of DIF.”

[2] Although the value of the matching variable runs from 0 to K, the all correct (i.e., K) and all incorrect (i.e., 0) score categories are not included in the DIF analysis in order to avoid having a denominator equal to 0.

In addition, once $\hat{\lambda}_{MH}$ is obtained using Equation 11.4, the delta statistic (MH D-DIF, used by SBAC in flagging criteria) can be computed as

$$\text{MH D-DIF} = -2.35 \times \hat{\lambda}_{MH} \quad (11.6)$$

For polytomously scored items, an extension of the $\mathrm{MH}\chi^2$ procedure was computed (Mantel, 1963). The statistic is computed as follows (Zwick, Donaghue, & Grima, 1993, p. 239):

$$Mantel\ \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k VAR(F_k)} \quad (11.7)$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable and is defined as

$$F_k = \sum_t y_t n_{Ftk}, \quad (11.8)$$

and the expectation of F_k under the hypothesis of no association is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t\, n_{+tk}\,, \quad (11.9)$$

and the variance of F_k under the assumption of no association is

$$\mathrm{Var}(\mathrm{F_k}) = \frac{\mathrm{n_{R+k} n_{F+k}}}{\mathrm{n^2_{++k}(n_{++k} - 1)}} \left\{ \left(\mathrm{n_{++k}} \sum_t \mathrm{y_t^2 n_{+tk}} \right) - \left(\sum_t \mathrm{y_t\, n_{+tk}} \right)^2 \right\}. \quad (11.10)$$

Using the Mantel approach for ordered categories, the data are organized into a $2 \times T \times K$ contingency table, where T is the number of response categories and K is the number of levels of the matching variable. $y_1, y_2, \ldots, y_T$ represent the T scores that can be obtained on the item; n_{Rtk} and n_{Ftk} represent the number of examinees in the reference and focal groups, respectively, who are at the kth level of the matching variable and received an item score of y_t. The "+" denotes summation over a particular index (e.g., n_{R+k} denotes the total number of reference group members at the kth level of the matching variable). Under the null hypothesis of no association, the Mantel statistic has a chi-square distribution with one degree of freedom. For dichotomous items, the Mantel statistic reduces to the MH statistic (without the continuity correction).

In addition to the MH statistic, an ES was calculated by dividing the SMD statistics by the overall (i.e., focal and reference groups combined) standard deviation (SD) of the item scores: $\mathrm{ES} = \mathrm{SMDSD}$. The SMD compares the mean of the reference and focal groups, adjusting for the distribution of reference and focal group members on the matching variable (Zwick et al., 1993), which for these analyses is the M-STEP raw score. SMD is computed as follows (Zwick et al., 1993):

$$SMD = \sum_k p_{Fk\,(m_{Fk} - m_{Rk})} \quad (11.11)$$

where p_{Fk} is the proportion of the focal group members at the kth level of the matching variable m_{Fk} and m_{Rk} indicate mean item score for the focal group and the reference group at the kth level of the matching variable, respectively.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

11.3.1 Flagging Criteria and Results for ELA and Mathematics

For ELA and mathematics, according to Smarter Balanced (for more information, see the *Smarter Balanced 2017–2018 Technical Report* [2018]), the minimum case count for each of the two groups (i.e., the focal group and the reference group) was set at 100 and the minimum case count for the combined group was set to 400.

The following flagging criteria were used for dichotomously scored items (e.g., MC items):

- Moderate DIF: significant MH chi-square statistic ($\mathrm{p} < 0.05$) and $1.0 \leq |\mathrm{MH\ D\text{-}DIF}| < 1.5$
- Large DIF: significant MH chi-square statistic ($\mathrm{p} < 0.05$) and $|\mathrm{MH\ D\text{-}DIF}| \geq 1.5$

The following flagging criteria were used for polytomously scored items:

- Moderate DIF: if the extension of the MH statistic is significant ($\mathrm{p} < .05$) and $|\mathrm{ES}|$ is > 0.17 and ≤ 0.25.
- Large DIF: if the extension of the MH statistic is significant ($\mathrm{p} < .05$) and $|\mathrm{ES}| > 0.25$.

A positive $\mathrm{MH\ D\text{-}DIF}$ or ES value indicates that the item favors the focal group, while a negative value indicates that the item favors the reference group instead.

Table 11-1 shows the item counts for ELA and mathematics DIF analyses based on the 2019 M-STEP administration. Tables 11-2 and 11-3 summarize the number of items having moderate or large DIF flags (i.e., B or C) by grade for each focal/reference group that included at least 100 students for ELA and mathematics, respectively. For example, consider grade 3 ELA. There were 19 items (or 2.68% of all eligible items) flagged for moderate DIF when comparing the performance of academically similar males and females. Specifically, there were two items favoring males and 17 items favoring females.

Again, any items included on the M-STEP ELA and mathematics assessments (including those items flagged for DIF) have been thoroughly reviewed by MDE staff, DRC test development staff, and Smarter Balanced staff.

Table 11-1. Item Counts Used in Differential Item Functioning Analyses: ELA and Mathematics

Content Area	Grade	*N* Items	Female/ Male	Asian/ White	Black or African American/ White	Hispanic or Latino/ White	SWD/ Non-SWD	LEP/ Non-LEP	EconDis/ Non-EconDis
ELA	3	777	709	665	692	671	696	696	705
ELA	4	771	709	664	694	674	693	689	705
ELA	5	711	643	614	624	617	627	624	637
ELA	6	692	644	612	621	615	623	621	634
ELA	7	599	563	535	547	536	548	545	557
Math	3	1082	970	630	768	662	788	781	963
Math	4	1182	1014	693	739	694	741	731	1016
Math	5	1117	982	619	808	708	961	953	985
Math	6	1012	894	764	870	831	877	870	886
Math	7	940	851	479	619	515	642	637	845

Table 11-2. Number of Differential Item Functioning Flagged Items: ELA

Grade	DIF Category	Female/ Male	Asian/White	Black or African American/ White	Hispanic or Latino/ White	Disabilities/ Without Disabilities	LEP/ Non-LEP	EconDis/ Non-EconDis
3	b-	2	23	13	6	1	11	5
3	b+	17	17	18	17	29	26	17
3	c-	1	5	0	0	1	1	0
3	c+	0	3	2	0	3	2	0
4	b-	8	13	6	7	10	9	1
4	b+	14	20	27	11	38	24	9
4	c-	1	5	0	0	0	1	0
4	c+	0	1	1	1	5	2	1
5	b-	8	16	6	3	8	7	0
5	b+	17	14	20	15	41	31	9
5	c-	1	3	0	0	1	0	0
5	c+	3	1	1	2	9	9	0
6	b-	6	9	7	4	6	3	0
6	b+	10	14	23	18	40	20	7
6	c-	2	3	0	0	0	1	0
6	c+	4	3	2	1	3	7	0
7	b-	9	12	10	4	7	12	0
7	b+	29	12	23	9	33	29	4
7	c-	2	0	0	0	0	0	0
7	c+	3	3	2	0	2	4	0

Table 11-3. Number of Differential Item Functioning Flagged Items: Mathematics

Grade	DIF Category	Female/ Male	Asian/White	Black or African American/ White	Hispanic or Latino/ White	Disabilities/ Without Disabilities	LEP/ Non-LEP	EconDis/ Non-EconDis
3	b-	55	33	47	12	12	19	10
3	b+	35	20	50	7	24	21	18
3	c-	1	14	3	2	0	2	0
3	c+	5	4	7	3	1	3	4
4	b-	34	25	12	7	13	9	12
4	b+	45	16	31	15	18	19	22
4	c-	7	5	3	1	0	2	0
4	c+	8	0	7	3	3	1	3
5	b-	43	66	24	12	17	23	9
5	b+	45	14	34	4	22	19	13
5	c-	5	23	6	0	5	3	0
5	c+	6	6	12	2	1	5	0
6	b-	53	53	30	6	17	19	8
6	b+	49	13	46	11	11	26	17
6	c-	3	22	3	1	4	4	1
6	c+	12	10	8	1	3	9	1
7	b-	20	46	28	2	6	20	6
7	b+	27	8	28	7	10	23	22
7	c-	2	29	4	0	4	4	0
7	c+	4	5	11	2	2	9	0

11.3.2 Flagging Criteria and Results for Social Studies

For social studies, the minimum case count was 30 for each of the two groups (i.e., the reference group and the focal group). The following flagging criteria, adapted from Penfield (2007), were used:

- Negligible DIF (a): if either MH common log-odds ratio ($\hat{\lambda}_{MH}$) is not significantly different from zero or $|\hat{\lambda}_{MH}| < 0.426$
- Moderate DIF (b): if $\hat{\lambda}_{MH}$ is significantly different from zero and $|\hat{\lambda}_{MH}| \geq 0.426$ and either (a) $|\hat{\lambda}_{MH}| \leq 0.638$, or (b) $|\hat{\lambda}_{MH}|$ is not significantly greater than 0.426
- Large DIF (C): if $|\hat{\lambda}_{MH}|$ is significantly greater than 0.426 and $|\hat{\lambda}_{MH}| > 0.638$.

Table 11-4 shows the item counts for social studies DIF analyses. Tables 11-5 and 11-6 summarize the number of items having moderate and large DIF flags (i.e., b or c). For example, consider grade 11 social studies. There was 1 item (or 1.3% of all items) flagged for large Female vs. Male, favoring Female.

Again, any items included on the M-STEP social studies assessments (including those items flagged for DIF) have been thoroughly reviewed by MDE staff, DRC test development staff, and Michigan content/bias committee members.

Table 11-4. Item Counts used in Differential Item Functioning Analyses: Social Studies[3]

Content Area	Grade	*N* Items	Female/ Male	Asian/ White	Black or African American/ White	Hispanic or Latino/ White	SWD/ Non-SWD	LEP/ Non-LEP	EconDis/ Non-EconDis
Social Studies	5	90	90	90	90	90	90	90	90
Social Studies	8	88	88	88	88	88	88	88	88
Social Studies	11	76	76	76	76	76	76	76	76

Table 11-5. Number of Differential Item Functioning Flagged Items: Social Studies

Grade	DIF Category	Female/ Male	Asian/White	Black or African American/ White	Hispanic or Latino/ White	Disabilities/ Without Disabilities	LEP/ Non-LEP	EconDis/ Non-EconDis
5	b-	0	0	4	1	2	2	1
5	b+	0	1	4	0	1	3	1
5	c-	0	1	0	0	0	0	0
5	c+	0	1	1	0	0	0	0
8	b-	1	0	1	0	0	2	1
8	b+	2	0	1	1	2	4	1
8	c-	0	0	1	0	0	0	0
8	c+	0	0	2	0	0	0	0
11	b-	2	1	4	0	2	6	2
11	b+	3	3	3	0	1	4	2
11	c-	0	1	0	0	0	1	0
11	c+	1	0	1	0	0	0	0

[3] For the 2019 administration, DIF analyses were conducted separately by mode. Some OP items appear on both online forms and paper/pencil forms, but it may occur that the very same item was not flagged for DIF with online forms but was flagged for paper/pencil forms. For that reason, here each item/mode unique combination was counted as one item.

11.4 Summary

In summary, the overall purpose of this chapter is to address fairness concerns that are relevant to the administration of M-STEP. The information in this chapter supports multiple best practices of the testing industry and particularly the following AERA, APA, & NCME (2014) standards:

- Standard 3.1—Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
- Standard 3.2—Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.
- Standard 3.3—Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.
- Standard 3.4—Test takers should receive comparable treatment during the test administration and scoring process.
- Standard 3.5—Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.
- Standard 3.6—Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.

Chapter 12: Reliability and Evidence of Construct-Related Validity

This chapter presents evidence supporting construct-related validity. Part of the test validity argument is that scores must be consistent and precise enough to be useful for the intended purposes. The concepts of reliability and precision are examined through analysis of measurement error in simulated and operational (OP) conditions.

This chapter demonstrates M-STEP's adherence to AERA, APA, & NCME (2014) *Standards* 2.0, 2.1, 2.3, 2.13, 2.14, 2.16, 2.19, and 4.3. Each standard will be discussed in the pertinent section of this chapter.

12.1 Reliability

Reliability refers to the consistency of the students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Often, however, it is impractical to administer multiple forms of the test, and reliability is estimated on a single administration of the test. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across items within a test during a single test administration (Crocker & Algina, 1986). Reliability is a necessary but insufficient condition of validity.

The AERA, APA, & NCME (2014) *Standards* indicates the following:

> The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer 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. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory [IRT] information functions, or various indices of classification consistency). (p. 33)

In accordance with the AERA, APA, & NCME (2014) *Standards* and in developing and maintaining tests of the highest quality, the reliability of each M-STEP test has been calculated.

There are several specific AERA, APA, & NCME (2014) standards that this chapter addresses. These include *Standards* 2.0, 2.3, 2.13, and 2.19. Each standard is articulated below.

Standard 2.0 Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use. (p. 42)

Standard 2.3 For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported. (p. 43)

The total score reliabilities are reported below in Sections 12.1.5 through 12.1.7 of this chapter. The overall standard errors of measurement (SEMs) and conditional standard errors of measurement (CSEMs) by decile are presented in Section 12.1.5.

Standard 2.13 The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (p. 45)

The SEM based on scale scores and the CSEM based on scale scores are discussed below in Section 12.1.5.

Standard 2.19 Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported. (p. 47)

12.1.1 Reliability and Standard Error of Measurement

According to the classical true score theory, which is a fundamental component of the CTT, an observed score is a sum of two parts—a random component of true score (T) and a random component of error score (E), or mathematically, $X = T + E$ (McDonald, 1999). This model has the following properties: (1) the expected error score is zero, (2) the correlation between the true score and the error score is zero, and (3) the correlation between the error scores on different but parallel forms is zero (Lord & Novick, 1968).

Based on this model, a student's observed test score is an imprecise estimate of his or her actual ability because a portion of that score is attributable to random error. A fundamental theoretical quantity in test theory, the *reliability coefficient* of observed scores is defined as the ratio of the variance of true scores to the variance of observed scores. Tests are therefore most reliable when the proportion of observed score variance that may be attributed to error variance is minimized. According to McDonald (1999), test-retest methods, parallel or alternate-form methods, and internal analysis are the three recognized methods for estimating the reliability coefficient.

Due to practical difficulties in applying the first two above-mentioned methods, only the internal consistency reliability approach is described here. Estimates of internal consistency reliability involve "dividing the test into two or more constituent parts and in some way estimating reliability from the consistency of performance across these part-tests" (Haertel, 2006, p. 71).

12.1.2 Cronbach's Coefficient Alpha

Historically, various internal consistency reliability estimates have been proposed, but the most widely used, for fixed forms, is Cronbach's (1951) coefficient alpha (Haertel, 2006). Using sample statistics, it is computed as follows (adapted from Haertel, 2006):

$$\alpha = \frac{I}{I-1}\left(1 - \frac{\sum_{i=1}^{I} s_i{}^2}{s_X{}^2}\right) \text{(12.1)}$$

where l represents the number of items on the test, S_{i}^2 represents the sample variance of item i, and S_X^2 represents the sample variance of the total raw score.

The use of coefficient alpha has several theoretical advantages (Haertel, 2006). First, since it equals the mean of all possible split-half reliability coefficients, which is another estimate of internal consistency reliability that involves the division of the total test into two "parallel" subtests, the use of coefficient alpha avoids the arbitrary choice of a split or division. Second, it is mathematically equivalent to one of the lower bounds of the theoretical reliability coefficient. The implication of this is that the theoretical reliability coefficient is higher than the observed coefficient alpha.

12.1.3 Standard Error of Measurement

SEM is related to reliability and is calculated with sample statistics as follows (Hays, 1994, p. 617):

$$\mathrm{SEM}(X) = s_X\sqrt{1 - r_{XX'}} \text{ (12.2)}$$

where $\mathrm{SEM}(X)$ represents the estimated SEM of the observed test score X, S_X denotes the estimated standard deviation (i.e., sample standard deviation) of the observed score, and $r_{XX'}$ represents the estimated reliability coefficient of a test. In this report, the observed coefficient alpha is used as the estimated reliability coefficient for social studies.

According to Equation 12.2, the SEM is inversely related to the reliability of a test: For any standard deviation of the observed score, the SEM decreases when the reliability coefficient increases. Thus, when an SEM is small, one has more confidence in the accuracy, or precision, of the observed test scores.

12.1.4 Marginal Reliability for ELA and Mathematics

In a CAT administration, each student receives a different test form; therefore, the calculation of coefficient alpha is not applicable. An observed reliability can be derived from SEMs, which are computed from the test form each student took. The method of standard error calculation for both total and score reporting category scores, as described in the Smarter Balanced Scoring Specifications for 2014–2015 (AIR, 2014), is displayed below:

The standard error (SE) for student i is

$$SE(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \quad (12.3)$$

where $l(\theta_i)$ is the test information function for student i, calculated as

$$I(\theta_i) = \sum_{j=1}^{I} D^2 a_j^2 \left(\frac{\sum_{l=1}^{m_j} l^2 Exp(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk}))}{1 + \sum_{l=1}^{m_j} Exp(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk}))} - \left(\frac{\sum_{l=1}^{m_j} l Exp(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk}))}{1 + \sum_{l=1}^{m_j} Exp(\sum_{k=1}^{l} Da_j(\theta_i - b_{jk}))} \right)^2 \right) \quad (12.4)$$

where m_j is the maximum possible score point (starting from 0) for the jth item, D is the scale factor, 1.7. Values of a_j and b_{jk} are item parameters for item j and score level k.

SE is calculated based only on the answered items. The upper bound of the SE is set to 2.5 on the theta metric. Any value larger than 2.5 is truncated at 2.5 on the theta metric.

12.1.5 Observed Reliability, SEM, and CSEM for ELA and Mathematics

The marginal reliability for ELA and mathematics was calculated using the 2019 Michigan administration data. The results are presented in Table 12-1.

Table 12-1. ELA and Mathematics Summative Scale-Score Marginal Reliability Estimates

Content Area	Grade	*N*	Mean # Items	SD(SS)	Mean SEM	Marginal Reliability
ELA	3	100,255	45.22	25.97	6.11	0.94
ELA	4	101,838	45.03	26.23	6.26	0.94
ELA	5	104,530	45.45	27.24	6.57	0.94
ELA	6	108,397	45.13	26.40	6.68	0.93
ELA	7	108,653	45.06	26.43	7.17	0.93
Mathematics	3	100,261	36.00	27.40	5.58	0.96
Mathematics	4	101,929	36.00	25.73	5.37	0.95
Mathematics	5	104,553	36.00	26.47	5.91	0.94
Mathematics	6	108,449	36.00	26.04	5.66	0.95
Mathematics	7	108,702	36.00	26.65	6.13	0.94

SD(SS) = standard deviation of scale score

Table 12-2 shows that the marginal reliability varies by overall score levels. All students take a similar number of items, but the information delivered by the items differs. The most information occurs where the pool item difficulty and students' ability match the best with abundant items for selection. As shown in Figures 8-1 and 8-2, Smarter Balanced pools, used by Michigan, are difficult relative to the student ability levels of the state population. Consistently, as shown in Table 12-2, students with lower scores (e.g., deciles 1 and 2) have lower reliability than those with higher scores (e.g., deciles 8 and 9).

Table 12-2. Marginal Reliability Overall and by Decile for ELA and Mathematics

Content Area	Grade	*N*	Var	Overall	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
ELA	3	100,255	674.63	0.94	0.91	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
ELA	4	101,838	688.21	0.94	0.92	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.93
ELA	5	104,530	742.10	0.94	0.93	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.93	0.93
ELA	6	108,397	697.20	0.93	0.89	0.93	0.93	0.94	0.94	0.95	0.95	0.95	0.95	0.93
ELA	7	108,653	698.39	0.93	0.88	0.92	0.93	0.93	0.93	0.94	0.94	0.93	0.93	0.92
Mathematics	3	100,261	750.97	0.96	0.92	0.95	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.96
Mathematics	4	101,929	661.96	0.95	0.90	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.96
Mathematics	5	104,553	700.47	0.94	0.84	0.91	0.93	0.95	0.96	0.96	0.97	0.98	0.97	0.97
Mathematics	6	108,449	677.98	0.95	0.87	0.93	0.94	0.95	0.96	0.96	0.96	0.97	0.97	0.97
Mathematics	7	108,702	710.08	0.94	0.79	0.91	0.93	0.95	0.95	0.96	0.97	0.98	0.98	0.98

Because of the CSEM differences by score level, demographic groups with lower average scores tend to have lower reliability than the population as a whole. Due to the small sample sizes of some of the subgroups (e.g., American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander), corresponding results should be interpreted with caution. Tables 12-3 to 12-6 show marginal reliability by demographic group and the MSE.

Table 12-3. Marginal Reliability of Total Summative Scores by Ethnic Group—ELA

Grade	Group	*N*	Var	MSE	Marginal Reliability
3	All	100,255	674.63	6.11	0.94
3	American Indian or Alaska Native	561	533.58	6.12	0.93
3	Asian	3,506	658.77	6.08	0.94
3	Black or African American	18,788	529.96	6.39	0.92
3	Hispanic or Latino	8,374	570.22	6.11	0.93
3	Native Hawaiian or Other Pacific Islander	97	620.66	6.03	0.94
3	Two or More Races	4,884	659.41	6.10	0.94
3	White	64,045	613.46	6.03	0.94
4	All	101,838	688.21	6.26	0.94
4	American Indian or Alaska Native	599	563.07	6.25	0.93
4	Asian	3,508	674.82	6.31	0.94
4	Black or African American	18,794	555.30	6.42	0.92
4	Hispanic or Latino	8,287	578.41	6.23	0.93
4	Native Hawaiian or Other Pacific Islander	85	601.17	6.18	0.94
4	Two or More Races	4,701	670.70	6.26	0.94
4	White	65,864	628.14	6.22	0.94
5	All	104,553	700.47	5.91	0.94
5	American Indian or Alaska Native	627	585.43	6.16	0.93
5	Asian	3,579	687.97	5.13	0.96
5	Black or African American	18,547	534.49	7.28	0.89
5	Hispanic or Latino	8,626	572.18	6.22	0.93
5	Native Hawaiian or Other Pacific Islander	96	573.78	5.63	0.94
5	Two or More Races	4,778	677.29	6.03	0.94
5	White	68,300	614.17	5.52	0.95
6	All	108,449	677.98	5.66	0.95
6	American Indian or Alaska Native	716	567.96	5.92	0.93
6	Asian	3,700	702.40	4.99	0.96
6	Black or African American	18,788	536.52	6.75	0.91
6	Hispanic or Latino	8,929	566.26	5.95	0.93
6	Native Hawaiian or Other Pacific Islander	94	618.74	5.44	0.95
6	Two or More Races	4,595	681.88	5.81	0.95
6	White	71,627	579.18	5.36	0.95
7	All	108,702	710.08	6.13	0.94
7	American Indian or Alaska Native	685	602.74	6.37	0.92

Grade	Group	*N*	Var	MSE	Marginal Reliability
7	Asian	3,696	777.94	5.00	0.96
7	Black or African American	18,672	534.97	7.79	0.87
7	Hispanic or Latino	8,772	597.61	6.61	0.92
7	Native Hawaiian or Other Pacific Islander	105	696.21	6.02	0.94
7	Two or More Races	4,559	694.76	6.35	0.93
7	White	72,213	621.29	5.68	0.94

Table 12-4. Marginal Reliability of Total Summative Scores by Ethnic Group—Mathematics

Grade	Group	*N*	Var	MSE	Marginal Reliability
3	All	100,261	750.97	5.58	0.96
3	American Indian or Alaska Native	561	533.33	5.53	0.94
3	Asian	3,596	699.63	5.43	0.96
3	Black or African American	18,809	633.39	6.07	0.94
3	Hispanic or Latino	8,219	623.22	5.64	0.95
3	Native Hawaiian or Other Pacific Islander	97	773.18	5.57	0.96
3	Two or More Races	4,888	722.02	5.59	0.96
3	White	64,091	659.63	5.44	0.95
4	All	101,929	661.96	5.37	0.95
4	American Indian or Alaska Native	601	498.48	5.40	0.94
4	Asian	3,596	647.69	5.15	0.96
4	Black or African American	18,818	542.86	5.98	0.93
4	Hispanic or Latino	8,201	550.82	5.47	0.94
4	Native Hawaiian or Other Pacific Islander	86	627.63	5.51	0.95
4	Two or More Races	4,702	665.35	5.44	0.95
4	White	65,925	572.27	5.18	0.95
5	All	104,553	700.47	5.91	0.94
5	American Indian or Alaska Native	627	585.43	6.16	0.93
5	Asian	3,579	687.97	5.13	0.96
5	Black or African American	18,547	534.49	7.28	0.89
5	Hispanic or Latino	8,626	572.18	6.22	0.93
5	Native Hawaiian or Other Pacific Islander	96	573.78	5.63	0.94
5	Two or More Races	4,778	677.29	6.03	0.94
5	White	68,300	614.17	5.52	0.95
6	All	108,449	677.98	5.66	0.95
6	American Indian or Alaska Native	716	567.96	5.92	0.93
6	Asian	3,700	702.40	4.99	0.96
6	Black or African American	18,788	536.52	6.75	0.91
6	Hispanic or Latino	8,929	566.26	5.95	0.93
6	Native Hawaiian or Other	94	618.74	5.44	0.95
6	Two or More Races	4,595	681.88	5.81	0.95
6	White	71,627	579.18	5.36	0.95

Grade	Group	*N*	Var	MSE	Marginal Reliability
7	All	108,702	710.08	6.13	0.94
7	American Indian or Alaska Native	685	602.74	6.37	0.92
7	Asian	3,696	777.94	5.00	0.96
7	Black or African American	18,672	534.97	7.79	0.87
7	Hispanic or Latino	8,772	597.61	6.61	0.92
7	Native Hawaiian or Other Pacific Islander	105	696.21	6.02	0.94
7	Two or More Races	4,559	694.76	6.35	0.93
7	White	72,213	621.29	5.68	0.94

Table 12-5. Marginal Reliability of Total Summative Scores by Group—ELA

Grade	Group	*N*	Var	MSE	Marginal Reliability
3	Economically Disadvantaged	56,092	589.42	6.19	0.93
3	LEP	9,513	560.45	6.11	0.93
3	Disabilities	11,748	518.90	6.33	0.92
3	All	100,255	674.63	6.11	0.94
4	Economically Disadvantaged	56,117	598.27	6.30	0.93
4	LEP	8,906	531.00	6.26	0.93
4	Disabilities	11,959	534.09	6.43	0.92
4	All	101,838	688.21	6.26	0.94
5	Economically Disadvantaged	56,442	634.73	6.51	0.93
5	LEP	7,688	489.17	6.45	0.91
5	Disabilities	12,349	523.00	6.58	0.92
5	All	104,530	742.10	6.57	0.94
6	Economically Disadvantaged	57,280	598.38	6.84	0.92
6	LEP	6,446	418.45	7.07	0.88
6	Disabilities	12,161	447.34	7.25	0.88
6	All	108,397	697.20	6.68	0.93
7	Economically Disadvantaged	55,647	607.45	7.29	0.91
7	LEP	6,520	437.11	7.48	0.87
7	Disabilities	11,787	480.92	7.69	0.87
7	All	108,653	698.39	7.17	0.93

Table 12-6. Marginal Reliability of Total Summative Scores by Group—Mathematics

Grade	Group	N	Var	MSE	Marginal Reliability
3	Economically Disadvantaged	56,025	671.24	5.76	0.95
3	LEP	9,494	667.44	5.58	0.95
3	Disabilities	11,798	794.21	6.28	0.95
3	All	100,261	750.97	5.58	0.96
4	Economically Disadvantaged	56,111	581.77	5.60	0.94
4	LEP	8,976	579.34	5.48	0.95
4	Disabilities	12,009	653.97	6.22	0.94
4	All	101,929	661.96	5.37	0.95
5	Economically Disadvantaged	56,389	598.61	6.50	0.92
5	LEP	7,745	565.79	6.47	0.92
5	Disabilities	12,374	604.81	7.71	0.89
5	All	104,553	700.47	5.91	0.94
6	Economically Disadvantaged	57,267	599.43	6.16	0.93
6	LEP	6,459	555.20	6.43	0.92
6	Disabilities	12,162	604.73	7.35	0.90
6	All	108,449	677.98	5.66	0.95
7	Economically Disadvantaged	55,639	607.32	6.93	0.91
7	LEP	6,572	567.14	7.48	0.89
7	Disabilities	11,791	540.42	8.69	0.84
7	All	108,702	710.08	6.13	0.94

In addition to the SEM, the CSEM expresses the degree of measurement error in scale-score units and are conditioned on the ability of the student. The CSEM is reported in support of AERA, APA, & NCME (2014) Standard 2.14, which states the following:

> When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 46)

The CSEMs are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985); therefore, Equations 12.3 and 12.4 are used to calculate the CSEM.

In further compliance with Standard 2.14, Table 12-7 shows the median CSEM near the achievement level cut scores for ELA and mathematics.

Table 12-7. Conditional Standard Errors of Measurement near (±10 Points) Achievement Level Cut Scores, ELA and Mathematics

Content Area	Level—Cut Score	Grade	*N*	Median	Standard Deviation
ELA	1—2	3	12,888	5.98	0.19
ELA	2—3	3	13,154	5.55	0.50
ELA	3—4	3	12,405	5.87	0.34
ELA	1—2	4	13,114	6.05	0.22
ELA	2—3	4	12,453	6.01	0.10
ELA	3—4	4	13,414	6.01	0.08
ELA	1—2	5	13,146	6.08	0.27
ELA	2—3	5	11,752	6.21	0.41
ELA	3—4	5	12,337	6.93	0.25
ELA	1—2	6	14,144	6.63	0.49
ELA	2—3	6	13,503	6.09	0.28
ELA	3—4	6	10,478	6.19	0.40
ELA	1—2	7	14,213	6.95	0.35
ELA	2—3	7	14,580	6.72	0.45
ELA	3—4	7	9,940	6.98	0.15
Mathematics	1—2	3	12,643	5.59	0.50
Mathematics	2—3	3	14,093	5.06	0.24
Mathematics	3—4	3	11,853	5.01	0.10
Mathematics	1—2	4	13,689	5.39	0.49
Mathematics	2—3	4	15,079	4.97	0.18
Mathematics	3—4	4	11,349	4.65	0.48
Mathematics	1—2	5	13,731	5.86	0.44
Mathematics	2—3	5	14,155	4.53	0.50
Mathematics	3—4	5	11,570	4.18	0.39
Mathematics	1—2	6	15,846	5.80	0.42
Mathematics	2—3	6	16,078	4.94	0.25
Mathematics	3—4	6	12,780	4.24	0.43
Mathematics	1—2	7	15,927	6.13	0.38
Mathematics	2—3	7	13,545	4.86	0.36
Mathematics	3—4	7	12,439	4.02	0.15

When using a CAT, the CSEM will vary for the same scale score; therefore, it is necessary to report averages. Table 12-8 presents the overall average CSEM and the average CSEM by scale-score decile for ELA and mathematics.

Table 12-8. Overall Average CSEM and Average CSEM by Decile, ELA and Mathematics

Content Area	Grade	Overall SEM	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
ELA	3	6.11	7.61	6.33	6.02	5.93	5.71	5.55	5.58	5.80	5.99	6.52
ELA	4	6.26	7.38	6.20	6.06	6.05	6.02	6.01	6.01	6.01	6.04	6.90
ELA	5	6.57	7.34	6.28	6.08	6.08	6.12	6.23	6.49	6.82	6.96	7.29
ELA	6	6.68	8.55	7.17	6.88	6.51	6.25	6.10	6.05	6.06	6.17	6.98
ELA	7	7.17	8.89	7.50	7.04	6.85	6.75	6.72	6.72	6.79	6.96	7.35
Mathematics	3	5.58	7.75	6.03	5.69	5.32	5.13	5.04	5.02	5.01	5.02	5.68
Mathematics	4	5.37	7.90	5.98	5.34	5.07	5.02	4.99	4.82	4.59	4.69	5.15
Mathematics	5	5.91	10.29	7.87	6.85	5.97	5.14	4.96	4.48	4.14	4.24	4.82
Mathematics	6	5.66	9.36	6.94	6.13	5.81	5.15	5.01	4.94	4.54	4.19	4.32
Mathematics	7	6.13	11.85	8.12	6.93	6.13	5.71	5.07	4.82	4.15	4.02	4.06

Figures 12-1 through 12-10 display the CSEM curves by grade and content area. The dashed vertical lines represent the cut scores. The CSEM tends to be higher at the ends of the scale-score range. The measurement error increases when there are few items at a particular ability level. The figures show that the CSEM tends to be at its minimum around cut scores between Levels 2 and 3 and Levels 3 and 4.

Figure 12-1. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 3 English Language Arts

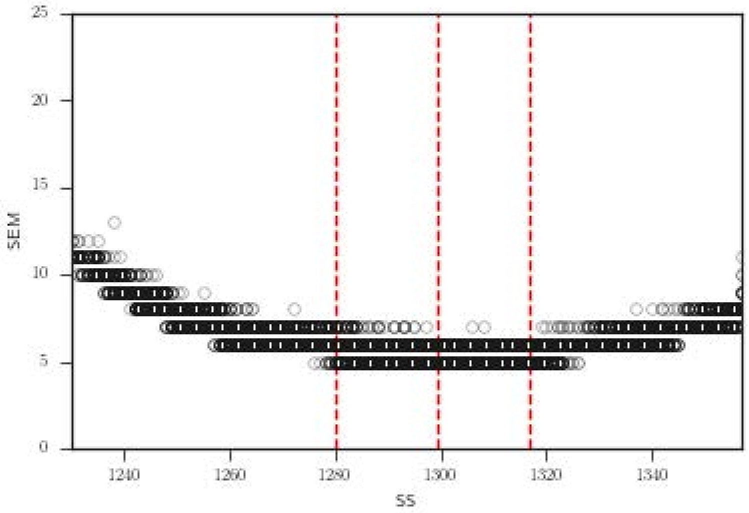


Figure 12-2. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 4 English Language Arts

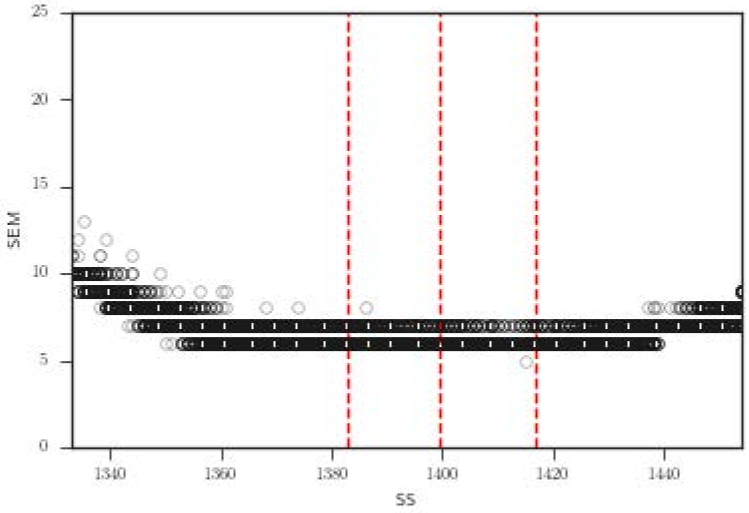


Figure 12-3. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 5 English Language Arts

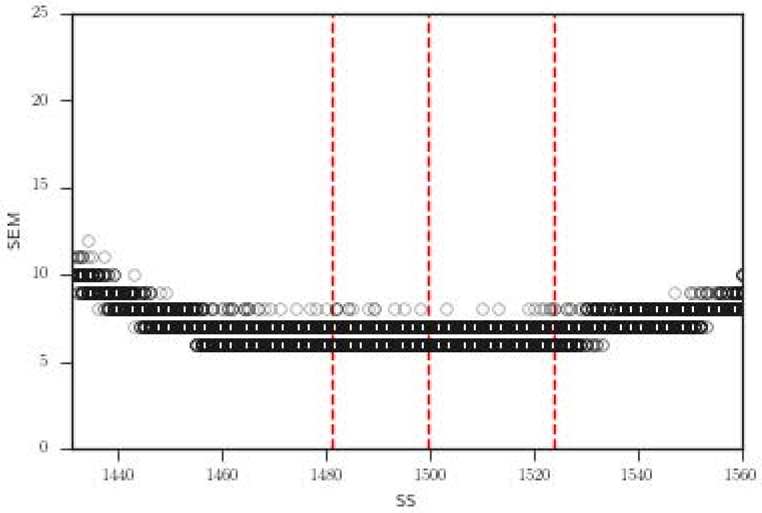


Figure 12-4. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 6 English Language Arts

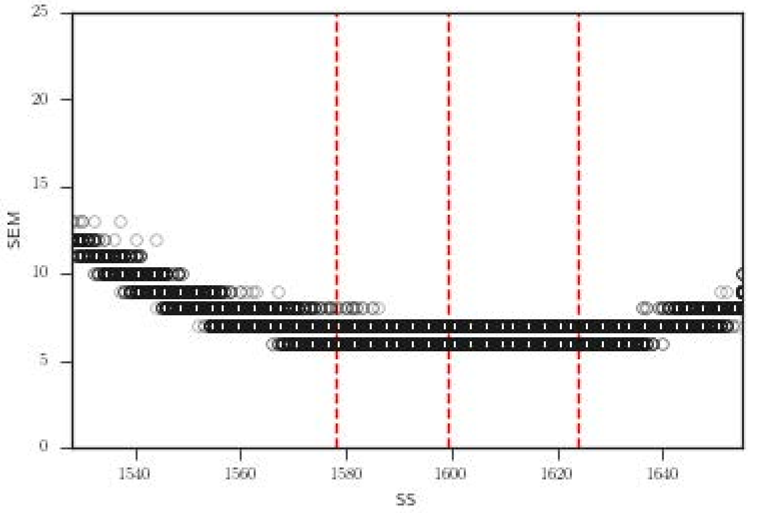


Figure 12-5. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 7 English Language Arts

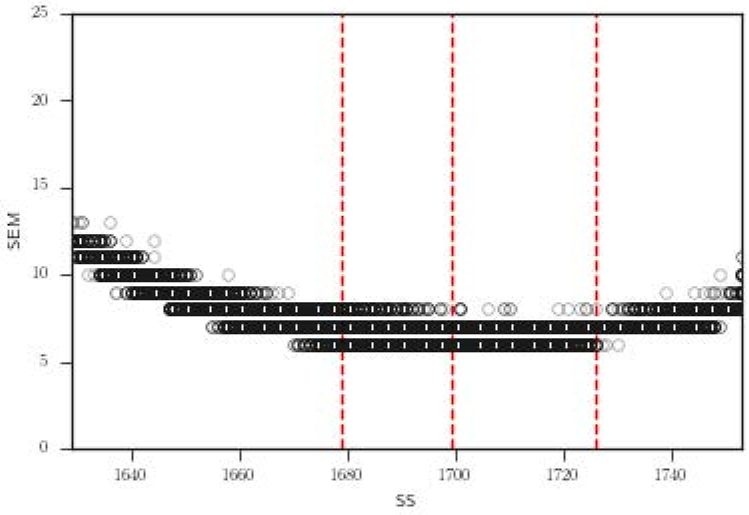


Figure 12-6. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 3 Mathematics

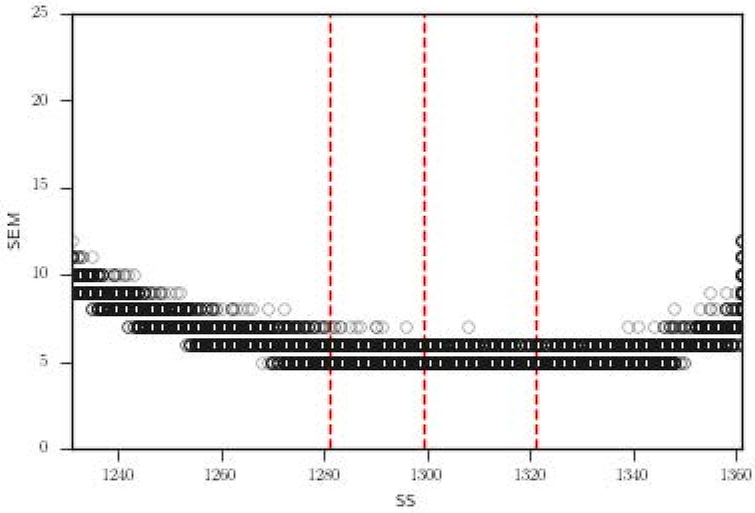


Figure 12-7. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 4 Mathematics

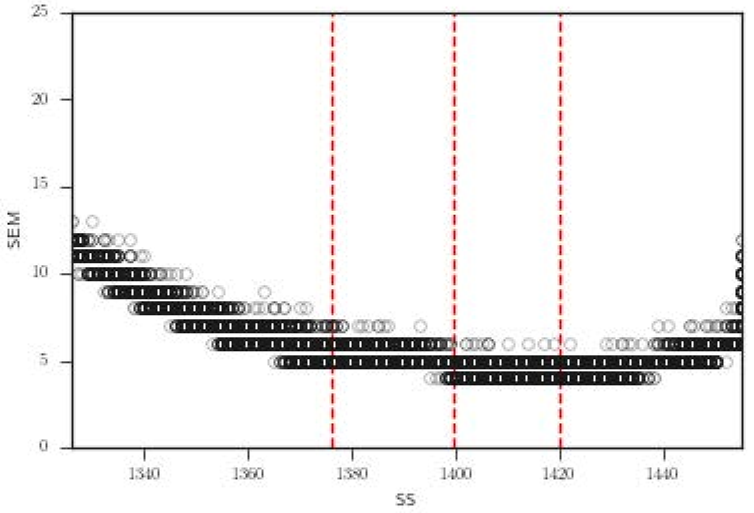


Figure 12-8. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 5 Mathematics

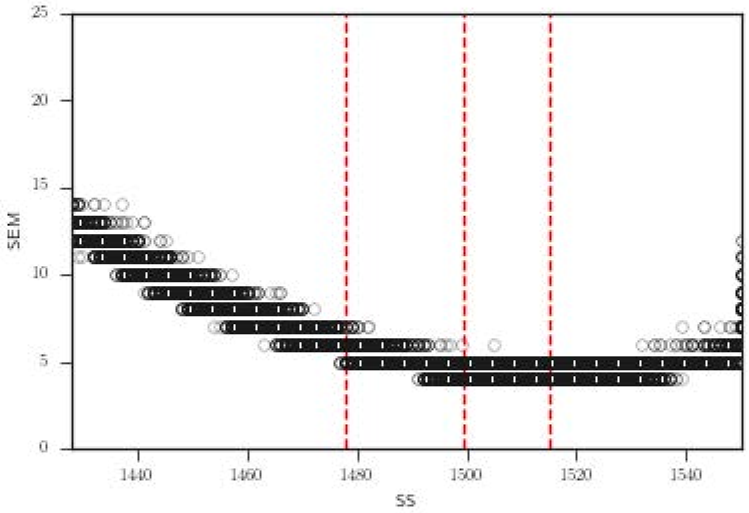


Figure 12-9. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 6 Mathematics

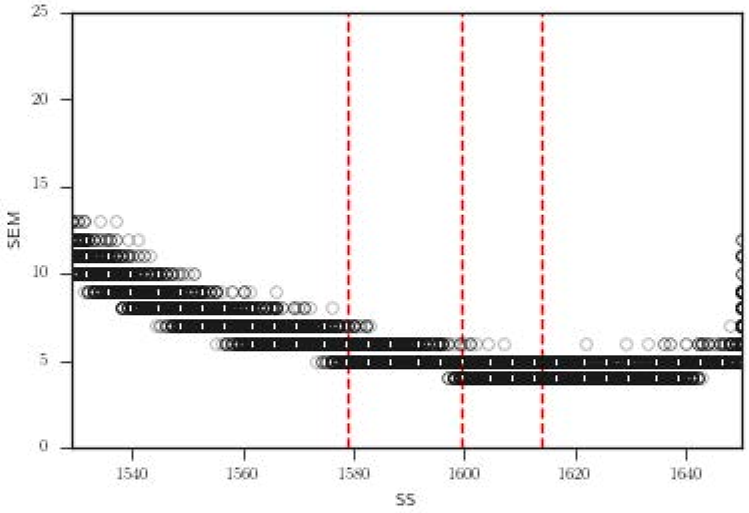
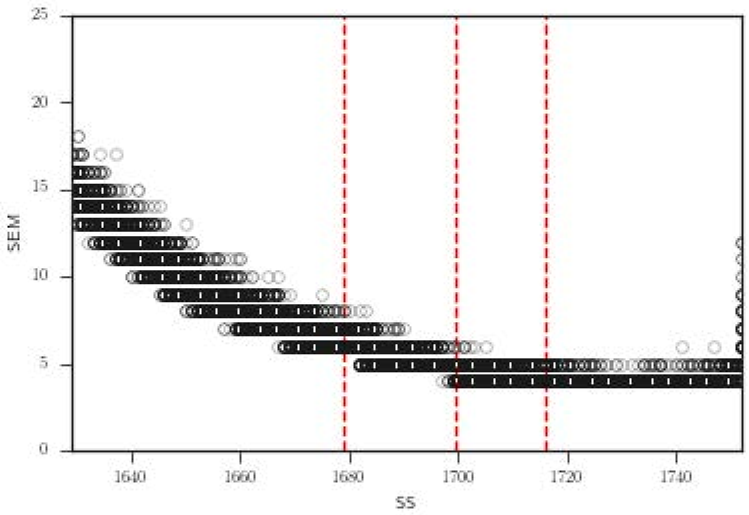


Figure 12-10. Conditional Standard Errors of Measurement for Overall Scale Scores, Grade 7 Mathematics



Smarter Balanced supports using fixed-form paper/pencil tests in schools that lack computer capacity or administering them to address potential religious concerns associated with using technology for assessments. Since the paper/pencil tests for ELA and mathematics consist of Smarter Balanced items, and there are few students who take the paper/pencil forms, DRC has chosen to be consistent with ELA and mathematics online counterparts by calculating the marginal reliability for those forms using Equation 12.6.

Table 12-9 shows the marginal reliability for the paper/pencil forms. As expected, overall estimated reliability coefficients are high and in the acceptable range for a large-scale, high-stakes test.

Table 12-9. Fixed-Form Marginal Reliability: ELA and Mathematics

Content Area	Grade	Number of Items per Form	Reliability	SEM
ELA	3	42	0.91	8.26
ELA	4	42	0.90	8.37
ELA	5	42	0.87	9.02
ELA	6	42	0.89	8.80
ELA	7	42	0.89	8.52
Mathematics	3	36	0.89	8.43
Mathematics	4	36	0.89	8.44
Mathematics	5	36	0.87	8.15
Mathematics	6	36	0.82	10.28
Mathematics	7	36	0.83	10.09

12.1.6 Reliability of Claims for ELA and Mathematics

Scale-score summary statistics (i.e., mean and standard deviation), marginal reliability coefficients, and mean CSEM were computed for each of the claims by grade and content area using M-STEP data. These statistics are presented in Tables 12-10 and 12-11 for ELA and mathematics, respectively. Reliability indices are a function of the number of test items. As expected, reliability coefficients are lower for a claim assessed by a small number of items compared to a claim assessed by a larger number of items. Consequently, the reliability for claims with larger CSEMs is lower than those with smaller CSEMs. These CSEMs are reported in the scale-score metric.

Table 12-10. Reliability, Mean, Standard Deviation, and Conditional Standard Error of Measurement (CSEM) of ELA Claims

Grade	Claim No.	Claim	Student *N* Count	Number of Items	Mean	Std. Dev.	Reliability	Mean CSEM
3	1	Reading	100,252	15-16	1295.48	27.43	0.84	10.87
3	2	Writing	100,252	13	1291.23	28.66	0.82	12.01
3	3	Listening	100,252	8-9	1297.94	36.59	0.62	22.60
3	4	Research	100,252	8	1291.89	33.93	0.76	16.67
4	1	Reading	01,836	15-16	1394.68	28.25	0.83	11.76
4	2	Writing	101,836	13	1393.16	28.79	0.84	11.67
4	3	Listening	101,836	8-9	1399.27	33.69	0.62	20.87
4	4	Research	101,836	8	1392.16	34.69	0.75	17.46
5	1	Reading	103,807	15-16	1496.75	29.05	0.82	12.23
5	2	Writing	103,807	13	1492.86	30.64	0.83	12.68
5	3	Listening	103,807	8-9	1499.79	35.31	0.64	21.25
5	4	Research	103,807	8	1494.25	32.50	0.78	15.30
6	1	Reading	106,229	15-16	1591.23	29.41	0.81	12.80
6	2	Writing	106,229	13	1589.37	29.49	0.79	13.37
6	3	Listening	106,229	8-9	1595.84	33.95	0.62	20.88
6	4	Research	106,229	8	1591.67	33.37	0.72	17.53
7	1	Reading	108,653	15-16	1695.20	27.71	0.80	12.39
7	2	Writing	108,653	13	1688.65	31.25	0.79	14.27
7	3	Listening	108,653	8-9	1697.10	33.72	0.63	20.41
7	4	Research	108,653	8	1692.52	32.83	0.70	17.97

Table 12-11. Reliability, Mean, Standard Deviation, and Conditional Standard Error of Measurement (CSEM) of Mathematics Claims

Grade	Claim No.	Claim	Student *N* Count	Number of Items	Mean	Std. Dev.	Reliability	CSEM
3	1	Concepts and Procedures	100,261	20	1296.94	29.23	0.93	7.49
3	3	Communicating Reasoning	100,261	8	1295.33	32.62	0.71	17.44
3	2 & 4	Problem Solving and Modeling and Data Analysis	100,261	8	1292.44	32.97	0.65	19.51
4	1	Concepts and Procedures	101,929	20	1393.52	27.04	0.94	6.88
4	3	Communicating Reasoning	101,929	8	1391.64	31.75	0.69	17.59
4	2 & 4	Problem Solving and Modeling and Data Analysis	101,929	8	1390.87	31.68	0.69	17.50
5	1	Concepts and Procedures	104,553	20	1487.51	28.24	0.91	8.41
5	3	Communicating Reasoning	104,553	8	1485.02	32.12	0.71	17.40
5	2 & 4	Problem Solving and Modeling and Data Analysis	104,553	8	1481.68	34.84	0.53	23.80
6	1	Concepts and Procedures	108,449	20	1588.33	27.10	0.93	7.25
6	3	Communicating Reasoning	108,449	8	1584.44	31.00	0.61	19.48
6	2 & 4	Problem Solving and Modeling and Data Analysis	108,449	8	1580.49	35.72	0.47	25.99
7	1	Concepts and Procedures	108,702	20	1688.43	27.97	0.91	8.19
7	3	Communicating Reasoning	108,702	8	1683.75	32.30	0.51	22.57
7	2 & 4	Problem Solving and Modeling and Data Analysis	108,702	8	1679.68	36.48	0.46	26.92

12.1.7 Reliability, SEM, and CSEM for Social Studies

Table 12-12 provides information on reliability (coefficient alpha) (see Equation 12.1) and SEM (see Equation 12.2) from the classical true score theory for social studies by grade and form, despite the fact that all OP items are the same for forms 1 through 3 per grade. This choice was made for two reasons: (1) conceptually, it makes more sense to report test-level results by form because each form represents one test, and (2) it shows variations of statistics across forms (if there are any) to inform related decisions (i.e., whether to combine online forms per grade for social studies) when computing classification accuracy and consistency.

As shown in Table 12-12, the values of coefficient alpha across forms and grades for social studies range from 0.83 to 0.89. Therefore, in general, based on coefficient alpha, M-STEP social studies tests have an acceptable degree of internal consistency. Moreover, very similar statistics across the three online forms per grade for social studies are observed. This supports the later decision to combine all three online forms per grade when examining classification accuracy and consistency for social studies.

Table 12-12. Test-Level Descriptive Statistics by Form: Social Studies Reliability and Standard Error of Measurement

Grade	*N* OP Items	Form	*N*	Reliability	SEM
5	45	1	34,846	0.86	3.07
5	45	2	34,862	0.85	3.07
5	45	3	34,848	0.85	3.07
5	45	4	767	0.83	3.08
8	44	1	35,945	0.88	3.01
8	44	2	35,897	0.88	3.01
8	44	3	35,830	0.88	3.02
8	44	4	394	0.83	3.06
11	38	1	34,266	0.89	2.70
11	38	2	34,220	0.89	2.71
11	38	3	34,240	0.89	2.71
11	38	4	621	0.89	2.76

Additionally, the CSEM was calculated for social studies. Related numerical information can be found in corresponding conversion tables reported in Chapter 8 (i.e., Table 8-11). Graphical representations can be found in Figures 12-11 to 12-13. According to these graphs, the CSEMs are not the lowest at the proficient cut scores (i.e., the vertical line, which indicates the cut between Level 2 and Level 3). However, the ability ranges from -2 to 2 in all graphs appear to have low SE. Note, however, that these graphs are made using the post-administration estimated item parameters.

Figure 12-11. Test (Conditional) Standard Error for Social Studies Grade 5 by Form

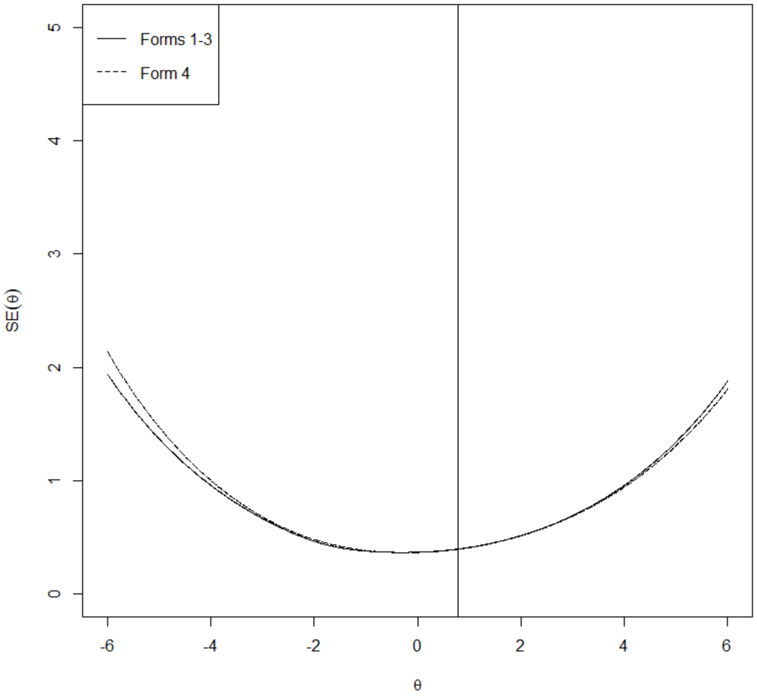


Figure 12-12. Test (Conditional) Standard Error for Social Studies Grade 8 by Form

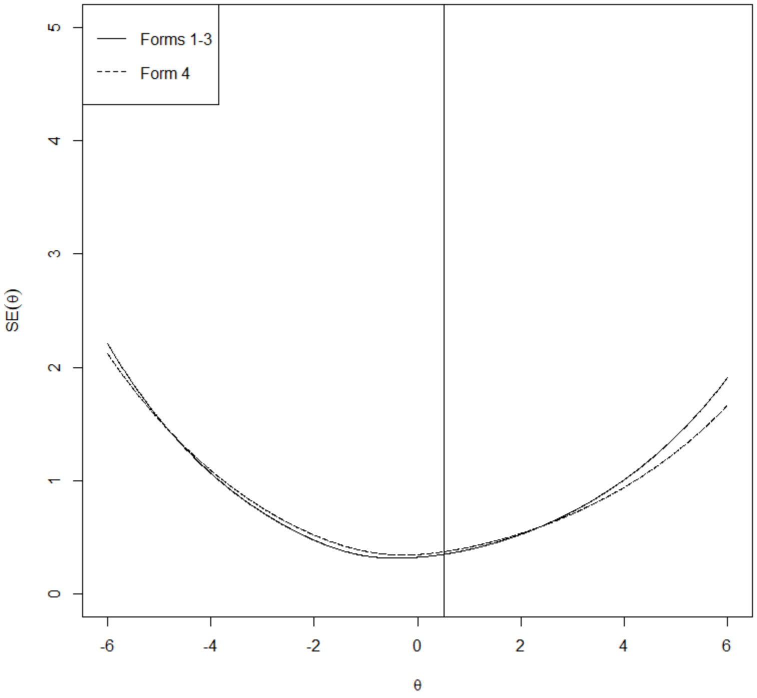
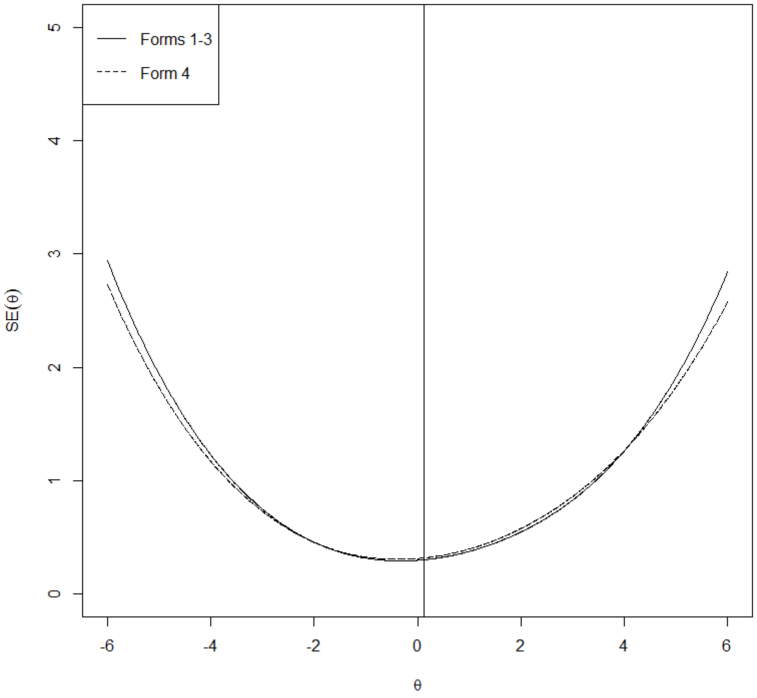


Figure 12-13. Test (Conditional) Standard Error for Social Studies Grade 11 by Form



12.2 Classification Accuracy and Consistency

Based on M-STEP scale scores, student performance in corresponding content areas is classified into one of the four performance levels (i.e., Advanced, Proficient, Partially Proficient, and Not Proficient). Among these, the most important classification is between the Proficient and Partially Proficient categories (i.e., the proficient or not cut). While it is always important to know the reliability of student scores in any examination, it is also important to assess the quality of the decisions, especially regarding the proficient or "not cut" categories. Such evaluation was performed through estimation of the probabilities of accurate and consistent classification of student performance.

Classification accuracy is defined as the extent to which the actual classifications of examinees agree with classifications that would be made on the basis of their true scores (Livingston & Lewis, 1995). It is common to estimate classification accuracy by utilizing a psychometric model to find true scores corresponding to observed scores. The magnitude of classification accuracy measures is influenced by key features of the test design, including the number of items, the number of cut scores, reliability, and associated SEM or CSEM.

12.2.1 ELA and Mathematics

To calculate classification accuracy for each student in ELA and mathematics, the calculations used by Smarter Balanced (see the *Smarter Balanced 2017-18 Technical Report*, 2018). For each student, the likelihood of scoring in each performance level is calculated. The student likelihoods are used to calculate the accuracy by level and the overall accuracy.

Tables 12-13 through 12-15 provide the classification accuracy for ELA and mathematics. The overall classification accuracy ranges from 0.82 to 0.86, and the accuracy by performance level ranges from 0.70 to 0.92. These results suggest that accurate performance level classifications for ELA and mathematics are being made for students in Michigan based on M-STEP. Note that any inconsistencies between the expected values and accuracy by level or overall accuracy are due to computation rounding error.

Table 12-13. Overall Classification Accuracy: ELA and Mathematics

Content Area	Grade	*N*	Overall Accuracy
ELA	3	100,974	0.84
ELA	4	102,564	0.83
ELA	5	105,294	0.83
ELA	6	109,138	0.83
ELA	7	109,181	0.82
Mathematics	3	101,200	0.85
Mathematics	4	102,838	0.85
Mathematics	5	105,486	0.86
Mathematics	6	109,297	0.85
Mathematics	7	109,278	0.85

Table 12-14. Classification Accuracy: ELA

Grade	Assigned Level	*N*	Observed Proportion	Expected Level 1	Expected Level 2	Expected Level 3	Expected Level 4	Accuracy by Level
3	1	30,688	0.30	0.28	0.02	0.00	0.00	0.92
3	2	24,728	0.24	0.03	0.19	0.03	0.00	0.76
3	3	22,646	0.22	0.00	0.03	0.17	0.03	0.76
3	4	22,912	0.23	0.00	0.00	0.03	0.20	0.88
4	1	34,255	0.33	0.31	0.02	0.00	0.00	0.93
4	2	21,284	0.21	0.03	0.15	0.03	0.00	0.70
4	3	22,096	0.22	0.00	0.03	0.16	0.03	0.74
4	4	24,929	0.24	0.00	0.00	0.03	0.21	0.88
5	1	34,002	0.32	0.30	0.02	0.00	0.00	0.92
5	2	22,635	0.21	0.03	0.16	0.03	0.00	0.73
5	3	30,009	0.29	0.00	0.03	0.23	0.03	0.80
5	4	18,648	0.18	0.00	0.00	0.03	0.15	0.84
6	1	34,539	0.32	0.29	0.03	0.00	0.00	0.91
6	2	29,021	0.27	0.03	0.20	0.03	0.00	0.76
6	3	30,803	0.28	0.00	0.03	0.23	0.02	0.81
6	4	14,775	0.14	0.00	0.00	0.02	0.11	0.84
7	1	32,394	0.30	0.27	0.03	0.00	0.00	0.90
7	2	30,083	0.28	0.04	0.20	0.04	0.00	0.74
7	3	33,040	0.30	0.00	0.03	0.25	0.02	0.81
7	4	13,664	0.13	0.00	0.00	0.02	0.10	0.83

Table 12-15. Classification Accuracy: Mathematics

Grade	Assigned Level	N	Observed Proportion	Expected Level 1	Expected Level 2	Expected Level 3	Expected Level 4	Accuracy by Level
3	1	27,823	0.27	0.25	0.02	0.00	0.00	0.92
3	2	26,063	0.26	0.03	0.20	0.03	0.00	0.77
3	3	27,568	0.27	0.00	0.03	0.22	0.02	0.82
3	4	19,746	0.20	0.00	0.00	0.02	0.17	0.89
4	1	25,336	0.25	0.22	0.02	0.00	0.00	0.91
4	2	34,477	0.34	0.03	0.28	0.03	0.00	0.82
4	3	25,939	0.25	0.00	0.03	0.21	0.02	0.82
4	4	17,086	0.17	0.00	0.00	0.02	0.15	0.88
5	1	38,435	0.36	0.34	0.03	0.00	0.00	0.92
5	2	30,309	0.29	0.03	0.23	0.03	0.00	0.81
5	3	18,963	0.18	0.00	0.02	0.14	0.02	0.79
5	4	17,779	0.17	0.00	0.00	0.02	0.15	0.89
6	1	37,405	0.34	0.32	0.03	0.00	0.00	0.92
6	2	33,476	0.31	0.03	0.24	0.03	0.00	0.79
6	3	20,722	0.19	0.00	0.03	0.14	0.02	0.76
6	4	17,694	0.16	0.00	0.00	0.02	0.14	0.89
7	1	39,218	0.36	0.33	0.03	0.00	0.00	0.91
7	2	30,963	0.28	0.04	0.22	0.03	0.00	0.78
7	3	21,135	0.19	0.00	0.02	0.16	0.02	0.81
7	4	17,962	0.16	0.00	0.00	0.02	0.15	0.89

12.2.2 Social Studies

For social studies, each test under consideration consists of equally weighted and dichotomously scored items only, and procedures from Hanson and Brennan (1990) were applied to derive classification accuracy and classification consistency measures. Moreover, the definitions for accuracy and consistency of decisions presented in Young and Yoon (1998) were adopted. Specifically, the *accuracy* of decisions is the extent to which decisions would agree with those made if each student could somehow be tested with all possible forms of an examination; and the *consistency* of decisions is the extent to which decisions would agree with those made if each student had taken a parallel form of the examination, equal in difficulty and covering the same content as the form the student actually took (Young & Yoon, 1998). These ideas are shown schematically in Figures 12-14 and 12-15 using M-STEP social studies as an example. In both figures, “Achieves Proficient Status” refers to the proficient and above category on the total raw score, and “Does Not Achieve Proficient Status” refers to all categories below the proficient cut.

Figure 12-14. Classification Accuracy

		Decision made on a form actually taken	**Decision made on a form actually taken**
		Does Not Achieve Proficient Status	Achieves Proficient Status
“True status” based on all-forms average	Does Not Achieve Proficient Status	Correct Classification	Misclassification
	Achieves Proficient Status	Misclassification	Correct Classification

Note. Adapted from Young and Yoon (1998).

Figure 12-15. Classification Consistency

		Decision made on the 2nd form taken	**Decision made on the 2nd form taken**
		Does Not Achieve Proficient Status	Achieves Proficient Status
Decision made on the 1st form taken	Does Not Achieve Proficient Status	Consistent Classification	Inconsistent Classification
	Achieves Proficient Status	Inconsistent Classification	Consistent Classification

Note. Adapted from Young and Yoon (1998).

In Figure 12-14, accurate classification occurs when the decision made on the basis of the form actually taken agrees with the decision made on the basis of the theoretical “all-forms” average. Misclassification occurs, for example, when a student who “Does Not Achieve Proficient Status” based on his or her “all-forms” average is classified incorrectly as “Achieves Proficient Status.”

Consistent classification occurs (see Figure 12-15) when two possible alternate forms agree on the classification of a student as either “Achieves Proficient Status” or “Does Not Achieve Proficient Status,” whereas inconsistent classification occurs when the decisions made by the forms differ.

The analyses made use of the techniques outlined and implemented by Hanson and Brennan (1990) and Brennan (2004). Specifically, a four-parameter beta distribution was used to model the true score, and Lord’s (1965) two-term approximation to the compound binomial distribution was used to model the conditional error. The BB-CLASS software (Version 1.1) was used to complete these analyses (Brennan, 2004).

Table 12-16 presents the analysis results of decision accuracy and consistency for classifying students at each grade level per test form as “Achieves Proficient Status” or “Does Not Achieve Proficient Status” based on their M-STEP social studies total raw scores. As mentioned above, the three online forms for social studies were combined (see Table 12-16) due to the fact that all OP items are exactly the same across these forms and the raw score statistics are very similar across forms (see Table 8-6).

In addition to classification accuracy and consistency, Table 12-16 provides information on the proportion of false positives and false negatives (i.e., the two types of misclassification). The false positive is the type of misclassification in which students should be classified in the “Does Not Achieve Proficient Status” category based on their “all-forms” average but end up in the “Achieves Proficient Status” category based on the actual form. The false negative is just the opposite: students who should be in the “Achieves Proficient Status” category based on their “all-forms” average end up in the “Does Not Achieve Proficient Status” category based on the actual form. The sum of the proportion values for accuracy, false positives, and false negatives should be equal to 1.00. Due to rounding, however, the sum of these values in the table may not be equal to 1.00.

As shown in Table 12-16, the proportion of false positives (i.e., labeling a student as proficient when he or she should be categorized as not proficient) ranged from 0.03 to 0.06 for social studies. Moreover, the proportion of false negatives (i.e., labeling a student as not proficient when he or she should be categorized as proficient) ranged from 0.01 to 0.03 for social studies.

The last column in Table 12-16 reports the proportion of students predicted by the model that would be assigned to the same category (i.e., either proficient or not proficient) if an alternate form of M-STEP social studies assessments (with similar content coverage and item difficulty as the actual form) had been administered. These values range from 0.87 to 0.94.

Table 12-16. Decision Accuracy and Consistency on M-STEP Social Studies Total Raw Score by Grade and Form

Grade	Form	Accuracy	False Positive	False Negative	Consistency
5	1–3	0.93	0.05	0.02	0.91
5	4	0.96	0.03	0.01	0.94
8	1–3	0.92	0.05	0.03	0.89
8	4	0.95	0.03	0.01	0.93
11	1–3	0.91	0.06	0.03	0.87
11	4	0.92	0.05	0.03	0.88

12.3 Assumption of Unidimensionality

Another measure of construct validity is unidimensionality. One of the underlying assumptions of the IRT models used to scale M-STEP content area tests is that the items being calibrated are unidimensional; that is, items comprising M-STEP in each grade/content area measure a single construct. For example, mathematics items should measure mathematics ability and not reading skills. Standard 1.13 of the AERA, APA, & NCME (2014) *Standards* states the following:

> If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided. (p. 26–27)

12.3.1 ELA and Mathematics

Smarter Balanced examined the unidimensionality for the Smarter Balanced/M-STEP ELA and mathematics assessments. Based on the findings of the dimensionality study, Smarter Balanced found that the use of the unidimensional item response theory (IRT) model and test design was appropriate. A detailed discussion and the results of the dimensionality study can be found in the online Smarter Balanced 2013–2014 Technical Report (2016).[1]

12.3.2 Social Studies

For M-STEP social studies, MDE conducted two analyses to evaluate the unidimensionality assumption with OP items only. The first set was an exploratory factor analysis (EFA) using the Mplus software with the WLSMV[2] estimator. Barendse, Oort, and Timmerman (2015) found that WLSMV is the preferred estimation method and is recommended to rely on the Root Mean Squared Error of Approximation (RMSEA) index (in which values less than 0.05 are desired) if the primary interest is in major factors. The second set of analyses is a principle component analysis (PCA) using *MATLAB* (2018). For PCA results, the magnitude of the first and second eigenvalues are examined. Both the eigenvalues-greater-than-one rule and the scree plot approach are considered. The RMSEA values for one-factor EFA models and the first two eigenvalues from

[1] https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

[2] "WLSMV-weighted least square parameter estimates using a diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistic that use a full weight matrix" (Muthén and Muthén, 2012, p. 603)

each PCA model are reported in Table 12-17.

As shown in Table 12-17, the dimensionality assessment for social studies is evaluated by administration mode at each grade level.[3] Both the EFA and PCA results failed to reject the unidimensionality assumption, which is a supporting piece of evidence for the use of unidimensional IRT models at each grade for social studies.

Table 12-17. RMSEA from 1-Factor EFA and the First Two Eigenvalues from PCA

Content Area	Grade	Form	RMSEA (1-Factor EFA)	PCA First Eigenvalue	PCA Second Eigenvalue
Social Studies	5	1–3	0.013	1.4830	0.2686
Social Studies	5	4	0.014	1.3197	0.3224
Social Studies	8	1–3	0.016	1.7576	0.3000
Social Studies	8	4	0.013	1.3869	0.3636
Social Studies	11	1–3	0.017	1.8331	0.2913
Social Studies	11	4	0.015	1.8357	0.3343

12.4 Validity Evidence

The *Standards for Educational and Psychological Testing* defines validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests” (AERA, APA, & NCME, 2014, p. 11). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence that either supports or challenges its validity, including design, content specifications, item development, psychometric quality, and inferences made from the results.

The validity of score interpretations for M-STEP is supported by multiple sources of evidence. Chapter 1 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) specifies the following sources of validity evidence that are important to gather and document in order to support validity claims for an assessment:

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of test use

[3] Note that for each grade, forms 1–3 are online forms and form 4 is a paper/pencil form. All OP items are the same across forms 1–3 for social studies at each grade. Form 4, however, has somewhat different OP items from the online forms because technology-enhanced items cannot be put on a paper/pencil form.

It is important to note that these categories are not mutually exclusive. One source of validity evidence often falls into more than one category, as discussed in more detail in this section. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout test development, administration, scoring, reporting, and beyond. As the technical report has progressed, it has covered the different phases of the testing cycle. Each part of the technical report has detailed the procedures and processes applied in Michigan and the corresponding results. Each part has also highlighted the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to specific sections of the *Standards*. The current section now addresses these final issues in validity: test content, response processes, internal test structure, relation to other variables, and consequences of test use.

12.4.1 Minimization of Construct-Irrelevant Variance and Construct Underrepresentation

Minimization of construct-irrelevant variance and construct underrepresentation is addressed in the following steps of the test development process: 1) specification, 2) item writing, 3) review, 4) field-testing, 5) test construction, and 6) item calibration. See Chapter 3 for more information on steps 1 through 5 and Chapter 8 for more information on calibration.

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs measured by the test. For example, when tests are not administered under standardized conditions (e.g., one administration may be timed, but another administration may not be timed), differences in student performance may be partially associated with the different administration conditions. Careful specification of content and review of the items representing that content are the first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data, is used to infer construct irrelevance. For additional details with respect to ELA and mathematics, please see the *Smarter Balanced 2017–2018 Technical Report* (2018).

Construct underrepresentation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. Specification and review, in which test blueprints are developed and reviewed, are primary steps in the development process and are designed to ensure that content is appropriately represented.

12.4.2 Evidence Based on Test Content

According to the *Standards*, evidence based on test content “can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores” (AERA, APA, & NCME, 2014, p. 14). Documentation of the content domains, how the content is sampled and represented, and alignment of items to the content were discussed in Chapter 3. The documentation showed how test specification documents derived from earlier developmental activities guided the final phases of test development and ultimately yielded the test forms that were administered to students.

Chapter 3 also showed that the participation of Michigan educators in that process provided a solid rationale for having confidence in the content and design of Michigan M-STEP as a tool from which to derive valid inferences about Michigan student performance. Particularly for social studies, use of classroom teachers brought into the process the enacted curriculum perspective and the written curriculum perspective. The test development process and the involvement of Michigan educators in that process formed an important part of the validity of the entire Michigan M-STEP assessment.

12.4.3 Evidence Based on Response Processes

According to the *Standards*, evidence based on response processes “generally comes from analyses of individual responses” (AERA, APA, & NCME, 2014, p. 15). Hence, the best opportunity for detecting and eliminating potential sources of invalidity occurs during the test development process (U.S. Department of Education, 2015). As indicated in Chapter 3, all items for M-STEP were carefully reviewed through multiple cycles of the item development process for issues related to ambiguity, bias, sensitivity, irrelevance, and inaccuracy to ensure a fit between the construct and the nature of the actual performance.

12.4.4 Evidence Based on Internal Test Structure

According to the *Standards*, evidence based on internal structure reflects “the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA, APA, & NCME, 2014, p. 16). Three important sources of internal structure evidence have been addressed within this technical document: measurement invariance, dimensionality, and reliability. Evidence of measurement invariance is provided in Chapter 11 by using DIF. Additional support for measurement invariance can be found in Section 12.2.5, which reports the subgroup reliability estimates. The dimensionality investigation mentioned in Section 12.3 also provides supporting evidence of the internal test structure.

12.4.5 Evidence Based on Relations to Other Variables

Convergent validity is a subtype of construct validity that can be estimated by the extent to which measures of constructs that theoretically should be related to each other are, in fact, observed as related to each other. Analyses of the internal structure of a test can indicate the extent to which the relationships among test items conform to the construct the test purports to measure. For example, M-STEP mathematics test is designed to measure a single overall construct—mathematics achievement. Therefore, the items comprising the M-STEP mathematics test should measure only mathematics.

For M-STEP assessments,[4] this technical report summarizes additional statistics that contribute to construct validity, reliability—as reported previously in this chapter and Chapter 8—and item fit. The internal consistency coefficient reported above is a measure of item homogeneity. For a group of items to be homogeneous, they must measure the same construct (construct validity) or represent the same content domain (content validity). Because IRT models were used to

[4] For ELA and mathematics, not all psychometric characteristics are provided in this report. Additional details can be found in the Smarter Balanced Technical Reports (2016 - 2018).

calibrate test items and to report student scores, item fit is also relevant to construct validity. The extent to which test items function as the IRT model prescribes is relevant to the validation of test scores. Additional evidence to support construct validity is examined by the correlations between the claim scores for ELA and mathematics in the next section.

12.4.6 Correlations among Claims as Evidence of Convergent Validity

In this section, the strength of the interrelationships among the claims are reported by computing the correlations between them. Two types of correlations are reported here: the uncorrected Pearson product-moment (PPM) correlation coefficients and the PPM corrected for attenuation (CAPPM).

AERA, APA, & NCME (2014) Standard 1.21 states the following:

> When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates. (p. 29)

We can correct for the attenuation of the PPM statistically using Spearman's formula:

$$CAPPM = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \quad (12.9)$$

where r_{xy} is the PPM between two claims, r_{xx} is the reliability of one of those claims, and r_{yy} is the reliability for the other claim.

Tables 12-18 and 12-19 report the PPM and CAPPM described above. The PPM among the claim scores is presented below the diagonal portion of the matrix, and the CAPPM is presented above the diagonal portion of the matrix in each table.

The uncorrected PPM in Tables 12-18 and 12-19 should be interpreted in the context of the reliability coefficient. In general, it is expected to see lower PPM coefficients between variables that are less reliable. In most cases, the PPM coefficients show that performance on one claim is moderately related to performance on another claim within the same grade and content area. In cases where there is a limited number of items per claim, caution should be used when comparing the PPM coefficients measuring the relationships between claims to those measuring the relationships between content areas.). We expect to see a more modest relationship (smaller correlation coefficients) reported between the claims as a consequence of the lower number of items measuring each of the reporting categories. The PPM between two claim scores may be artificially low because of measurement error.

Across all tables, the CAPPM indicates strong relationships between the claims. In some cases, the CAPPM is greater than 1.00. "Disattenuated values greater than 1.00 indicate that measurement error is not randomly distributed" (Schumacker, 1996). The strong relationships suggested by the CAPPM in Tables 12-18 and 12-19 are further evidence of the validity of the test construct. Since the overall content area is composed of the claim scores, and the content area is expected to measure a single dimension, it is expected that these claim scores are also highly related.

Table 12-18. Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Claims: English Language Arts

Grade	Claim No.	Claim	Number of Items	1	2	3	4
3	1	Reading	15-16		0.94	0.96	0.91
3	2	Writing	13	0.79		0.93	0.91
3	3	Listening	8-9	0.70	0.66		0.92
3	4	Research	8	0.73	0.72	0.63	
4	1	Reading	15-16		0.93	0.94	0.91
4	2	Writing	13	0.77		0.90	0.89
4	3	Listening	8-9	0.67	0.65		0.90
4	4	Research	8	0.72	0.70	0.61	
5	1	Reading	15-16		0.93	0.94	0.93
5	2	Writing	13	0.77		0.91	0.92
5	3	Listening	8-9	0.68	0.66		0.91
5	4	Research	8	0.74	0.74	0.64	
6	1	Reading	15-16		0.94	0.96	0.94
6	2	Writing	13	0.75		0.93	0.92
6	3	Listening	8-9	0.68	0.65		0.94
6	4	Research	8	0.72	0.70	0.63	
7	1	Reading	15-16		0.94	0.96	0.95
7	2	Writing	13	0.75		0.93	0.93
7	3	Listening	8-9	0.69	0.66		0.95
7	4	Research	8	0.71	0.69	0.63	

Table 12-19. Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Claims: Mathematics

Grade	Claim No.	Claim	Number of Items	1	3	2 & 4
3	1	Concepts and Procedures	20		0.90	0.94
3	3	Communicating Reasoning	8	0.73		0.94
3	2 & 4	Problem Solving and Modeling and Data Analysis	8	0.73	0.64	
4	1	Concepts and Procedures	20		0.92	0.92
4	3	Communicating Reasoning	8	0.74		0.94
4	2 & 4	Problem Solving and Modeling and Data Analysis	8	0.74	0.65	
5	1	Concepts and Procedures	20		0.94	1.01
5	3	Communicating Reasoning	8	0.75		1.02
5	2 & 4	Problem Solving and Modeling and Data Analysis	8	0.70	0.63	
6	1	Concepts and Procedures	20		0.94	1.11
6	3	Communicating Reasoning	8	0.70		1.10
6	2 & 4	Problem Solving and Modeling and Data Analysis	8	0.73	0.59	
7	1	Concepts and Procedures	20		0.94	1.11
7	3	Communicating Reasoning	8	0.70		1.10
7	2 & 4	Problem Solving and Modeling and Data Analysis	8	0.73	0.59	

12.4.7 Divergent (Discriminant) Validity

Measures of different constructs should not be highly correlated with each other. Divergent validity is a subtype of construct validity that can be assessed by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures of unrelated or distantly related constructs are examined in support of divergent validity.

To assess the divergent validity of M-STEP, correlations were computed between the ELA and mathematics scale scores for students who took both assessments. These correlation results are shown in Table 12-20. The correlation coefficients ranged from 0.77 (between ELA and mathematics in grade 5) to 0.79 (between ELA and mathematics in grades 6 and 7). The correlation coefficients suggest that individual student scores for ELA and mathematics are highly related. Despite high correlations, the tests are not perfectly related to each other, suggesting that different constructs are being tapped; however, the test scores do appear as highly related to one another, suggesting they may be tapping into a similar knowledge base or general underlying ability.

Table 12-20. Inter-correlation of ELA and Mathematics Scale Scores

Grade	Inter-Correlation
3	0.78
4	0.78
5	0.77
6	0.79
7	0.79

12.4.8 Evaluation of Item Exposure for CAT ELA and Mathematics

Controlling item exposure is of concern with CAT administrations, which impacts the validity of the interpretation of the text scores. Overexposed items could be a threat to validity because students may become familiar with the items over time and, thus, decrease the difficulty of the item, which would impact the ability estimate (Georgiadou, Triantafillou, & Economides, 2007). Item exposure rates were obtained using all completed, online, adaptive tests for which item data were available. The exposure rate for a given item is the proportion of tests (in the grade and content area) on which the item appeared.

Table 12-21 presents a summary of the item exposure results for ELA and mathematics. Within each grade, the table presents the number of items in the OP pool (N) and various descriptive statistics, including the mean, standard deviation (SD), range (Min, Max), and median of the observed exposure rates. Table 12-21 shows that, on average, the same item appeared in 5% of the grade 3 tests; in other words, 5% of grade 3 examinees saw the same item. As a rule of thumb, Smarter Balanced attempts to maintain a maximum exposure rate of 25% (meaning that 25% of examinees will see the same item). Table 12-21 shows that the mean and median exposure rates for ELA and mathematics CAT items are well below 25%.

Table 12-21. Summary of ELA Item Exposure Rates by Grade and Component

Content Area	Grade	*N*	Mean	SD	Min	Max	Median
ELA	3	860	0.05	0.09	0.00	0.47	0.01
ELA	4	820	0.05	0.09	0.00	0.51	0.01
ELA	5	775	0.06	0.09	0.00	0.49	0.02
ELA	6	746	0.06	0.10	0.00	0.68	0.02
ELA	7	663	0.07	0.11	0.00	0.61	0.02
Mathematics	3	1,211	0.03	0.04	0.00	0.21	0.01
Mathematics	4	1,269	0.03	0.04	0.00	0.22	0.01
Mathematics	5	1,193	0.03	0.04	0.00	0.20	0.01
Mathematics	6	1,094	0.03	0.05	0.00	0.23	0.01
Mathematics	7	967	0.04	0.06	0.00	0.27	0.01

Table 12-22 provides further information about the exposure rates by showing the number of items in the OP pool (N) and proportion of items with exposure rates falling into certain ranges (bins with a width of 20%), including those that were completely unexposed (Unused). The majority of CAT items, for both ELA and mathematics, had item exposure rates between 0% and 20%.

There were a handful of items in ELA with higher-than-desirable exposure rates. This occurred when there were few items measuring elements in the blueprint. There were also items in both content areas that were unused. There is a trade-off between blueprint fidelity and exposure, with the adaptive CAT engine weighting blueprint fidelity more heavily. In addition, for ELA, it was requested to use all or almost all items with a passage so students were not given numerous passages to read to meet the blueprint.

Table 12-22. Percentage of CAT Items by Exposure Rate

Content Area	Grade	Total Number of Items	Unused*	0%–20%	21%–40%	41%–60%	61%–80%	81%–100%
ELA	3	860	10.12	92.67	6.16	1.16	0.00	0.00
ELA	4	820	6.10	90.85	8.17	0.98	0.00	0.00
ELA	5	775	8.65	92.00	7.23	0.77	0.00	0.00
ELA	6	746	7.64	89.01	9.38	1.34	0.27	0.00
ELA	7	663	9.80	88.08	9.80	1.96	0.15	0.00
Mathematics	3	1,211	11.15	99.92	0.08	0.00	0.00	0.00
Mathematics	4	1,269	7.09	99.68	0.32	0.00	0.00	0.00
Mathematics	5	1,193	7.12	100.00	0.00	0.00	0.00	0.00
Mathematics	6	1,094	8.23	99.54	0.46	0.00	0.00	0.00
Mathematics	7	967	3.21	97.72	2.28	0.00	0.00	0.00

*Note: “Unused” is also included in the 0% to 20% range.

12.4.9 Evidence Based on Consequences of Test Use

The *Standards* incorporate the intended and unintended consequences of test use into the concept of validity. It indicates that information about the consequences of testing does not in and of itself detract from the validity of intended test interpretations (AERA, APA, & NCME, 2014, p. 19). Rather, according to the *Standards*, a more searching inquiry into the sources of those consequences given the intended purposes of an assessment is a basis for evaluating the quality of the validity evidence. The test data alone do not provide sufficient verification of this type of evidence. For this reason, it is not straightforward to measure and collect evidence on the consequential aspects of validity.

To address the intended consequences of M-STEP, the purposes of M-STEP must be specified. MDE has carefully articulated the intended purposes of M STEP as driving features of the selection of Smarter Balanced items, the development of social studies tests, and the implementation of the testing program. The specific purposes associated with M-STEP include the following:

- M-STEP accurately describes both student achievement (i.e., how much students know at the end of the year) and student growth (i.e., how much students have improved since the previous year) to inform program evaluation and school-, district-, and state-accountability systems and to provide valid, reliable, and fair measures of students’ progress toward, and attainment of, the knowledge and skills required to be college- and career-ready.
- M-STEP informs state and federal accountability.
- M-STEP assessments are fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement.

12.5 Summary

In summary, Chapter 12 of this report demonstrates M-STEP's adherence to the AERA, APA, & NCME (2014) *Standards* regarding reliability and construct-related validity. The analyses described above address multiple best practices of the testing industry, particularly the following standards:

- Standard 2.0—Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.
- Standard 2.1—The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation.
- Standard 2.3—For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.
- Standard 2.13—The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score.
- Standard 2.14—When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.
- Standard 2.16—When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.
- Standard 2.19—Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.
- Standard 4.3—Test developers should document the rationale and supporting evidence for the administration, scoring, and reporting rules used in computer-adaptive, multistage-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or sets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and in controlling item exposure.

References

American Institutes for Research (2016). Smarter Balanced scoring specification: Summative and Interim Assessments: ELA/Literacy Grades 3-8;11 and Mathematics Grade 3-8;11. Los Angeles, CA: Smarter Balanced Assessment Consortium.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling,* 22(1), 87–101.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores (pp. 374–472). Reading, MA: Addison-Wesley Publishing.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443–459.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (Version 1.1)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Cai, L. (2017). flexMIRT (Version 3.51) [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245–276.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 221–256). Westport, CT: American Council on Education and Praeger.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.

Candell, G. L. & Drasgow, F. (1988). An iterative procedure for linking metrics bias in item response theory. *Applied Psychological Measurement*, *12(3)*, 253–260.

Cattell, R. B. (1966). *The scree test for the number of factors*. Multivariate Behavioral Research, 1, 245–276.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the “problem” of sample size: A clarification. Psychological Bulletin, 109(3), 512–519.

Darling-Hammond, L., & Pecheone, R. (2010). Developing an Internationally Comparable Balanced Assessment System that Supports High-Quality Learning. Retrieved from http://www.k12center.org/publications.html.

Data Recognition Corporation (2017). *Technology User Guide*. Maple Grove, MN: Author.

Dorans, N. J., & Schmitt, M. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton: Educational Testing Service.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

ETS. (2012). Smarter Balanced Assessment Consortium: Bias and sensitivity guidelines. Princeton, NJ: ETS.

Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8).

Green, D. R. (1975). *Procedures for assessing bias in achievement tests*. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.

Hansen, E.G. & Mislevy, R.J. (2008). *Design Patterns for Improving Accessibility for Test Takers With Disabilities*. Princeton, NJ, ETS Research Report No. RR-08-49.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27(4)*, 345–359.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2(5)*. Available from http://www.jtla.org.

Lewis D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), IRT-based standard-setting procedures utilizing behavioral anchoring. Symposium conducted at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Mercado, R. L, & Schultz, E. M. (2012). *The bookmark standard setting procedure*. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations. New York, NY: Routledge.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika*, *30*, 239–270.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mantel, N. (1963) Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association*, *58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

MATLAB and Statistics Toolbox Release 2018b, The MathWorks, Inc., Natick, Massachusetts, United States.

McDonald, R. P. (1999). T*est theory: A unified treatment*. Hillsdale, NJ: Erlbaum.

Michigan Department of Education (2019). *2018–2019 Guide to State Assessments*. Retrieved from https://www.michigan.gov/documents/mde/Guide_to_State_Assessments_622260_7.pdf

Michigan Department of Education (2018). *Assessment Integrity Guide*. Retrieved from https://www.michigan.gov/documents/mde/Assessment_Integrity_Guide_291950_7.pdf

Michigan Department of Education (2019). *Spring 2019 Interpretive Guide to M-STEP Reports*. Retrieved from https://www.michigan.gov/documents/mde/2019_Interpretive_Guide_to_M-STEP_Reports_661956_7.pdf

Michigan Department of Education (2019). *Spring 2019 Interpretive Guide to MME (Michigan Merit Exam) Reports*. Retrieved from https://www.michigan.gov/documents/mde/Interpretive_Guide_to_MME_Reports_671111_7.pdf

Michigan Department of Education (2019). *Spring 2019 Michigan Grade 8 Testing Interpretive Guide to Reports*. Retrieved from https://www.michigan.gov/documents/mde/2019_Michigan_Grade_8_Testing_Interpretive_Guide_to_Reports_665028_7.pdf

Michigan Department of Education (2016). *Supports and Accommodations Guidance Document for M-STEP, MI-Access, WIDA, PSAT, SAT, and ACT WorkKeys*. Retrieved from https://www.michigan.gov/documents/mde/Michigan_Accommodations_Manual.final_480016_7.pdf

Michigan Department of Education and Michigan Virtual University (2019). *MDE Assessment Security*. Retrieved from http://bit.ly/MDEAssessmentSecurity

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25(4)*, 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: *Interdisciplinary Research and Perspectives*, *1(1)*, 3–62.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software].

Muthén, B. O., & Muthén, L. K. (2012). Mplus user's guide: Statistical analysis with latent variables. Los Angeles, CA: Muthén & Muthén.

Schumacker, R. E. (1996). *Disattenuating correlation coefficients*. Rasch Measurement Transactions, 10, 479.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, *29*, 150–151.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education*, *20*, 335–355.

Smarter Balanced Assessment Consortium. (2014a). *Accessibility and accommodations framework*. *Retrieved from Smarter Balanced Accessibility and Accommodations Framework*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium (2014b). *Usability, accessibility, and accommodations guidelines*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium (2014f). *Interpretation and use of scores and achievement levels*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2014g). *Reporting system user guide*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2015a). *Content specifications for the summative assessment of the common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2015b). *Content specifications for the summative assessment of the common core state standards for mathematics*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2015c). *Item and task specifications*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2015d). *The Smarter Balanced Assessment Consortium: Achievement level setting final report*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2016). *2014–2015 technical report*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2017). *2016–2017 technical report*. Los Angeles, CA: Author.

Smarter Balanced Assessment Consortium. (2018). *2017–2018 technical report*. Los Angeles, CA: Author.

U.S. Department of Education. (2007). Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind act of 2001. Retrieved from US Department of Education Policy and Guidance

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.

Yen, W. M., & Fitzpatrick, A. R. (2006). *Item response theory*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 111–153). Westport, CT: Praeger

Young, M. J., & Yoon, B. (1998, April). Estimating the consistency and accuracy of classifications in a standards-referenced assessment. (CSE Technical Report 475). Center for the Study of Evaluation, Standards, and Student Testing. Los Angeles, CA: University of California, Los Angeles.

Zhang, T., Haertel, G., Javitz, H., Mislevy, R., Murray, E., & Wasson, J. (2009). *A design pattern for a spelling bee assessment for students with disabilities*. A paper presented at the annual conference of the American Psychological Association, Montreal, Canada.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). *Assessment of differential item functioning for performance tasks*. Journal of Educational Measurement, 30, 233–251.