H.B. 25-1278 Study: Assessment Shortening and Adaptive Technology

Submitted to:

Office of the Governor

Colorado State Board of Education

House and Senate Education Committees of the Colorado General Assembly

By:
Christina Wirth-Hawkins, Ph.D., Chief Assessment Officer; Jasmine Carey, Ph.D., Director of Achievement Data and Psychometrics; Sara Loerzel, M.Ed., Director of Colorado-Developed Assessments

Assessment and Accountability Analytics Division
Colorado Department of Education
201 E. Colfax Ave., Denver, CO 80203
wirth-hawkins_c@cde.state.co.us

November 2026

Table of Contents

Executive Summary 3

Introduction and Background 4

Figure 1. CMAS Maximum Testing Time by Grade Over Years (Hours) 4

Figure 2. Maximum Statewide Testing Time Across States (Hours) 5

Research Questions and Methods 5

Phase 1: Shorter Fixed-Form Assessments 5

Phase 2: Computer Adaptive Testing 6

Phase 3: Operational and Administrative Changes 6

Findings by Study Phase 6

Phase 1: Shorter Fixed-Form Assessments 6

Phase 2: Computer Adaptive Testing 7

Phase 3: Operational and Administrative Changes 8

Technical Review, Stakeholder Feedback, and Basis for Assessment Adjustments 8

Table 1: Stakeholder Groups Consulted 9

Spring 2027 Actions and Implementation 10

Table 2: Assessment Reductions for 2027 10

Table 3: Overall Reduction for Students 11

Table 4: Sample Spring 2027 CMAS Schedule One Unit Per Day Per Student 11

Table 5: Sample Spring 2027 CMAS Schedule Up to Two Units Per Day Per Student 12

Conclusion 12

Appendices 14

Appendix A: Technical Study: Colorado Measures of Academic Success (CMAS) Testing Time Reduction Study 15

Appendix B: Technical Advisory Panel (TAP) Feedback Summary 77

Appendix C: Accountability Work Group (AWG) Feedback Summary 80

Executive Summary

House Bill 25-1278 directed the Colorado Department of Education (CDE) to examine ways to shorten statewide assessments, including computer adaptive testing. CDE completed the study with Pearson, evaluated findings with technical experts and Colorado stakeholders, and established CMAS reductions for spring 2027. The changes are based on technical evidence and reflect the stakeholder priority of reducing disruption to instruction while preserving assessment quality and the usefulness of results.

Study Priorities and Approach

The study examined three approaches: shorter fixed-form assessments (Phase 1), adaptive testing (Phase 2), and operational changes (Phase 3). Across all phases, considerations included technical quality, standards coverage, the reliability and usefulness of student achievement and growth results, cost, and administration. Analyses included simulations using spring 2025 mathematics and English language arts (ELA) data from grades 4 and 6, student completion times, allowances in other states, and school scheduling demands. The priority was reducing disruption to instruction while preserving assessment quality and the usefulness of results. Appendices A–C provide detailed technical analyses and stakeholder feedback summaries.

Findings and Stakeholder Direction

Removing items reduced reliability and precision, particularly for detailed skill results in both ELA and mathematics. ELA reading passages and writing tasks limit time savings because removing questions does little to shorten reading and writing time. Adaptive designs did not demonstrate substantial time savings while preserving required content coverage and reporting. Operational analyses identified opportunities to reduce sessions and scheduled time without removing assessed content.

The Technical Advisory Committee (TAC) reviewed study methods and findings. The Technical Advisory Panel (TAP) and Accountability Work Group (AWG) considered technical findings, administration demands, and reporting needs. Stakeholders supported responsible reductions but prioritized retention of valid and reliable scores, constructed response items, student growth scores, trends, and detailed reporting over modest time savings. They identified fewer sessions as particularly valuable for reducing scheduling, proctoring, device, and make-up demands.

Spring 2027 Actions and Follow-Through

Mathematics and elementary and middle school science will each move from three 65-minute units to two 75-minute units, eliminating one session and reducing maximum available time by 45 minutes, or 23 percent, per subject. Students taking both subjects will have two fewer sessions and 1.5 fewer hours of maximum available testing time. These changes will reduce testing burden and will be achieved without removing test questions, reducing the content assessed, or changing the information reported. ELA will retain its design and schedule because studied reductions offered limited savings with meaningful losses in measurement quality.

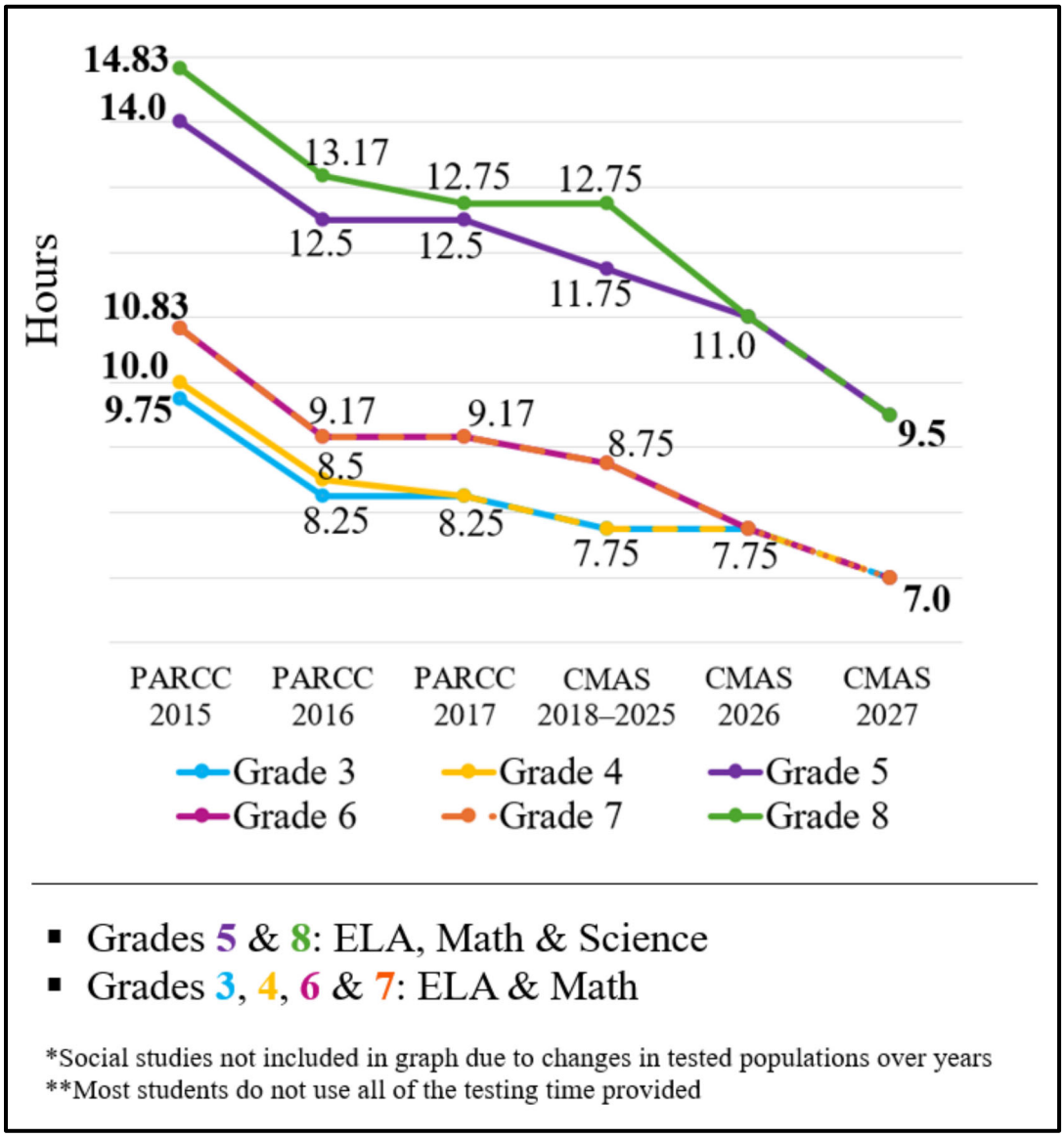
CDE will continue examining opportunities to reduce testing time, including timing analyses, shorter reading passages, fewer field-test items, and administrative efficiencies while considering tradeoffs and stakeholder feedback.

Introduction and Background

The 1241 Accountability Task Force recommended investigating shorter statewide assessments and adaptive testing. House Bill 25-1278 directed CDE to study those approaches and report findings and recommendations to state leaders by November 1, 2026. The central policy question is how to reduce testing burden while retaining an assessment that covers the breadth and depth of the Colorado Academic Standards, meets state and federal assessment and accountability requirements, and supports the results used by families, educators, the community, and the school and district performance frameworks.

Colorado has reduced maximum state assessment time since 2015. By 2026, maximum testing time in grades 3 through 8 had fallen by approximately 20 to 28 percent, depending on grade.

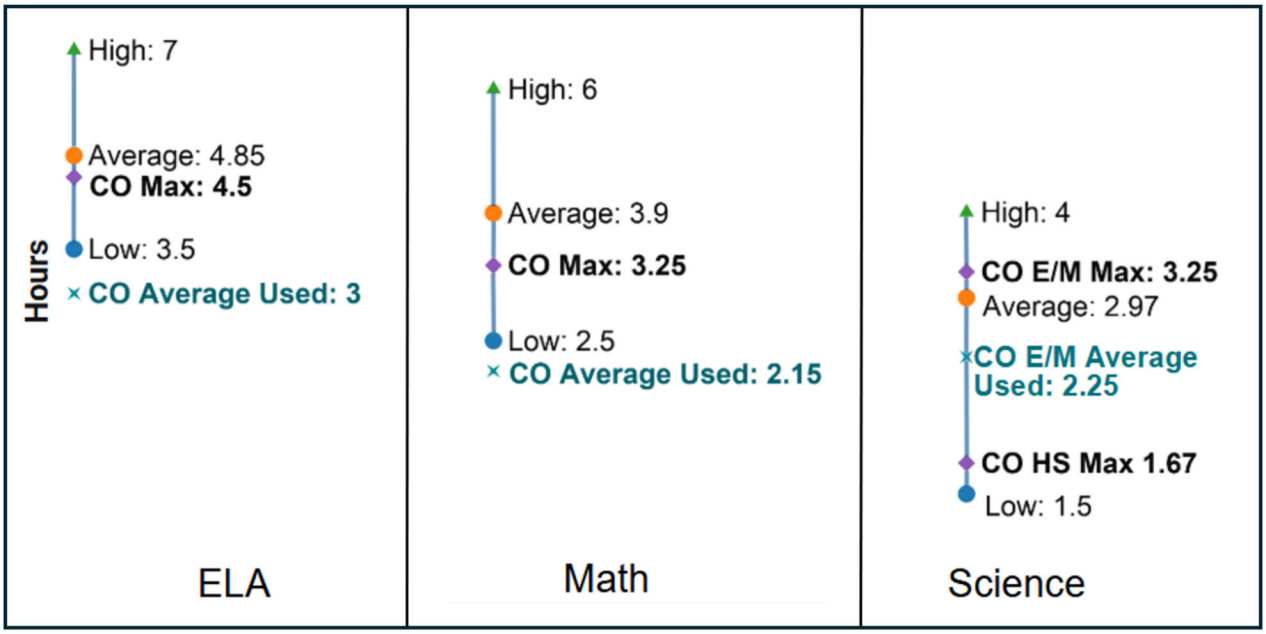
Figure 1. CMAS Maximum Testing Time by Grade Over Years (Hours)



Maximum scheduled time is distinct from the time an individual student actually uses. Districts determine whether students may leave the testing room after finishing or remain and read quietly.

A comparison of published time allowances placed Colorado generally near or below the average of comparison states. For each content area assessment – ELA, math, and science – the analysis included all states with publicly available maximum testing times.

Figure 2. Maximum Statewide Testing Time Across States (Hours)



These figures compare scheduled maximums, not actual student working time, and states differ in assessment design and reporting expectations. Colorado's 2026 mathematics maximum time allowance was 3.25 hours across test sessions, compared with an average of 3.9 hours among comparison states. On average, Colorado students used less than the provided time, completing math test sessions in 2 hours and 9 minutes (2.15 hours).

Research Questions and Methods

CDE organized the study into three phases examining complementary approaches to reducing testing time: shortening existing assessments, changing test delivery, and improving administration. Across all three phases, the study considered technical quality, standards coverage, the reliability and usefulness of student achievement and growth results, cost, and administration. Test length was an important consideration because assessments must maintain sufficient reliability to support valid and reliable student growth calculations, in addition to meaningful achievement results. Technical experts and Colorado stakeholders helped interpret the findings and weigh potential changes against the assessment features and information they prioritized. Appendix A provides the full technical study, including the simulation methods, assessment blueprints, statistical analyses, and detailed results underlying the findings summarized in this report.

Phase 1: Shorter Fixed-Form Assessments

How much testing time could be saved by removing items, and what would be lost in score precision, standards coverage, and reporting? Using spring 2025 operational content and response data from grades

4 and 6, the analyses compared current and reduced-length mathematics and ELA forms. Measures included completion time, reliability, measurement error, performance-level classification, overall score relationships, and detailed reporting. Detailed comparisons of full-length and shortened assessments, including reliability, classification accuracy, measurement precision, and impacts on reporting, are provided in Appendix A.

Phase 2: Computer Adaptive Testing

Could adaptive delivery shorten assessments while preserving technical quality, standards coverage, and current uses? Simulations examined item-level adaptive testing in mathematics and multistage adaptive testing in ELA, where questions are grouped with passages. The analyses evaluated test length and measurement precision under content-constrained designs and an illustrative unconstrained mathematics condition. These simulations evaluated potential designs rather than operational administrations. Appendix A provides the adaptive testing simulation designs and results, including comparisons of test length, measurement precision, and performance under different design constraints.

Phase 3: Operational and Administrative Changes

Could scheduled time and testing sessions be reduced without removing assessed content? CDE examined student completion times, unit structure, scheduling, and make-up logistics, alongside published time allowances in other states. This phase considered both available testing time and the additional demands each session places on schools.

CDE considered the evidence across all three phases together with Technical Advisory Committee (TAC) review and Technical Advisory Panel (TAP) and Accountability Work Group (AWG) feedback. The spring 2027 assessment changes draw on Phase 3 opportunities while retaining content and reporting that findings from Phases 1 and 2 and stakeholder priorities supported preserving.

Findings by Study Phase

Phase 1: Shorter Fixed-Form Assessments

Mathematics Findings

A simulated mathematics blueprint with approximately 30 percent fewer score points retained a strong relationship with the full form, but reliability and classification precision declined. Grade 4 overall reliability declined from 0.93 to 0.89 and grade 6 from 0.92 to 0.88. Although these reductions in overall reliability were relatively modest, they represented a loss of measurement precision that warranted consideration given the use of overall assessment results for student performance reporting, student growth calculations, and school and district accountability.

The impacts on detailed reporting were more significant. For example, the reduced assessments contained only one item measuring Modeling and Application, leaving insufficient information to support continued reporting of student performance in that category. While a shorter blueprint could be considered if these tradeoffs were acceptable, similarity in overall scores alone did not establish that all current reporting uses could be preserved.

ELA Findings

The current ELA test design requires students to read passages and respond to passage-based questions, including selected-response reading items and two open-ended written responses. Because of the time taken by students to read passages and craft written responses, removing items does little to reduce the time needed for these tasks. In the simulated reduced forms, median observed time declined by 28 minutes in grade 4 and by three minutes in grade 6. At grade 4, overall reliability declined from 0.88 to 0.83, representing a loss of measurement precision, while vocabulary subclaim reliability declined from 0.61 to 0.38, a level too low to support continued reporting of student performance in that category. These tradeoffs, particularly the loss of detailed reporting, do not support shortening of ELA assessments through item removal. More meaningful time reductions would require changes to the underlying test design, such as using fewer or shorter reading passages or modifying writing tasks, rather than simply removing individual items. Changes of this scope would require new content development and field testing before implementation and would likely disrupt comparability with prior years' results, requiring a new baseline for interpreting student performance over time. Such changes could also affect student growth calculations and school and district accountability determinations. As a result, any substantial redesign would require a multiyear development and implementation process, with careful consideration of impacts on standards coverage, score reliability, student growth calculations, and accountability. Importantly, even a substantial redesign would not necessarily result in meaningful reductions in testing time while preserving the assessment's current reporting functions.

Implications for Reporting

The finer the reported result, the more sensitive it was to shortening. Removing items narrowed the evidence for particular standards and subclaims. A high correlation between shortened and full scores therefore does not establish that the shortened form supports every current reporting use equally well.

That distinction matters because CMAS is used beyond an overall performance level. Families and educators see more specific results, while districts and the state use achievement, growth, and trends to understand performance over time. A reduced form can appear similar on an overall score comparison while leaving less evidence for detailed interpretations and less stable classification near performance-level cut scores.

Phase 2: Computer Adaptive Testing

Adaptive selection can better match question difficulty to a student and improve measurement precision across the performance range. It must still cover the breadth and depth of the standards for each student. With these content rules, simulations did not establish a substantial reduction in average test length. Without them, simulated mathematics tests ranged from 13 to 245 items, and the average number needed to reach the precision target was roughly the same as the full form. Such an unconstrained design could provide substantially different standards coverage across students and is not a viable accountability design.

Current CMAS item banks were built for fixed forms. An operational adaptive program would require substantial item development across standards and achievement levels which would include associated additional cost, design and validation of content controls, and changes to detailed academic standard-level and item-level reporting. The study supports continued technical research of adaptive designs, but it does not support adopting them in the near term to reduce testing time.

In the mathematics simulation, the test adapted after each question; the ELA used multistage adaptation because its questions are tied to reading passages and must be administered together. An operational adaptive test would need enough calibrated items or passage sets at each difficulty level to cover the standards for all students, as well as exposure controls and evidence that the resulting scores and reports support current uses.

Phase 3: Operational and Administrative Changes

Observed completion times and maximum allowances answer different questions. Most students finished before the maximum, but some used considerably more time than the majority. Analyses of the existing three-unit design did not support simply reducing allowances across all assessments. Considering timing alongside unit structure and stakeholder feedback identified a different approach: consolidate mathematics and elementary and middle school science into two 75-minute units, reducing the overall allowance while retaining assessed content.

Reducing the number of test units also eases logistical demands that student working-time measures do not capture, including room and device scheduling, directions, transitions, proctoring, and make-up arrangements. Together, the mathematics and science changes remove two sessions for students taking both subjects. ELA's passage and writing structure limits opportunities to reduce testing time while retaining the current design. Based on the study findings and stakeholder feedback, reducing testing time through operational changes that preserve assessed content and reporting, rather than shortening the assessments by removing content, is recommended. Consistent with this approach, CDE will implement reduced testing time and fewer sessions for mathematics and science beginning in spring 2027 and evaluate the changes using operational evidence following implementation.

Technical Review, Stakeholder Feedback, and Basis for Assessment Adjustments

CDE engaged the TAC on study design, methods, and technical findings and shared findings with the TAP and AWG to consider administration and the use of results in schools, districts, and the state accountability system.

In doing so, the study paired independent technical review with feedback from a range of Colorado education stakeholders, including school and district leaders, superintendents, assessment and accountability practitioners, and technical experts.

Table 1: Stakeholder Groups Consulted

Group	Role in the process	Contribution
Technical Advisory Committee (TAC)	Independent national and local technical experts in psychometrics, assessment design, and special populations	Reviewed study design, methods, technical findings, risks, and defensibility of potential changes.
Technical Advisory Panel (TAP)	Colorado assessment stakeholders and practitioners with technical expertise	Considered technical and practical tradeoffs, including validity and reliability, growth and trends, reporting, administration, scheduling, and performance frameworks, and student experience.
Accountability Work Group (AWG)	Colorado district, school, and community accountability stakeholders	Considered technical and practical tradeoffs through a practitioner lens, with particular attention to school realities, including scheduling, make-ups, instructional disruption, reporting, growth, and student experience.

Summaries of the TAP and AWG discussions, including the considerations and recommendations that informed CDE's decisions, are provided in Appendices B and C, respectively. Technical review identified measurement risks; stakeholder discussions weighed those risks against disruption to instruction and the information needed from CMAS. Across roles, participants expressed a strong and consistent preference: **reduce time and administrative burden where it can be done responsibly, while preserving the technical rigor, usefulness, and reporting value of CMAS.**

- Stakeholders identified fewer testing units as especially meaningful because each unit requires additional scheduling, proctoring, directions, make-up opportunities, and interruption to instruction.
- They strongly valued preserving valid and reliable results, student growth scores, information, writing tasks, and detailed reporting, and did not support sacrificing these for modest reductions in testing time.
- They supported using actual student timing data to shorten scheduled time while retaining sufficient time for students who need it.
- They noted that computer adaptive testing does not necessarily shorten testing and could create comparability challenges when/if paper-pencil testing is needed.

CDE used evidence from all three phases, together with stakeholder input, to determine changes to the 2027 assessment administration. The findings supported reducing mathematics and science sessions and overall scheduled testing time while retaining assessed content and current reporting. ELA remains unchanged because its reading and writing requirements and current test design limit opportunities for meaningful time reductions. The study also found that adaptive testing did not emerge as an effective strategy for meaningfully reducing testing time. Given Colorado's standards coverage and reporting requirements, adaptive testing would require larger item banks and additional development costs while providing limited time savings.

Spring 2027 Actions and Implementation

Based on the technical findings and stakeholder feedback, CDE determined that reducing mathematics and science testing sessions and maximum available time, while retaining the current ELA design and not adopting adaptive testing, provides the most appropriate approach to reducing testing burden while preserving current assessment functions. Elementary and middle school mathematics and science will each move from three 1-hour, 5-minute units to two 1-hour, 15-minute units. For each subject, this reduces testing from three sessions to two and decreases the maximum available testing time by 45 minutes - from 3 hours, 15 minutes to 2 hours, 30 minutes, a reduction of approximately 23 percent. Students taking both subjects will have two fewer testing sessions and 1 hour, 30 minutes less maximum available testing time. The mathematics and science changes are based on analyses of student completion times and administration requirements and reductions will be achieved without removing test questions, reducing the content assessed, or changing the level of information reported. These times represent the maximum time schools will make available; many students finish before the full time is used. These adjustments reflect the combined technical findings, timing analyses, and stakeholder feedback described above and documented in Appendices A–C.

Table 2: Assessment Reductions for 2027

Assessment	Spring 2026 (Maximum Time)	Spring 2027 (Maximum Time)	Reduction
Mathematics	3 units × 1 hour 5 minutes **3 hours, 15 minutes total**	2 units × 1 hour 15 minutes **2 hours, 30 minutes total**	1 fewer unit + 45 fewer minutes
Science (Grades 5 & 8)	3 units × 1 hour, 5 minutes **3 hours 15 min total**	2 units × 1 hour 15 minutes **2 hours, 30 minutes total**	1 fewer unit + 45 fewer minutes
English Language Arts	Current schedule	No change	

Table 3: Overall Reduction for Students

Grades	Assessments	Spring 2026 (Maximum Time)	Spring 2027 (Maximum Time)	Reduction
3, 4, 6, 7*	**English Language Arts + Mathematics**	7 hours, 45 minutes	7 hours	45 minutes + 1 less unit
5, 8	**English Language Arts + Mathematics + Science**	11 hours	9.5 hours	1.5 hours + 2 fewer units
*Social Studies not included due to changes in tested populations over the years.				

Table 4: Sample Spring 2027 CMAS Schedule
One Unit Per Day Per Student

Allows school completion of all CMAS testing in seven days. Each student takes only one CMAS unit per day; the most days on which an individual student would spend a portion of the school day testing (50 to 90 minutes) is seven days. The average Colorado school day is 330 minutes for elementary schools and 360 minutes for secondary schools.

- Grades 3, 4 and 6 – five days (three ELA units, two math units)
- Grade 7 – six days (three ELA units, two math units, one social studies unit)
- Grades 5 and 8 – seven days (three ELA units, two math units, two science units)
- Grade 11 – two days (two science units)

Available State Testing Window April 5–23, 2027				
5 ELA Unit 1 (Grades 3-8) 90 minutes	6 ELA Unit 2 (Grades 3-8) 90 minutes	7 ELA Unit 3 (Grades 3-8) 90 minutes	8 Math Unit 1 (Grades 3-8) 75 minutes	9 Math Unit 2 (Grades 3-8) 75 minutes
12 Science Unit 1 (Grades 5, 8, 11) OR Social Studies Unit 1 (Grade 7) 50-75 minutes	13 Science Unit 2 (Grades 5, 8, 11) 50-75 minutes	14	15	16
19	20	21	22	23

Table 5: Sample Spring 2027 CMAS Schedule
Up to Two Units Per Day Per Student

Allows school completion of all CMAS testing in four days. Each student takes up to two CMAS units per day; the most days on which an individual student would spend a portion of the school day testing (50 to 165 minutes) is four days. The average Colorado school day is 330 minutes for elementary schools and 360 minutes for secondary schools.

- Grades 3, 4 and 6 – two and a half days (three ELA units, two math units)
- Grade 7 – three days (three ELA units, two math units, one social studies unit)
- Grades 5 and 8 – three and a half days (three ELA units, two math units, two science units)
- Grade 11 – one day (two science units)

Available State Testing Window April 5–23, 2027				
5 AM: ELA Unit 1 (Grades 3-8) PM: Math Unit 1 (Grades 3-8) 165 minutes	6 AM: ELA Unit 2 (Grades 3-8) PM: Math Unit 2 (Grades 3-8) 165 minutes	7 AM: ELA Unit 3 (Grades 3-8) PM: Science Unit 1 (Grades 5, 8, 11) 50-165 minutes	8 AM: Science Unit 2 (Grades 5, 8, 11) OR Social Studies Unit 1 (Grade 7) 50-75 minutes	9
12	13	14	15	16
19	20	21	22	23

These changes preserve assessed content coverage and current detailed reporting. ELA will retain its current design and schedule. Fewer units reduce transitions, directions, proctoring, and make-up arrangements; calendar effects will vary with local schedules. CDE will update administration directions, scheduling guidance, communications, and training in advance of the spring window.

These changes represent actionable reductions identified through the study that can be implemented while maintaining current assessment content and reporting expectations. The analyses also established that more substantial reductions would involve tradeoffs in assessment quality, content coverage, reporting, or comparability that warrant careful consideration before further changes are made. After spring 2027, CDE will examine completion times and field experience, including timing analyses from the administration, and review the evidence with its standing advisory groups.

Conclusion

CDE completed the three-phase study required by House Bill 25-1278 and used technical evidence and stakeholder input to establish concrete changes for spring 2027. Mathematics and science changes reduce sessions and maximum available testing time while preserving assessed content and reporting. ELA

remains unchanged because the reductions studied produced limited time savings while meaningfully weakening measurement and reporting.

The study also identified constraints on further reductions. Shorter mathematics forms reduced reliability and detailed reporting, while larger ELA reductions would require changes to passages, writing tasks, or the overall test design that would take several years to implement and could disrupt trend data and accountability. Computer adaptive testing similarly did not emerge as an effective time-reduction strategy because content and reporting requirements limited potential time savings while adding development costs.

Additional reductions are possible, but will require careful consideration of impacts on standards coverage, precision, achievement and growth reporting, and longitudinal comparisons. CDE will continue examining targeted improvements, including shorter reading passages, fewer field-test items, and operational efficiencies, through ongoing analyses and its established technical and stakeholder advisory processes. This work will also inform the periodic reviews of assessment technology required by House Bill 25-1278 beginning in 2030.

Appendices

- *Colorado Measures of Academic Success Testing Time Reduction Study Results.* Pearson, 2026.
- *Technical Advisory Panel (TAP) Feedback and Recommendations Summary.* April 2026.
- *Accountability Work Group (AWG) Feedback and Recommendations Summary.* May 2026.

Appendix A: Technical Study: Colorado Measures of Academic Success (CMAS) Testing Time Reduction Study

Colorado Measures of Academic Success (CMAS) Testing Time Reduction Study

March 2026 | Prepared by Pearson

Table of Contents

Table of Contents 17

List of Tables 18

List of Figures 19

1. Introduction 21

    1.1. Federal Accountability and the Every Student Succeeds Act (ESSA) 21

    1.2. Computer Adaptive Testing (CAT) 22

2. Methods 24

    2.1. Test Shortening 24

    2.1.1. Student Data 24

    2.1.2. Assessment Reduction 24

    2.1.3. Item Response Theory Model 28

    2.1.4. Scoring Tables 28

    2.1.5. Reporting Scales and Performance Levels 28

    2.2. Item-Level CAT Simulation (Mathematics) 29

    2.3. Multistage Adaptive Testing Simulation (ELA) 30

3. Results 32

    3.1. Test Shortening 32

    3.1.1. Scale Score Summary Statistics 32

    3.1.2. Performance Level Agreement 32

    3.1.3. Correlations and Scale Score Differences 35

    3.1.4. Summative Test Characteristic, Test Information, and CSEM Curves 40

    3.1.5. Reliability Indices 47

    3.1.6. Timing Analyses 49

    3.2. CAT Simulation (Mathematics) 58

    3.3. Multistage Adaptive Testing Simulation (ELA) 65

4. Summary 71

5. Conclusion 74

6. References 76

List of Tables

Table 1.1. CAT Designs 22
Table 2.1. Number of Simulated Examinees 24
Table 2.2. Blueprint and Adjusted Points—Grade 4 Mathematics 24
Table 2.3. Blueprint and Adjusted Points—Grade 6 Mathematics 25
Table 2.4. Blueprint and Adjusted Points—Grade 4 ELA 26
Table 2.5. Blueprint and Adjusted Item Counts—Grade 4 ELA 26
Table 2.6. Blueprint and Adjusted Counts by Subdomain—Grade 4 ELA 26
Table 2.7. Blueprint and Adjusted Points—Grade 6 ELA 27
Table 2.8. Blueprint and Adjusted Item Counts—Grade 6 ELA 27
Table 2.9. Blueprint and Adjusted Counts by Subdomain—Grade 6 ELA 27
Table 3.1. Summary Statistics for Full and Adjusted Scores 32
Table 3.2. Performance Level Agreement 33
Table 3.3. Overall Performance Level Percent Agreement—Grade 4 Mathematics 33
Table 3.4. Overall Performance Level Percent Agreement—Grade 6 Mathematics 34
Table 3.5. Overall Performance Level Percent Agreement—Grade 4 ELA 34
Table 3.6. Overall Performance Level Percent Agreement—Grade 6 ELA 34
Table 3.7. Pearson Correlations between Full and Reduced Length Tests 35
Table 3.8. Pearson Correlations between Subclaims and Overall Scores for the Full and Reduced Length Tests—Mathematics 35
Table 3.9. Pearson Correlations between Subclaims and Overall Scores for the Full and Reduced Length Tests—ELA 35
Table 3.10. Coefficient Alpha for the Full and Reduced Length Tests—Mathematics 48
Table 3.11. Coefficient Alpha for the Full and Reduced Length Tests—ELA 48
Table 3.12. Classification Accuracy and Consistency for the Full and Reduced Length Tests 48
Table 3.13. Classification Accuracy at the Cut Scores for the Full and Reduced Length Tests 49
Table 3.14. Classification Consistency at the Cut Scores for the Full and Reduced Length Tests 49
Table 3.15. Timing Descriptive Statistics for the Full and Reduced Length Tests 50
Table 3.16. Summary Statistics for Full and Reduced Length Scores—Mathematics 58
Table 3.17. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 4 Math 58
Table 3.18. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 4 Math 59
Table 3.19. Overall Performance Level Percent Agreement Comparing the Unconstrained CAT Simulation Results to Initial “True” Ability—Grade 4 Math 59
Table 3.20. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 6 Math 60
Table 3.21. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 6 Math 60
Table 3.22. Overall Performance Level Percent Agreement Comparing the Unconstrained CAT Simulation Results to Initial “True” Ability—Grade 6 Math 61
Table 3.23. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form and Unconstrained Conditions—Grade 4 Math 61
Table 3.24. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form and Unconstrained Conditions—Grade 6 Math 62
Table 3.25. Summary Statistics for Full and Reduced Length Scores—ELA 65

Table 3.26. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 4 ELA 66
Table 3.27. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 4 ELA 66
Table 3.28. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 6 ELA 67
Table 3.29. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial “True” Ability—Grade 6 ELA 67
Table 3.30. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form Conditions—Grade 4 ELA 68
Table 3.31. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form Conditions—Grade 6 ELA 68

List of Figures

Figure 2.1. CMAS ELA Multistage Configuration 31
Figure 3.1. Overall Full vs. Reduced Length Tests—Grade 4 Math 37
Figure 3.2. Overall Full vs. Reduced Length Tests—Grade 6 Math 37
Figure 3.3. Overall Full vs. Reduced Length Tests—Grade 4 ELA 38
Figure 3.4. Overall Full vs. Reduced Length Tests—Grade 6 ELA 38
Figure 3.5. Overall Scale Score Differences Between Adjusted and Full Length Tests—Grade 4 Math 39
Figure 3.6. Overall Scale Score Differences Between Adjusted and Full Length Tests—Grade 6 Math 39
Figure 3.7. Overall Scale Score Differences Between Reduced and Full Length Tests—Grade 4 ELA..... 40
Figure 3.8. Overall Scale Score Differences Between Reduced and Full Length Tests—Grade 6 ELA..... 40
Figure 3.9. Overall Percentage TCCs for Full Reduced Length Tests—Grade 4 Math 42
Figure 3.10. Overall Percentage TCCs for Full Reduced Length Tests—Grade 6 Math 42
Figure 3.11. Overall Percentage TCCs for Full Reduced Length Tests—Grade 4 ELA 43
Figure 3.12. Overall Percentage TCCs for Full Reduced Length Tests—Grade 6 ELA 43
Figure 3.13. Overall Information Curves for Full and Reduced Length Tests—Grade 4 Math 44
Figure 3.14. Overall Information Curves for Full and Reduced Length Tests—Grade 6 Math 44
Figure 3.15. Overall Information Curves for Full and Reduced Length Tests—Grade 4 ELA..... 45
Figure 3.16. Overall Information Curves for Full and Reduced Length Tests—Grade 6 ELA..... 45
Figure 3.17. Overall CSEM Curves for Full and Reduced Length Tests—Grade 4 Math 46
Figure 3.18. Overall CSEM Curves for Full and Reduced Length Tests—Grade 6 Math 46
Figure 3.19. Overall CSEM Curves for Full and Reduced Length Tests—Grade 4 ELA..... 47
Figure 3.20. Overall CSEM Curves for Full and Reduced Length Tests—Grade 6 ELA..... 47
Figure 3.21. Histogram of Testing Time for The Full-Length Tests—Grade 4 Math 50
Figure 3.22. Histogram of Testing Time for The Reduced-Length Tests—Grade 4 Math 51
Figure 3.23. Histogram of Testing Time for The Full-Length Tests—Grade 6 Math 52
Figure 3.24. Histogram of Testing Time for The Reduced-Length Tests—Grade 6 Math 53
Figure 3.25. Histogram of Testing Time for The Full-Length Tests—Grade 4 ELA..... 54
Figure 3.26. Histogram of Testing Time for The Reduced-Length Tests—Grade 4 ELA 55
Figure 3.27. Histogram of Testing Time for The Full-Length Tests—Grade 6 ELA..... 56
Figure 3.28. Histogram of Testing Time for The Reduced-Length Tests—Grade 6 ELA 56
Figure 3.29. Overall Full-Length Fixed Form Simulation vs. “True” Ability—Grade 4 Math 62

Figure 3.30. Overall Reduced-Length Fixed Form Simulation vs. “True” Ability—Grade 4 Math 63
Figure 3.31. Unconstrained Simulation vs. “True” Ability—Grade 4 Math 63
Figure 3.32. Overall Full-Length Fixed Form Simulation vs. “True” Ability—Grade 6 Math 64
Figure 3.33. Overall Reduced-Length Fixed Form Simulation vs. “True” Ability—Grade 6 Math 64
Figure 3.34. Unconstrained Simulation vs. “True” Ability—Grade 6 Math 65
Figure 3.35. Overall Full-Length Fixed Form Simulation vs. “True” Ability—Grade 4 ELA 68
Figure 3.36. Overall Reduced-Length Fixed Form Simulation vs. “True” Ability—Grade 4 ELA 69
Figure 3.37. Overall Full-Length Fixed Form Simulation vs. “True” Ability—Grade 6 ELA 69
Figure 3.38. Overall Reduced-Length Fixed Form Simulation vs. “True” Ability—Grade 6 ELA 70

1. Introduction

The Colorado Measures of Academic Success (CMAS) assessments are a critical component of the state's education and accountability system. In an ongoing effort to ensure that state assessments are both technically robust and as efficient as possible, the Colorado legislature has sought expert guidance on potential improvements. This research study was conducted in direct response to an assessment study requirement established by Colorado House Bill 25-1278 and the Colorado 1241 Accountability, Accreditation, Student Performance, and Resource Inequity Task Force Report (November, 2024) that recommended the state investigate methods for shortening statewide assessments and explore the implementation of adaptive assessment technologies (Colorado General Assembly, 2023).

The core objective of this study is to provide the Colorado Department of Education (CDE) with a data-driven analysis of various pathways to potentially reduce student testing time while upholding the rigorous technical and content standards required for federal accountability. This report examines three distinct approaches:

- Shorten the current fixed-form linear assessments
- Implement a multi-stage adaptive testing (MST) design for English language arts (ELA)
- Implement an item-level computer adaptive testing (CAT) design for mathematics

This investigation focuses on CMAS Mathematics for item-level adaptive tests, as their structure is most amenable to that design. An MST is explored for CMAS ELA because they are comprised of passage-based item sets.

1.1. Federal Accountability and the Every Student Succeeds Act (ESSA)

The design and administration of the CMAS assessments are governed by the requirements of the federal Every Student Succeeds Act (ESSA), which reauthorized the Elementary and Secondary Education Act of 1965 (ESEA). ESSA grants states more flexibility than its predecessor, No Child Left Behind (NCLB), but it maintains stringent requirements for statewide assessment and accountability systems to ensure educational equity and academic transparency. Under ESSA, states like Colorado are required to:

- Administer annual assessments in English language arts (ELA) and mathematics for all students in grades 3–8 and once in high school.
- Ensure these assessments are aligned with the state's challenging academic content standards.
- Demonstrate that the assessments are valid, reliable, and fair for all students, including subgroups such as students with disabilities, English learners, and students from major racial and ethnic groups.

Colorado uses CMAS as a central component of its School and District Performance Frameworks under the Education Accountability Act (SB09-163; C.R.S. 22-11-101 et seq.). Generally:

- Achievement indicators: CMAS proficiency and scale scores contribute to school and district ratings.
- Growth indicators: Median student growth percentiles (MGPs) derived from longitudinal CMAS performance are a prominent component of ratings.
- Accountability ratings: Performance, Improvement, Priority Improvement, and Turnaround determinations are informed by CMAS achievement and growth, alongside other indicators (e.g., graduation, PSAT/SAT).

- Identification for support: CMAS results inform identification of schools and districts for comprehensive or targeted support and improvement.

Colorado uses scores from the CMAS as a primary indicator within its broader accountability framework, which identifies schools for support and improvement. Student performance is categorized into performance levels, and these results are used to evaluate school quality and student growth.

To ensure these high stakes are met with technically sound instruments, each state's assessment system must pass a rigorous federal peer review process. This process scrutinizes evidence of the test's validity, reliability, fairness, and alignment to standards. Failure to meet these federal requirements can have significant consequences for the state, which may include:

- The withholding of federal Title I funds, which support schools with high percentages of children from low-income families.
- The imposition of corrective action plans by the U.S. Department of Education.
- A loss of public confidence in the state's ability to measure student achievement and hold schools accountable for performance.

Therefore, any proposed changes to the CMAS, including efforts to shorten the test, must be carefully evaluated to ensure they do not compromise the technical quality necessary to pass federal peer review and maintain the integrity of Colorado's accountability system.

1.2. Computer Adaptive Testing (CAT)

Adaptive testing is a sophisticated assessment methodology that tailors the test to the individual student's ability level. Unlike a traditional fixed-form test where all students receive the same set of questions, a CAT selects items for each student in real-time based on their responses to previous items. The fundamental goal is to achieve a more precise estimate of a student's knowledge and skills, often with fewer items than a traditional test.

There are several forms of adaptive testing, each with different characteristics, as described in Table 1.1. The primary advantage of adaptive testing is its potential for greater measurement precision, as it avoids asking questions that are far too easy or too hard for the student. This can lead to a more engaging student experience and, in many cases, a shorter test. However, CAT presents significant technical and logistical challenges. It requires a very large bank of items that have been pre-calibrated and are psychometrically sound, and it involves complex algorithms for item selection and scoring.

Table 1.1. CAT Designs

CAT Design	Description
Item-Level CAT	This is the most granular and flexible form. After every single item a student answers, the system re-estimates their ability and selects the next item from a large item bank that is optimally informative for a student at that estimated ability level.
Testlet-Based CAT	In this model, items are pre-assembled into small groups called "testlets" (e.g., a set of questions related to a single reading passage). The adaptation occurs between “testlets” rather than after each item. The system selects the next testlet based on the student's performance on the previous one.
Multi-Stage Adaptive Testing (MST)	This is a more structured adaptive design. The test is organized into a small number of stages (typically two or three). All students begin with a common "routing" module in Stage 1. Based on their performance, they are routed to one of several parallel

CAT Design	Description
	modules in Stage 2, for instance, an easier, medium, or harder module. This provides a balance between adaptation and the content control offered by fixed forms.

A common misconception is that adaptive testing always results in shorter tests. While CAT can achieve a precise ability estimate relatively quickly, its use in a high-stakes accountability system like ESSA is heavily constrained by the need for **content validity**. ESSA requires that every student's test assesses the full breadth and depth of the state's academic standards. Therefore, a CAT must be programmed to continue administering items until all content blueprint requirements are met for every student, even if a precise score has already been determined. This "content-balancing" constraint means that while the *average* test length may decrease, it does not guarantee a shorter test for every single student and places a limit on potential time savings. Unconstrained CATs can be shortest but typically will not satisfy ESSA content requirements for accountability and peer review.

Given this requirement then, a transition to adaptive testing for CMAS would require a minimal blueprint be met that assesses the full breadth and depth of the Colorado Academic Standards (CAS). Within the context of this research then, the initial focus of the study was to explore options around existing test blueprint reductions that our content experts feel would still be viable under ESSA and explored the technical impacts to measurement characteristics and score reporting.

This report presents the methodology, analyses, findings, and discussion from a study designed to evaluate the viability of shortening the CMAS assessments. The study systematically explored the trade-offs between testing time, measurement precision, and content validity across areas.

- **Phase 1: Shortened Linear Analysis:** The first phase established a psychometric baseline using the current full-length CMAS based on the operational tests and student response data from the CMAS spring 2025 administration. For comparison, tests based on shortened versions of the same assessments were crafted in relation to ESSA requirements and analyzed for comparison to baseline.
- **Phase 2: Item-Level Computer Adaptive (CAT) and Multi-Stage Adaptive (MST) Simulations:** The second phase explored using the same full-length and reduced blueprints from Phase 1 to run adaptive test simulations. The simulations for item-level adaptive also included a simulation to explore the scenario where no content constraints were imposed on the simulated examinees. These analyses evaluated the balance between achieving measurement precision and satisfying the strict content-balancing constraints required by ESSA.

For these analyses, the CMAS Mathematics and ELA assessments in grades 4 and 6 were chosen for the study. The following section describes the methods used to carry out the research.

2. Methods

Test Shortening

This section discusses the data used for the analyses, the adjustments to the mathematics and ELA grades 4 and 6 test forms, the item response theory (IRT) model, generating the scoring tables, and the reported scales and performance levels.

2.1.1. Student Data

The data for this report was spring 2025 CMAS Mathematics and ELA assessment results in grades 4 and 6 for Colorado students. This administration consisted of one computer-based (CBT) operational form and one paper-based (PBT) operational form in addition to several accommodated forms. This study only included the CBT operational form. Table 2.1 presents the total simulated examinees.

Table 2.2. Number of Simulated Examinees

Assessment	#Simulees
Mathematics 4	52,518
Mathematics 6	52,236
ELA 4	52,252
ELA 6	52,414

2.1.2. Assessment Reduction

2.1.2.1. Mathematics

Items for the reduced blueprint were selected to retain the proportion of items from each of the types and claims. Table 2.2 and Table 2.3 present the number of score points by type and claim for the full spring 2025 grades 4 and 8 mathematics assessments and the adjusted assessments. For both grades, the tests were reduced from 51 to 34 points (30%).

Table 2.3. Blueprint and Adjusted Points—Grade 4 Mathematics

	Blueprint Points	Reduced Points	Blueprint Percentages	Reduced Percentages
Type 1	31	21	61-62%	62%
1.1	19	13	37-38%	38%
1.2	12	8	24%	24%
1.4	-	-	-	-
Claim A	24	14	47-48%	41%
Claim B	7	7	14-16%	21%
Type 2	10-11	7	20-22%	21%
2.3	3-6	3	6-12%	9%
2.4	4-8	4	8-16%	12%
Claim C	10-11	7	20-22%	21%
Type 3	9	6	18%	18%
3.3	3	0 or 6	6%	0 or 18%
3.6	6	0 or 6	12%	0 or 18%
Claim D	9	6	18%	18%
Calculator	–	–	–	–
Non-calculator	50-51	34	100%	100%
Total	50-51	34	100%	100%

Table 2.4. Blueprint and Adjusted Points—Grade 6 Mathematics

	Blueprint Points	Reduced Points	Blueprint Percentages	Reduced Percentages
Type 1	31	21	61-62%	62%
1.1	19	11	37-38%	33%
1.2	8	6	16%	18%
1.4	4	4	8%	12%
Claim A	20	14	39-40%	41%
Claim B	11	7	22%	21%
Type 2	10-11	7	20-22%	21%
2.3	3-6	3	6-12%	9%
2.4	4-8	4	8-16%	12%
Claim C	10-11	7	20-22%	21%
Type 3	9	6	18%	18%
3.3	3	0 or 6	6%	0 or 18%
3.6	6	0 or 6	12%	0 or 18%
Claim D	9	6	18%	18%
Calculator	36-37	24	72-73%	72%
Non-calculator	14	10	27-28%	28%
Total	50-51	34	100%	100%

2.1.2.2. ELA

Shortening the testing time for ELA is different than for mathematics in that so much of the testing time involves reading and writing. There are fewer items on the ELA assessments and less room to reduce the blueprints and maintain representativeness in line with ESSA.

The ELA blueprints at grades 4 and 6 include two prose-constructed responses (PCRs) that measure written expression and conventions (with raw scores of 3 and 4 points for grade 4 and 4 points for grade 6 across the two PCRs). Based on content requirements reflective of the standards, both PCRs must remain in the assessments. As a result, potential reductions in testing time must occur primarily through adjustments to the number of non-PCR items. However, non-PCR items represent a relatively small portion of total testing time at the item level. Consequently, reductions in this portion of the test provide only limited opportunities to substantially decrease total testing time while maintaining blueprint coverage and measurement precision. So, the shortened option presented here represents a balanced approach that maintains the required PCRs while reducing the weight of the written expression trait from three to two, preserves representation across test blueprint claims, and modestly reduces the number of non-PCR items.

Table 2.4 – Table 2.9 present the number of score points by type and claim for the full spring 2025 grades 4 and 6 ELA assessments and the adjusted ELA assessments. For the grade 4 ELA assessments, the reduced form was from 23 to 14 items. Grade 6 was reduced from 24 to 19 items.

Table 2.5. Blueprint and Adjusted Points—Grade 4 ELA

	Blueprint Points vs Reduction Points											
	RI	RI Reduction	RL	RL Reduction	V	V Reduction	WE	WE Reduction	WKL	WKL Reduction	Total	Total Reduction
Grade 4			(PCR Points)									
LAT	-		6 or 8+(4)	2+4	4	2	12	8	3	3	0	17
Short (info)	6	4	-		2 or 4	0 or 2	-		-			4
NWT	-		6 or 8	6			9	6	3.00	3		15
Long/ Paired (info)	12	8	-		2 or 4	2	-		-			10
Total	18	12	18	12	10	4	21	14	6	6	73	48
Percentage Points	25%	25%	or 20 25-27%	25%	or 8 11-14%	8.33%	29%	25.00%	8%	12.50%	-	

*RI= Reading Information, RL= Reading Literature, V=Vocabulary, WE=Written Expression trait, WKL=Convention trait, LAT = Literary Analysis Task, NWT= Narrative Writing Task

Table 2.6. Blueprint and Adjusted Item Counts—Grade 4 ELA

	Items									
	RI	RI Reduction	RL	RL Reduction	V	V Reduction	PCR	PCR Reduction	Total	Total Reduction
Grade 4										
LAT	-		3 or 4	1	2	1	1	1	6 or 7	3
Short (info)	3	2	-		1 or 2	0 or 1	-		4 or 5	2
NWT	-		3 or 4	3	-		1	1	4 or 5	4
Long/ Paired (info)	6	4	-		1 or 2	0 or 1	-		7 or 8	4
Total	9	6	7 or 8	4	4 or 5	2	2	2	23	14
Percentage Points										

* RI= Reading Information, RL= Reading Literature, V=Vocabulary, PCR=Prose Constructed Response, LAT = Literary Analysis Task, NWT= Narrative Writing Task

Table 2.7. Blueprint and Adjusted Counts by Subdomain—Grade 4 ELA

Grade	KI Lit (RL)	KI Lit(RL) Reduction	KI Info (RI)	KI info (RI) Reduction	Int	Int Reduction	C&S	C&S Reduction	V	V Reduction
4	22-24	14	22-24	16	8-14	8	12-18	8	8-10	4

* KI Lit= Key Ideas and Details, RI= Reading Information, RL= Reading Literature, Int= Integration of Knowledge and Ideas, C&S= Craft and Structure, V=Vocabulary

Table 2.8. Blueprint and Adjusted Points—Grade 6 ELA

	Blueprint Points vs. Reduction Points											
	RI	RI Reduction	RL	RL Reduction	V	V Reduction	WE	WE Reduction	WKL	WKL Reduction	Total	Total Reduction
Grade 6	(PCR Points)											
RST	12+(4)	10+4	-		4	2	12	8	3	3		27
Short (info)	6	4			2 or 4	2	-		-			6
NWT	-		6 or 8	6	-		12	8	3	3		17
Paired/ Long (lit)	-		8 or 10	8	2 or 4	2	-		-			10
Total	22	18	16 or 18	14	8 or 10	6	24	16	6	6	78	60
Percentage Points	28%	30%	21-23%	23.33%	10-13%	10.00%	31%	26.67%	8%	10.00%	-	

*RI=Reading Information, RL=Reading Literature, V=Vocabulary, WE=Written Expression trait, WKL=Convention trait, RST = Research Simulation Task, NWT= Narrative Writing Task

Table 2.9. Blueprint and Adjusted Item Counts—Grade 6 ELA

	Items									
	RI	RI Reduction	RL	RL Reduction	V	V Reduction	PCR	PCR Reduction	Total	Total Reduction
Grade 6										
RST	6	5	-		2	1	1	1	9	7
Short (info)	3	2	-		1 or 2	1	-		4 or 5	3
NWT	-		3 or 4	3	-		1	1	4 or 5	4
Paired/ Long (lit)	-		4 or 5	4	1 or 2	1	-		6 or 7	5
Total	9	7	8 or 9	7	4 or 5	3	2	2	24	19

* RI= Reading Information, RL= Reading Literature, V=Vocabulary, PCR=Prose Constructed Response, RST = Research Simulation Task, NWT= Narrative Writing Task

Table 2.10. Blueprint and Adjusted Counts by Subdomain—Grade 6 ELA

Grade	KI Lit (RL)	KI Lit(RL) Reduction	KI Info (RI)	KI info (RI) Reduction	Int	Int Reduction	C&S	C&S Reduction	V	V Reduction
6	18-20	20	24-28	22	10-14	8-12	18-24	18	8-10	6

* KI Lit= Key Ideas and Details, RI= Reading Information, RL= Reading Literature, Int= Integration of Knowledge and Ideas, C&S= Craft and Structure, V=Vocabulary

2.1.3. Item Response Theory Model

The spring 2025 CMAS Mathematics and ELA assessments were post-equated. The item parameter estimates from the post-equating analyses were used for both the full mathematics and ELA assessments and the adjusted assessments when creating the scoring tables. The operational IRT analyses were conducted by Pearson. The operational items in the incomplete data matrix (IDM) were concurrently calibrated with the two-parameter logistic/generalized partial credit model (2PL/GPC: Muraki, 1992). The 2PL/GPC is denoted as

$$p_{im}(\theta_j) = \frac{\exp\left[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})\right]}{\sum_{v=0}^{M_i-1} \exp\left[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{iv})\right]}$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $p_{im}(\theta_j)$ is the probability of a test taker with θ_j getting score m on item i; M_i is the number of score categories of item i with possible item scores as consecutive integers from 0 to $M_i - 1$; D is the IRT scale constant (1.7).

2.1.4. Scoring Tables

IRT ability estimates (θs) are calculated using estimates of item parameters and thetas are substituted for parameters in the formulas in the generalized partial credit model for both dichotomous and polytomous items. These estimates are then linearly transformed to the reporting scale using a slope and intercept which vary between assessments and were determined by relating two of the established cut scores on the theta metric to their fixed values on the reporting scale. The estimated conditional standard error of measurement (CSEM) for each scale score is computed.

2.1.5. Reporting Scales and Performance Levels

The CMAS Mathematics and ELA reporting scales designate student performance into one of five Performance Levels that delineate the knowledge, skills, and practices students are able to demonstrate. Level 1 indicates the lowest level of performance and Level 5 indicates the highest level of performance:

- Level 1: Did not yet meet expectations
- Level 2: Partially met expectations
- Level 3: Approached expectations
- Level 4: Met expectations
- Level 5: Exceeded expectations

Summative scale scores, which reflects performance across all items on the assessment, range from 650 to 850 and categorizes students into one of five summative performance levels with a 725 representing the threshold of Level 3, and 750 representing the threshold of Level 4 that represents college and career readiness (CCR). The threshold score for Level 5 is 796 and 788 for the grades 4 and 6 mathematics assessments, respectively. The threshold score for Level 5 is 790 for both the grades 4 and 6 ELA assessments.

2.2. Item-Level CAT Simulation (Mathematics)

A CAT simulator simulates individual examinees responding to actual test questions. It uses the same algorithms to carry out simulations as are used for operational CAT administrations. Simulation is invaluable for evaluating different administrative settings ahead of a live administration to ensure it will function optimally. In addition to providing validation feedback to adjust settings, it can also be used to carry out research investigations such as for this study. All CAT simulations for this study were run using the Pearson CAT engine (version 3.21.16).

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

For this study the item-level CAT simulations used the existing CMAS operational item banks as inputs to estimate simulated student ability. Both the item bank and selection algorithm function within an IRT-based statistical framework (described above). The items in the banks have been calibrated so that their difficulty level, or ability to discriminate between students of similar proficiencies are known. Based on these items' statistics, the point on the ability continuum where their measurement accuracy is highest and their degree of accuracy relative to other items can be calculated. As simulated examinees work through the test, their ability is measured after each item has been answered, and the next item is selected adaptively from the pool by choosing the item with the highest degree of precision at each student's current level of ability (subject to content coverage requirements and item exposure controls). For the simulation study, this is carried out by modeling each examinee based on their starting ability level. That is, for each simulated student, there is an assigned ability level prior to beginning the simulated assessment administration.

The weighted penalty model (WPM; Shin et al., 2009) for content balancing is the algorithm the Pearson CAT engine uses for CAT simulations. The WPM selects each ordinal item by balancing test specification and item information, and the conditional randomesque method (CRM; Shin and Chien, 2017) is used to control item exposure. The CAT is configured to control for such things as starting person estimates (student theta), the range of theta (IRT ability), item exposure control, conditional ranges of theta, group sizes, and the balance of weights assigned to item information and the blueprint constraints. The metadata and statistical parameters for the items in the CAT item banks are summarized and coded into the configuration file, as are the target item count ranges for each blueprint category (content strands, standards, and item types).

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

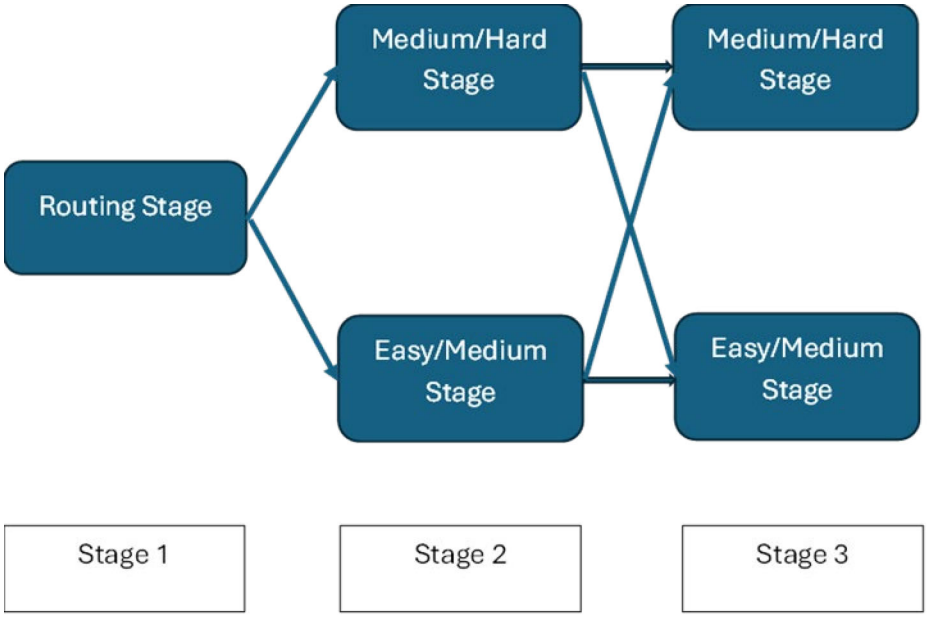
The CMAS operational item banks were used for these mathematics simulations. These banks have been built and maintained to support the current linear based testing program. Otherwise, the simulator was configured to match the blueprints for each respective math test described in the previous section. Simulations were carried out for the full-length form, the shortened form, and unconstrained for each grade. “True” ability levels for the simulated examinees were the actual student theta abilities of the 2025 CMAS administration (described above). That is, the overall ability scores for each math grade and subject tests were input into the simulator directly. The stopping rule for the full-length and shortened form simulations was to meet the blueprint and test length. For the unconstrained simulation, the stopping rule was set to meet a measurement precision level of 0.20. In all cases the starting ability level estimation was set to 0. Lastly, CMAS tests contain items that are scored by human scorers. In a live operational administration, these items would not contribute to estimation of student ability, as those item types would be scored after the administration. However, since this is not an issue for simulated cases, these items did contribute to the student estimates as if being scored automatically.

2.3. Multistage Adaptive Testing Simulation (ELA)

For the ELA multistage adaptive testing (MST) simulations, the same Pearson CAT engine was also used. Unlike item-level adaptive, as was used for the Math simulations, MST results in effectively presenting all examinees the same overall number of items that meet the full operational blueprint. The adaptivity comes into play in that for a given stage, item sets are grouped so that they comprise sets of differing difficulty level. Similar to item-level CAT, a student is routed to a set that best aligns with their ability level. That is, as with item-level CAT estimation based on item-response theory, MST estimates ability in the same way, based on all items in the given stage as opposed to on an item-by-item basis. That estimate is compared to a predetermined threshold value that results in the student being given a set of items that are easier or more difficult in the next stage.

For the CMAS MST simulation, the design was to reflect each MST as comprised of the full and reduced-length blueprints. Each simulated student would receive the same number of items matched to the blueprints presented in Table 2.4 – Table 2.9. Additionally, the operational passage sets (set of items referencing a given passage) served to identify what parts of each respective blueprint would comprise each stage. For both grades, one operational passage set lent to serving as the initial routing stage, as the average difficulty level was close to the middle of the ability distribution. A standard MST design that could be set up for simulation was a design that has three stages where the first is the routing stage and the second and third stages each have two item sets – one easier and one harder. This is referred to as a 1-2-2 MST and is depicted in Figure 2.1. Lastly, using the passage sets as driving the structure of the stage design meant there were roughly 10 items in the first two stages, and roughly six items in the third for each full-length MST simulation.

Figure 2.1. CMAS ELA Multistage Configuration



The final simulation design was attained by examining operational passage sets in each respective item bank and matching on several metadata variables that are most important during current operational test construction activities. Beyond this, eligible passage sets were identified by the average item difficulty and targeted for either the easier or harder stage. In this fashion, passage sets equivalent to blueprint but easier and harder were identified and used to carry out the respective simulations for Grades 4 and 6 full and reduced length tests. For each stage, the threshold routing criteria was set to 0.0 (roughly the middle of the ability distribution). And like the Math simulations, “true” ability was set by importing the operational CMAS scores (abilities) for examinees from the spring 2025 administration.

3. Results

3.1. Test Shortening

This section presents the results for the reduced-length assessments in comparison to the full-length spring 2025 administration. The results include scale score summary statistics, overall performance level agreement, standard performance level agreement, correlations, test characteristic curves (TCCs), information functions, CSEM curves, reliability indices, and timing analyses.

3.1.1. Scale Score Summary Statistics

The overall and reduced length scale score and scale score CSEMs were calculated based on all operational items. Table 3.1 presents summary statistics (count, mean, standard deviation, minimum, and maximum) for the full and reduced length scale scores and CSEM. The average scale scores were similar for the reduced and full-length test forms for both mathematics and ELA. The average scale score difference was negligible, and the average conditional standard errors were slightly higher for the reduced length tests compared to the full length for both content areas.

Table 3.11. Summary Statistics for Full and Adjusted Scores

Assessment		Count	Mean	SD	Min.	Max.
Mathematics 4	Full Scale Score	52,493	737.15	34.26	650	850
	Reduced Scale Score	52,493	737.18	35.12	650	850
	Full Scale Score CSEM	52,493	8.67	3.04	7	72
	Reduced Scale Score CSEM	52,493	10.64	4.76	8	87
Mathematics 6	Full Scale Score	52,221	732.67	33.60	650	850
	Reduced Scale Score	52,221	732.57	34.26	650	850
	Full Scale Score CSEM	52,221	9.43	4.40	7	69
	Reduced Scale Score CSEM	52,221	11.30	7.12	8	84
ELA 4	Full Scale Score	52,252	741.33	37.07	650	850
	Reduced Scale Score	52,252	740.37	39.99	650	850
	Full Scale Score CSEM	52,252	12.10	2.50	11	60
	Reduced Scale Score CSEM	52,252	16.22	5.97	14	74
ELA 6	Full Scale Score	52,414	742.75	33.66	650	850
	Reduced Scale Score	52,414	742.64	34.21	650	850
	Full Scale Score CSEM	52,414	9.40	1.99	8	54
	Reduced Scale Score CSEM	52,414	10.37	2.73	9	66

3.1.2. Performance Level Agreement

Table 3.2 presents the percent of students assigned the exact same performance level for both the full and the reduced assessments, as well as the percent of students assigned to different performance levels between the full and reduced assessments. If the reduced performance levels were a higher ability level compared to the full performance level the number and percent of students are listed as “Higher Level for Reduced.” If the reduced performance levels were a lower ability level compared to the full performance level the number and percent of students are listed as “Lower Level for Reduced.”

For mathematics, the percent of exact agreement in the overall performance level designation between the full-length assessment and the adjusted assessment was roughly 86% for both grades. The percent of students in the higher level or the lower level was roughly 7%. For ELA, the percent of exact agreement in the overall performance level designation between the full and the reduced length assessment was roughly 14% lower for grade 4 than grade 6. The percent of students in the higher level or the lower level was roughly 7%. The percent of exact agreement in the overall performance level designation between the full-length assessment and the assessment was higher for grade 6 than grade 4 by 14%.

Table 3.12. Performance Level Agreement

Assessment	Exact Agreement	Higher Level for Reduced	Lower Level for Reduced
Mathematics 4	86.66%	6.79%	6.55%
Mathematics 6	86.12%	6.89%	6.99%
ELA 4	75.77%	12.05%	12.18%
ELA 6	90.12%	4.67%	5.21%

Table 3.3 – Table 3.6 present the number and percent of students by the full length performance level designation and the reduced length performance level designation for the overall tests. The values bolded in the tables represent exact agreement. For all standards and the overall test, if the performance level designation was not exact, the difference was always within an adjacent performance level. This was not true for ELA grade 4, where there was one instance at Level 2 where there were 10 cases where agreement rates were non-adjacent.

Table 3.13. Overall Performance Level Percent Agreement—Grade 4 Mathematics

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**7,133 (13.59%)**	954 (1.82%)	0 (0%)	0 (0%)	0 (0%)	8,087 (15.41%)
2025 L2	1,050 (2.00%)	**9,845 (18.75%)**	965 (1.84%)	0 (0%)	0 (0%)	11,860 (22.59%)
2025 L3	0 (0%)	1,029 (1.96%)	**10,898 (20.76%)**	975 (1.86%)	0 (0%)	12,902 (24.58%)
2025 L4	0 (0%)	0 (0%)	986 (1.88%)	**15,820 (30.14%)**	669 (1.27%)	17,475 (33.29%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	373 (0.71%)	**1,796 (3.42%)**	2,169 (4.13%)
Total	8,183 (15.59%)	11,828 (22.53%)	12,849 (24.48%)	17,168 (32.71%)	2,465 (4.70%)	**52,493 (100%)**

Table 3.14. Overall Performance Level Percent Agreement—Grade 6 Mathematics

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**8,081 (15.47%)**	1,013 (1.94%)	0 (0%)	0 (0%)	0 (0%)	9,094 (17.41%)
2025 L2	1,086 (2.08%)	**10,028 (19.20%)**	1,094 (2.09%)	0 (0%)	0 (0%)	12,208 (23.38%)
2025 L3	0 (0%)	1,197 (2.29%)	**12,114 (23.20%)**	1,038 (1.99%)	0 (0%)	14,349 (27.48%)
2025 L4	0 (0%)	0 (0%)	991 (1.90%)	**12,542 (24.02%)**	453 (0.87%)	13,986 (26.78%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	376 (0.72%)	**2,208 (4.23%)**	2,584 (4.95%)
Total	9,167 (17.55%)	12,238 (23.44%)	14,199 (27.19%)	13,956 (26.72%)	2,661 (5.10%)	**52,221 (100%)**

Table 3.15. Overall Performance Level Percent Agreement—Grade 4 ELA

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**6,542 (12.52%)**	837 (1.60%)	0 (0%)	0 (0%)	0 (0%)	7,379 (14.12%)
2025 L2	1,834 (3.51%)	**5,950 (11.39%)**	2,020 (3.87%)	10 (0.02%)	0 (0%)	9,814 (18.78%)
2025 L3	17 (0.03%)	1,722 (3.30%)	**9,083 (17.38%)**	2,356 (4.51%)	0 (0%)	13,178 (25.22%)
2025 L4	0 (0%)	6 (0.01%)	1,991 (3.81%)	**13,735 (26.29%)**	1,071 (2.05%)	16,803 (32.16%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	796 (1.52%)	**4,282 (8.19%)**	5,078 (9.72%)
Total	8,393 (16.06%)	8,515 (16.30%)	13,094 (25.06%)	16,897 (32.34%)	5,353 (10.24%)	52,252 **(100%)**

Table 3.16. Overall Performance Level Percent Agreement—Grade 6 ELA

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**5,239 (10.00%)**	712 (1.36%)	0 (0%)	0 (0%)	0 (0%)	5,951 (11.35%)
2025 L2	876 (1.67%)	**8,561 (16.33%)**	575 (1.10%)	0 (0%)	0 (0%)	10,012 (19.10%)
2025 L3	0 (0%)	805 (1.54%)	**11,867 (22.64%)**	684 (1.30%)	0 (0%)	13,356 (25.48%)
2025 L4	0 (0%)	0 (0%)	759 (1.45%)	**17,772 (33.91%)**	474 (0.90%)	19,005 (36.26%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	291 (0.56%)	**3,799 (7.25%)**	4,090 (7.80%)
Total	6,115 (11.67%)	10,078 (19.23%)	13,201 (25.19%)	18,747 (35.77%)	4,273 (8.15%)	**52,414 (100%)**

3.1.3. Correlations and Scale Score Differences

The correlation between the full-scale scores and adjusted scale scores for the overall. Correlations were .985 for both grades in mathematics. For ELA, correlations were high but lower for grade 4 (0.96) than for grade 6 (0.99). Table 3.7, Table 3.8, and Table 3.9 present correlations between subclaim and overall test scores for the full and reduced length tests.

Table 3.17. Pearson Correlations between Full and Reduced Length Tests

Assessment	Full and Reduced Scale Scores Correlation
Mathematics 4	0.985
Mathematics 6	0.985
ELA 4	0.959
ELA 6	0.991

Table 3.18. Pearson Correlations between Subclaims and Overall Scores for the Full and Reduced Length Tests—Mathematics

Assessment	Blueprint	Subclaim	Subclaim B	Subclaim C	Subclaim D	Full Test
Mathematics 4	Full	A	0.778	0.817	0.777	0.967
		B		0.718	0.685	0.849
		C			0.729	0.903
		D				0.864
	Reduced	A	0.738	0.728	0.697	0.939
		B		0.696	0.650	0.863
		C			0.685	0.861
		D				0.839
Mathematics 6	Full	A	0.778	0.783	0.774	0.944
		B		0.753	0.729	0.884
		C			0.784	0.905
		D				0.891
	Reduced	A	0.737	0.718	0.693	0.937
		B		0.673	0.653	0.857
		C			0.691	0.853
		D				0.846

Table 3.19. Pearson Correlations between Subclaims and Overall Scores for the Full and Reduced Length Tests—ELA

Assessment	Blueprint	Subclaim	RL	RV	WE	WKL	Full Test
ELA 4	Full	RL	0.653	0.660	0.677	0.594	0.890
		RI		0.618	0.518	0.503	0.826
		RV			0.543	0.506	0.783
		WE				0.783	0.846
		WKL					0.766
	Reduced	RL	0.507	0.480	0.592	0.489	0.809
		RI		0.477	0.456	0.437	0.745
		RV			0.439	0.413	0.646
		WE				0.786	0.877
		WKL					0.780
ELA 6	Full	RL	0.754	0.673	0.623	0.592	0.855
		RI		0.724	0.725	0.683	0.920

Assessment	Blueprint	Subclaim	RL	RV	WE	WKL	Full Test
		RV			0.591	0.560	0.802
		WE				0.866	0.887
		WKL					0.832
	Reduced	RL	0.736	0.598	0.616	0.586	0.842
		RI		0.643	0.732	0.689	0.909
		RV			0.536	0.508	0.719
		WE				0.867	0.904
		WKL					0.847

Figure 3.1 – Figure 3.4 present the scatterplot of full scale scores versus reduced test scale scores and the frequency distribution of the differences in scale scores between the full and reduced tests. Figure 3.5 – Figure 3.8 present the histograms of the full versus reduced length scale score differences.

Figure 3.2. Overall Full vs. Reduced Length Tests—Grade 4 Math

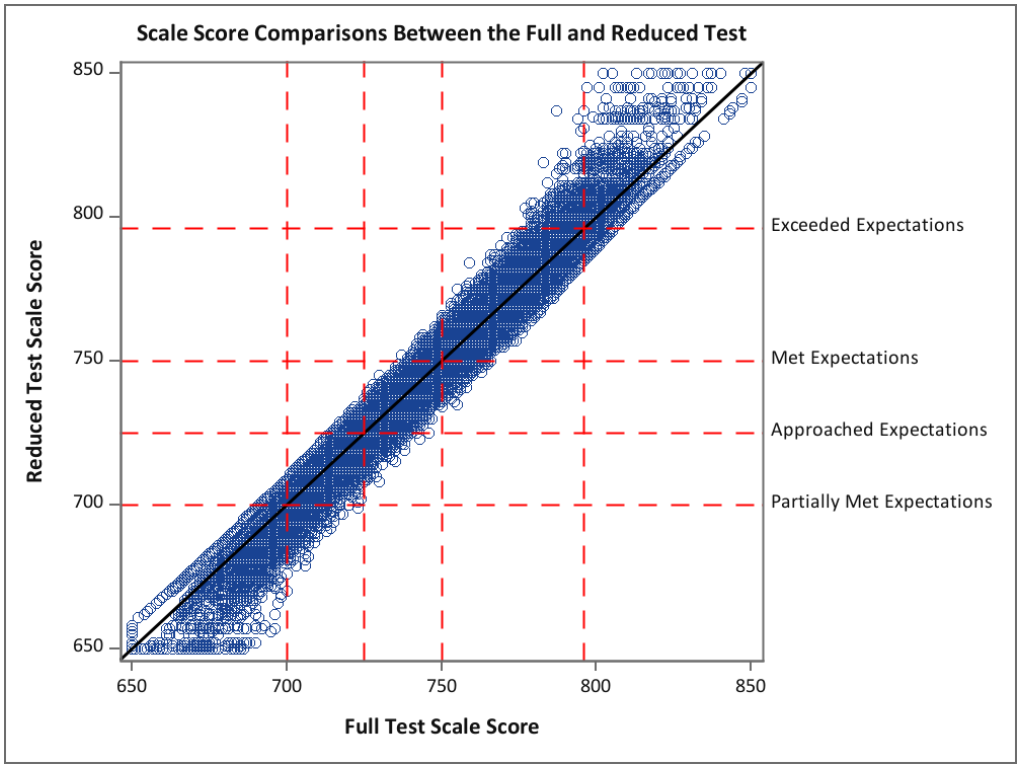


Figure 3.3. Overall Full vs. Reduced Length Tests—Grade 6 Math

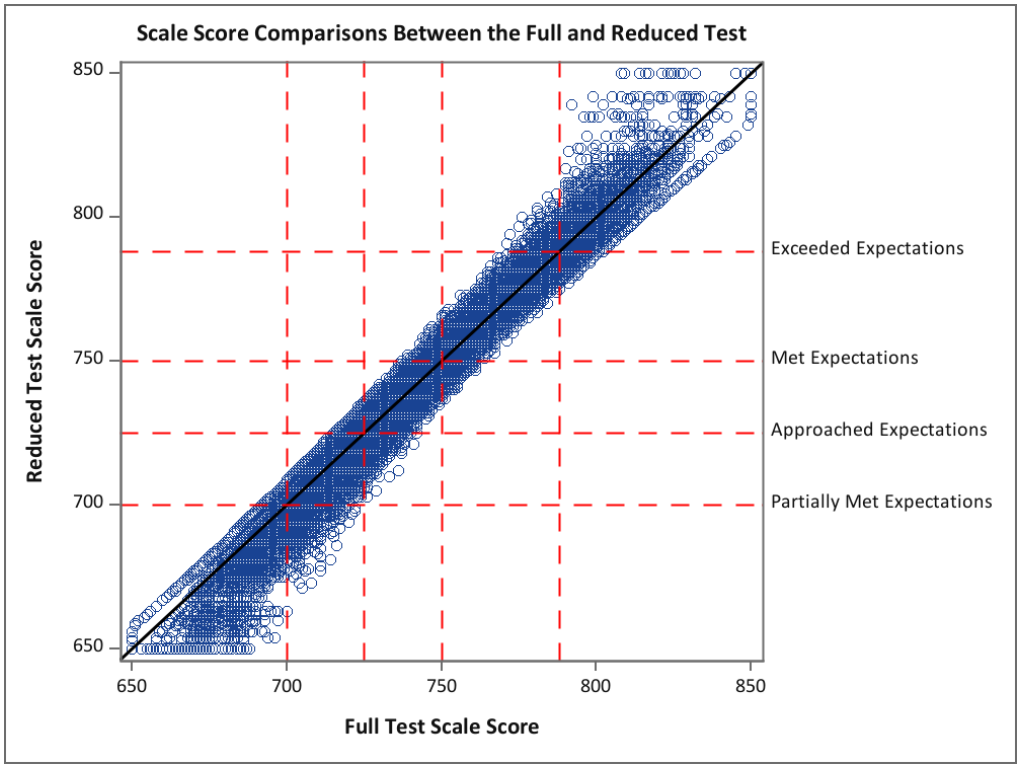


Figure 3.4. Overall Full vs. Reduced Length Tests—Grade 4 ELA

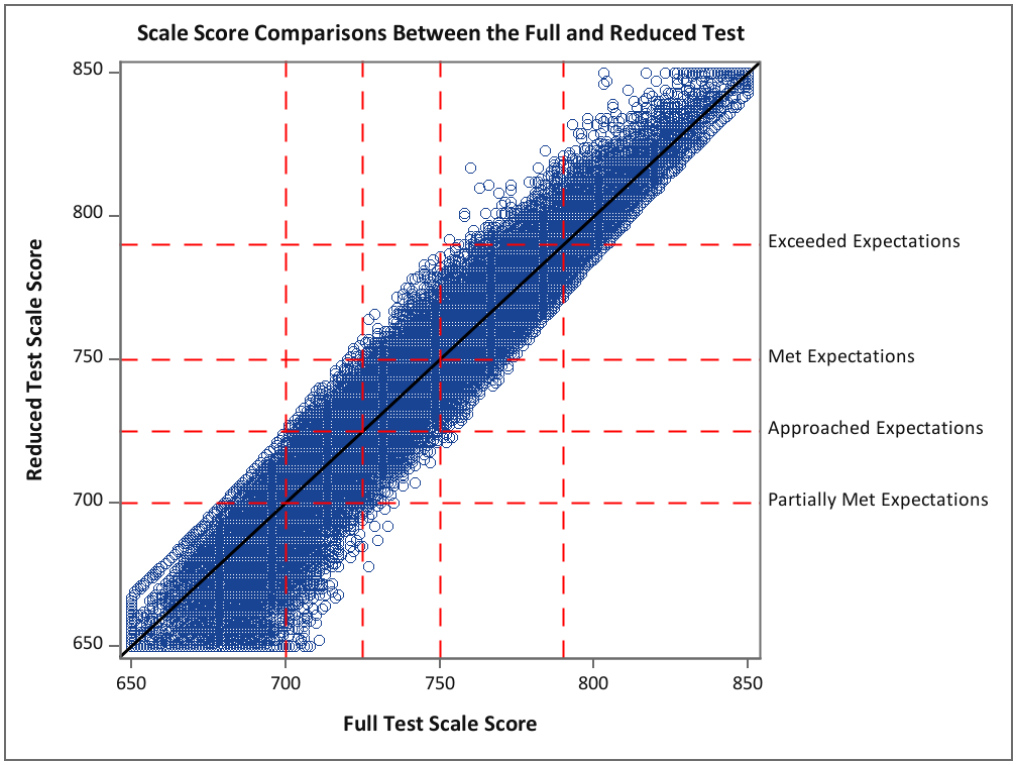


Figure 3.5. Overall Full vs. Reduced Length Tests—Grade 6 ELA

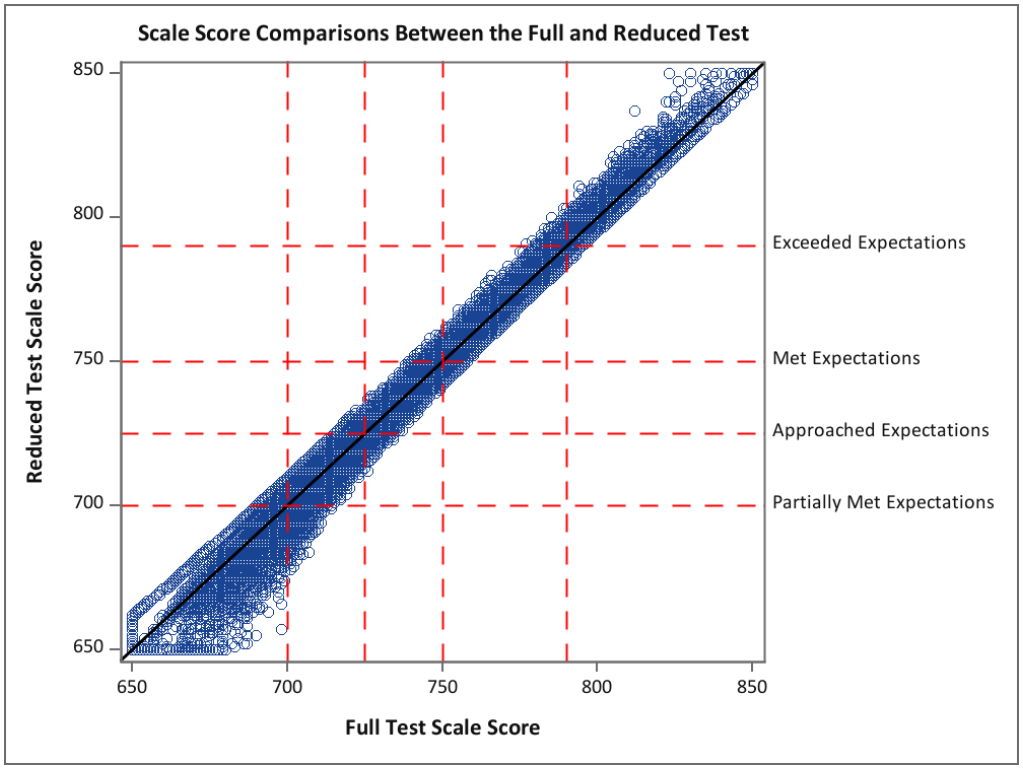


Figure 3.6. Overall Scale Score Differences Between Adjusted and Full Length Tests—Grade 4 Math

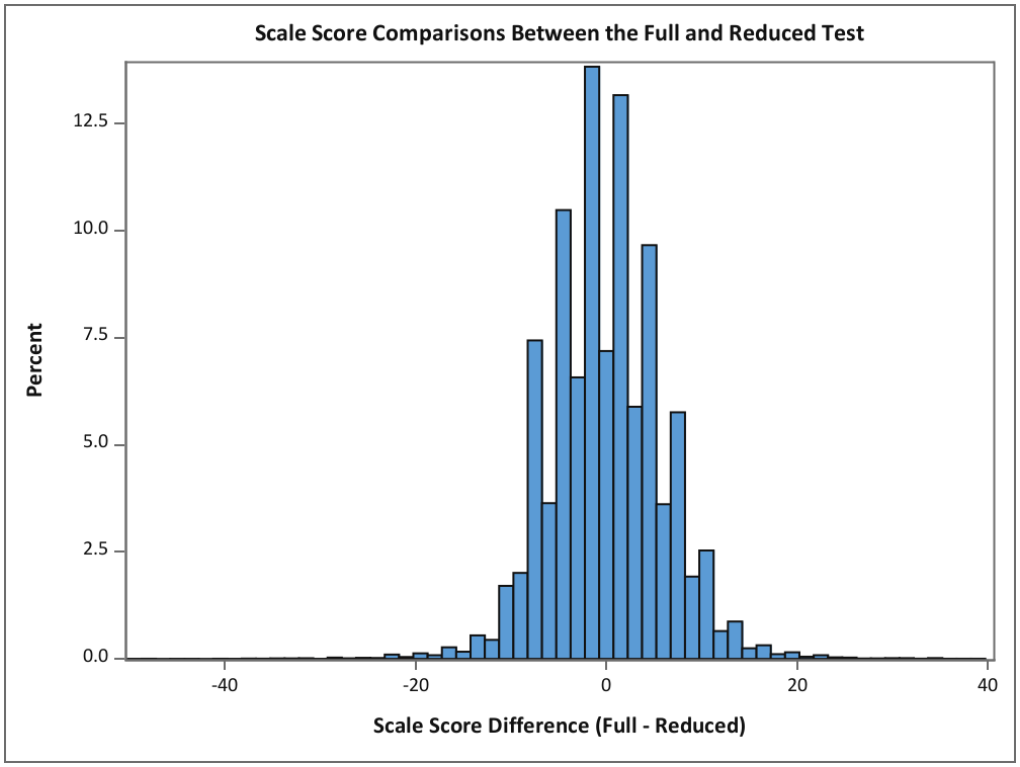


Figure 3.7. Overall Scale Score Differences Between Adjusted and Full Length Tests—Grade 6 Math

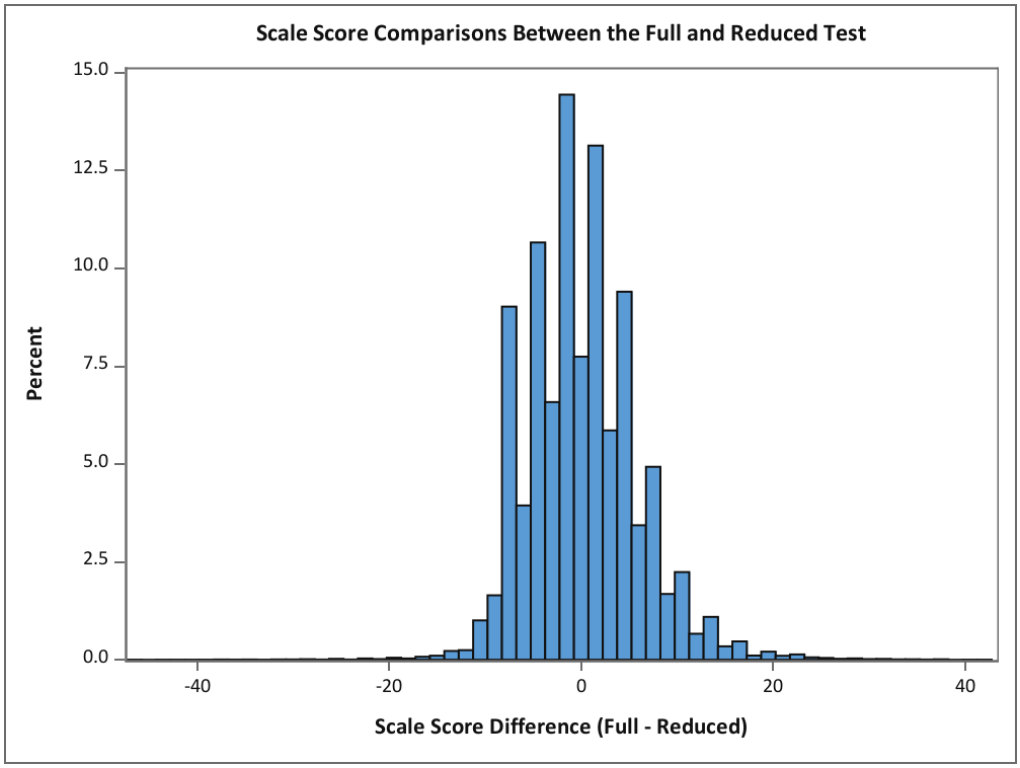


Figure 3.8. Overall Scale Score Differences Between Reduced and Full Length Tests—Grade 4 ELA

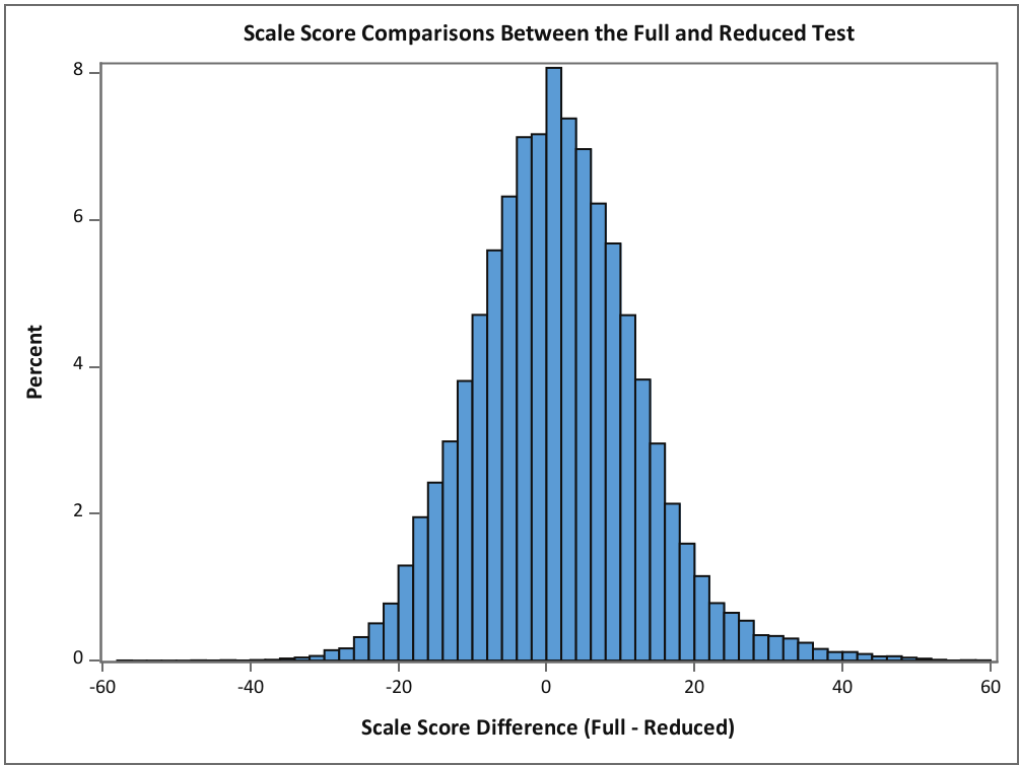
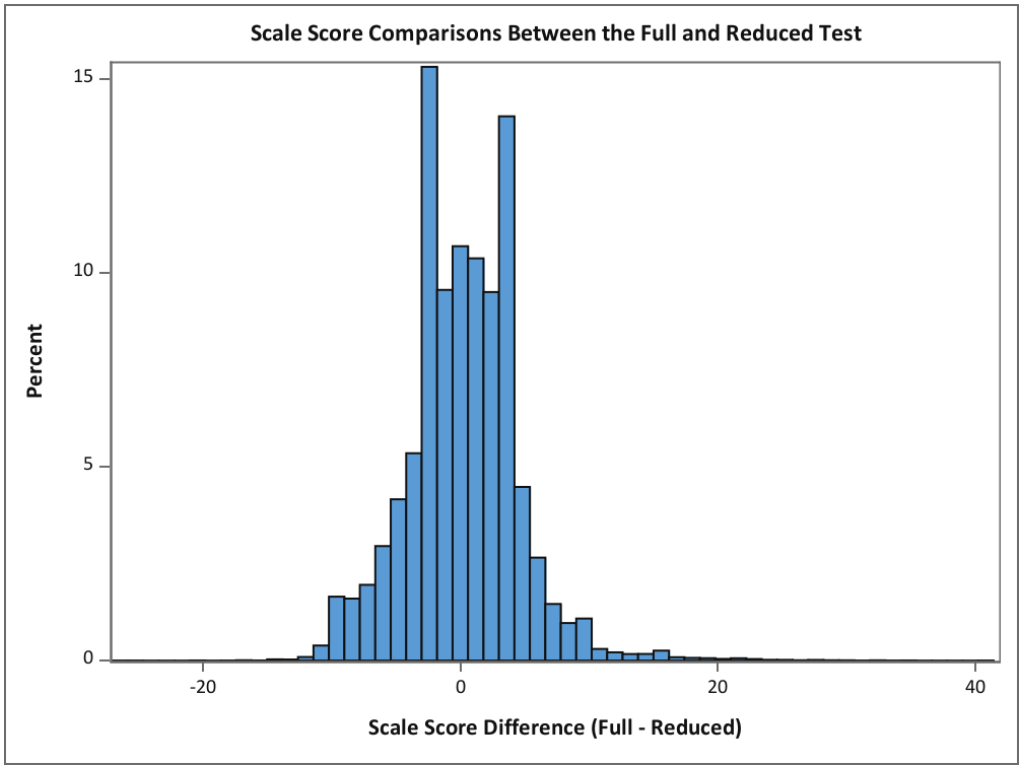


Figure 3.9. Overall Scale Score Differences Between Reduced and Full Length Tests—Grade 6 ELA



3.1.4. Summative Test Characteristic, Test Information, and CSEM Curves

The TCCs were generated for the raw score to theta scale for the overall theta scale. Due to the difference in raw score total and the number of operational items across the full and reduced forms, TCCs are

provided based on the percent of the total maximum possible score points. Figure 3.9 – Figure 3.12 present the TCCs for the full and reduced length tests. In general, the TCCs for the adjusted test forms are similar to the full test form TCCs.

Figure 3.13 – Figure 3.16 present the test information curves for the full and reduced length tests, and Figure 3.17 – Figure 3.20 present the CSEM curves for the full and reduced length tests. Each of these illustrates the impact of the test reductions on measurement precision, where the reduced length forms are less precise and of lower information.

Figure 3.10. Overall Percentage TCCs for Full Reduced Length Tests—Grade 4 Math

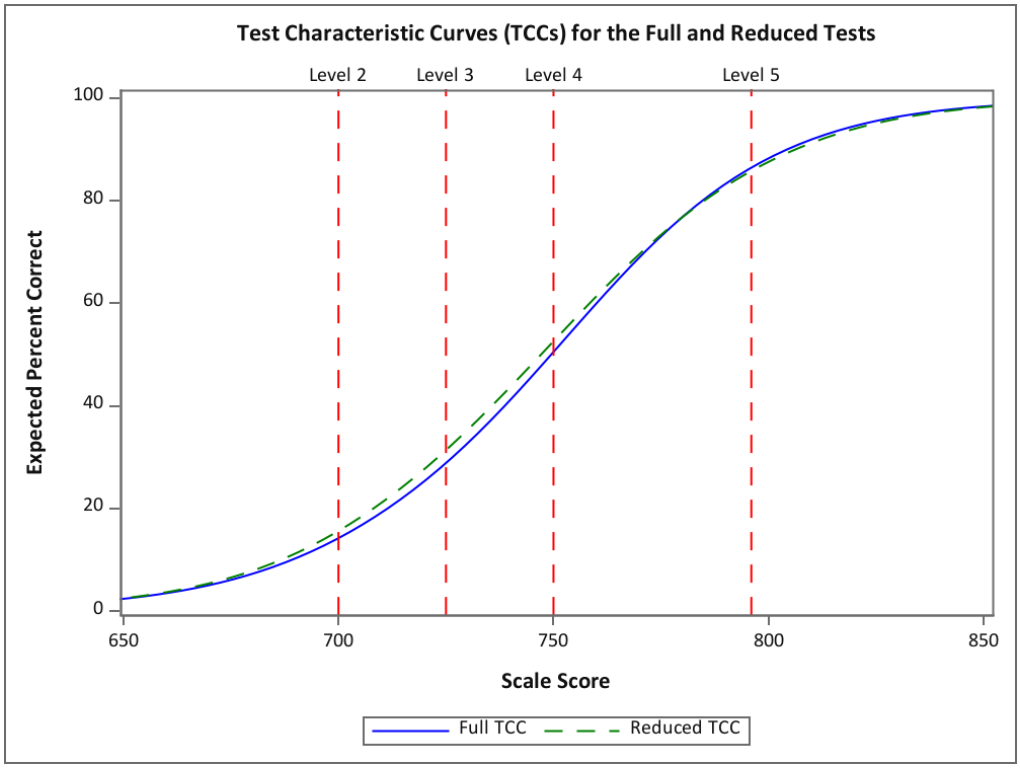


Figure 3.11. Overall Percentage TCCs for Full Reduced Length Tests—Grade 6 Math

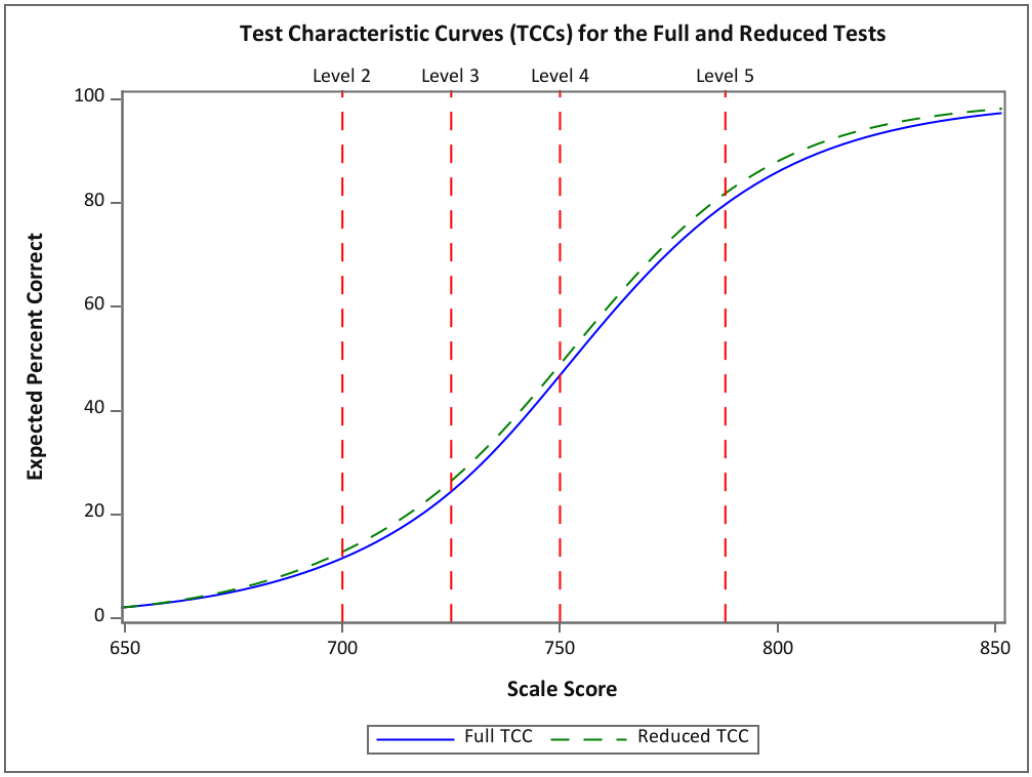


Figure 3.12. Overall Percentage TCCs for Full Reduced Length Tests—Grade 4 ELA

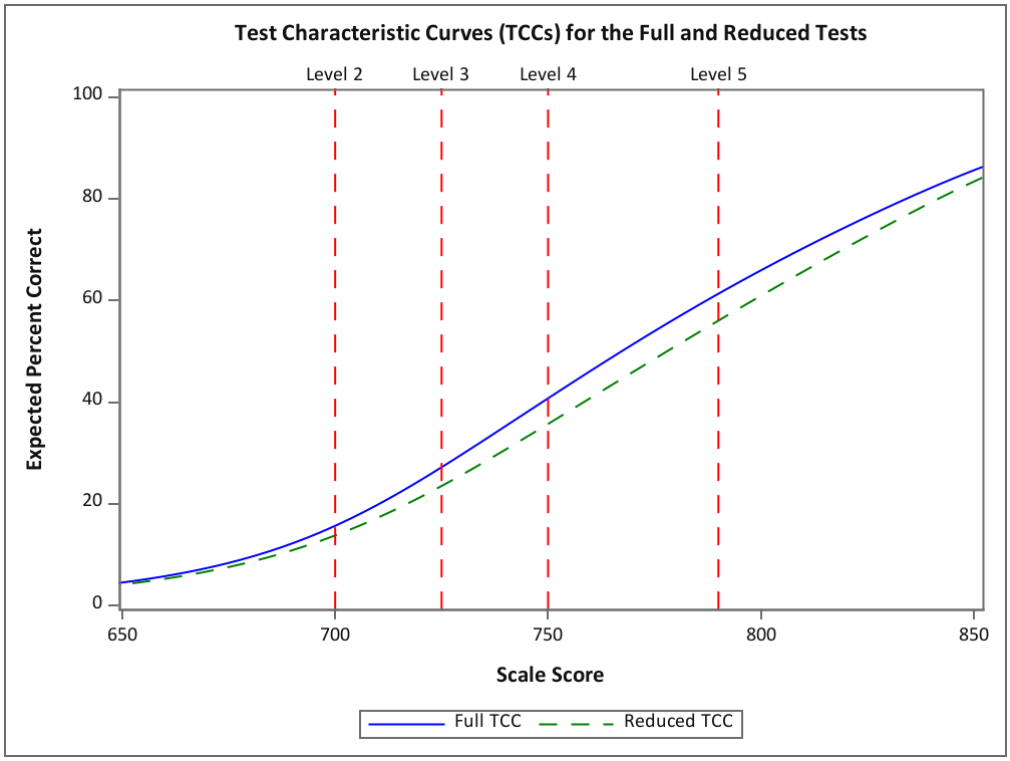


Figure 3.13. Overall Percentage TCCs for Full Reduced Length Tests—Grade 6 ELA

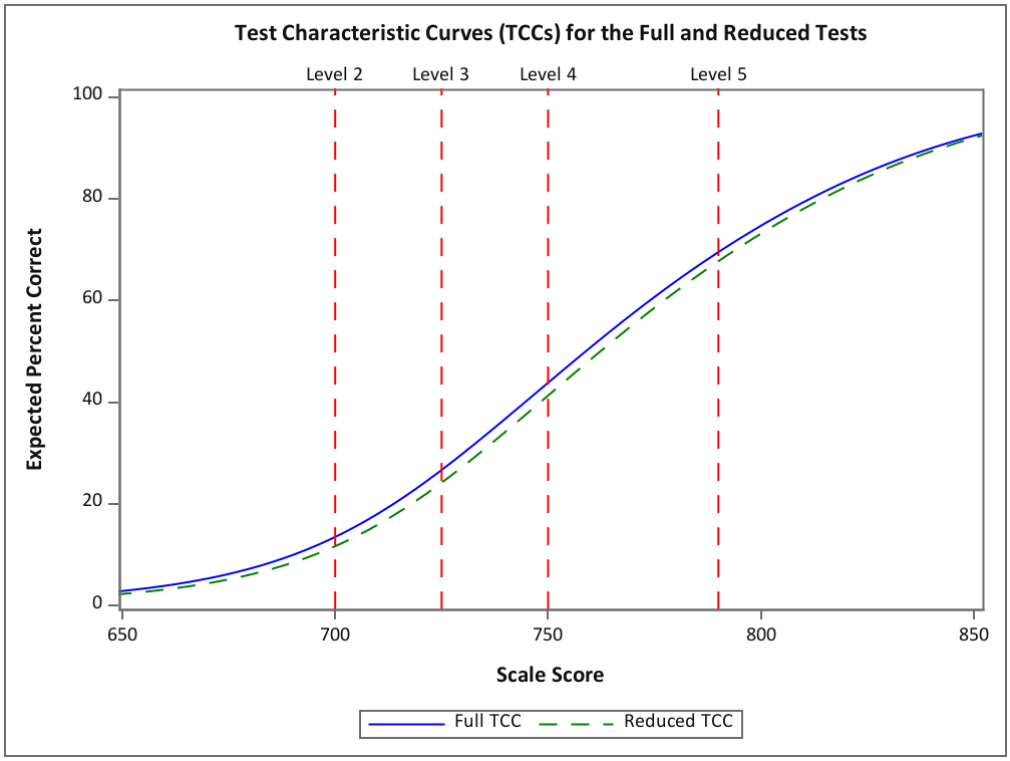


Figure 3.14. Overall Information Curves for Full and Reduced Length Tests—Grade 4 Math

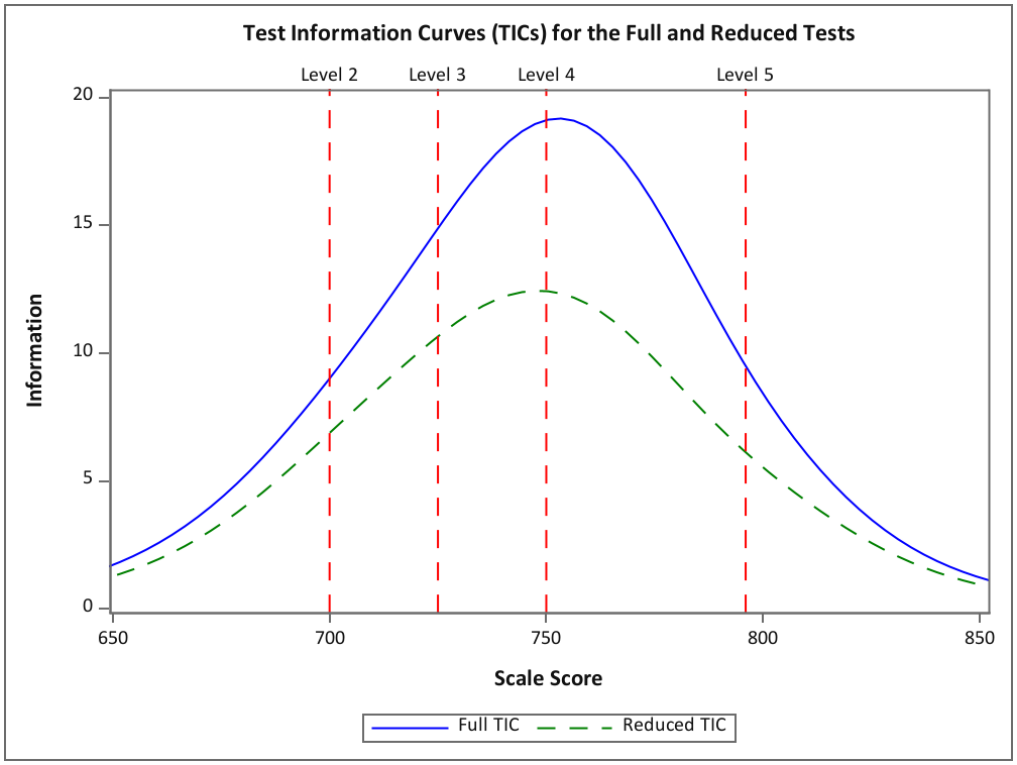


Figure 3.15. Overall Information Curves for Full and Reduced Length Tests—Grade 6 Math

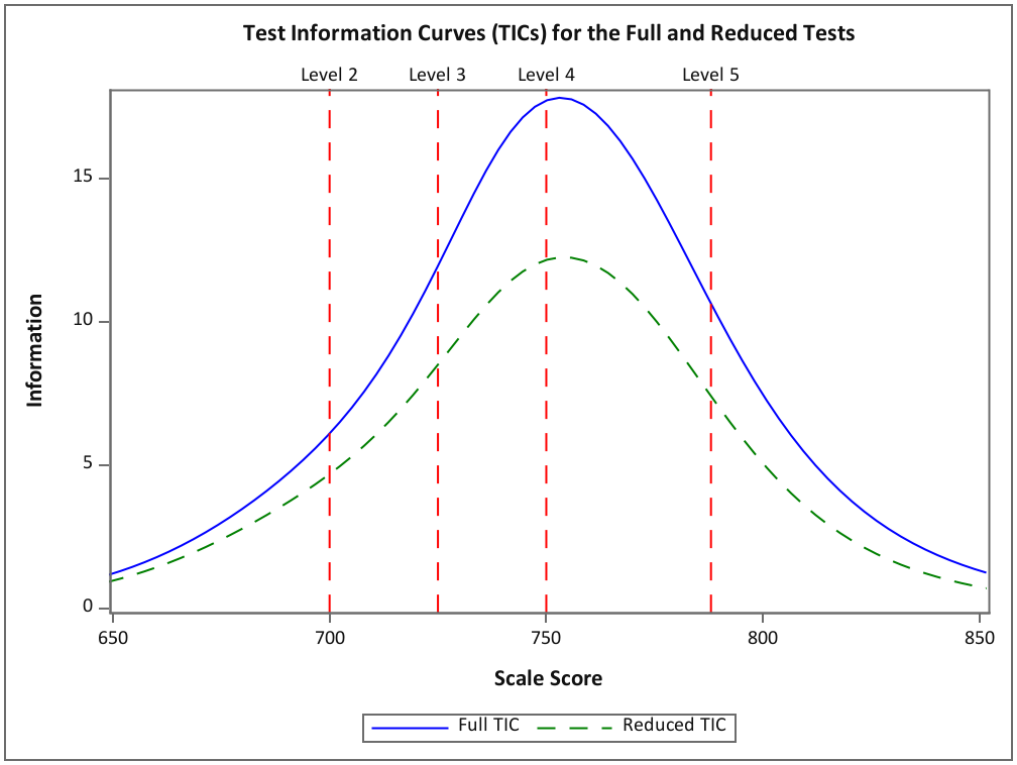


Figure 3.16. Overall Information Curves for Full and Reduced Length Tests—Grade 4 ELA

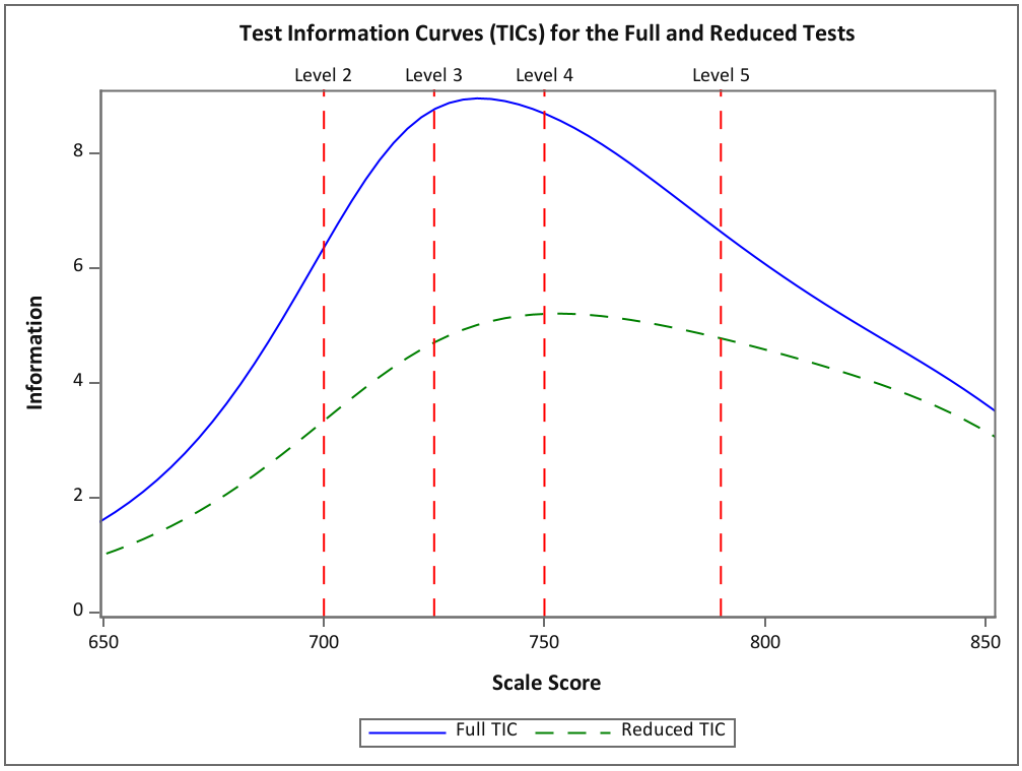


Figure 3.17. Overall Information Curves for Full and Reduced Length Tests—Grade 6 ELA

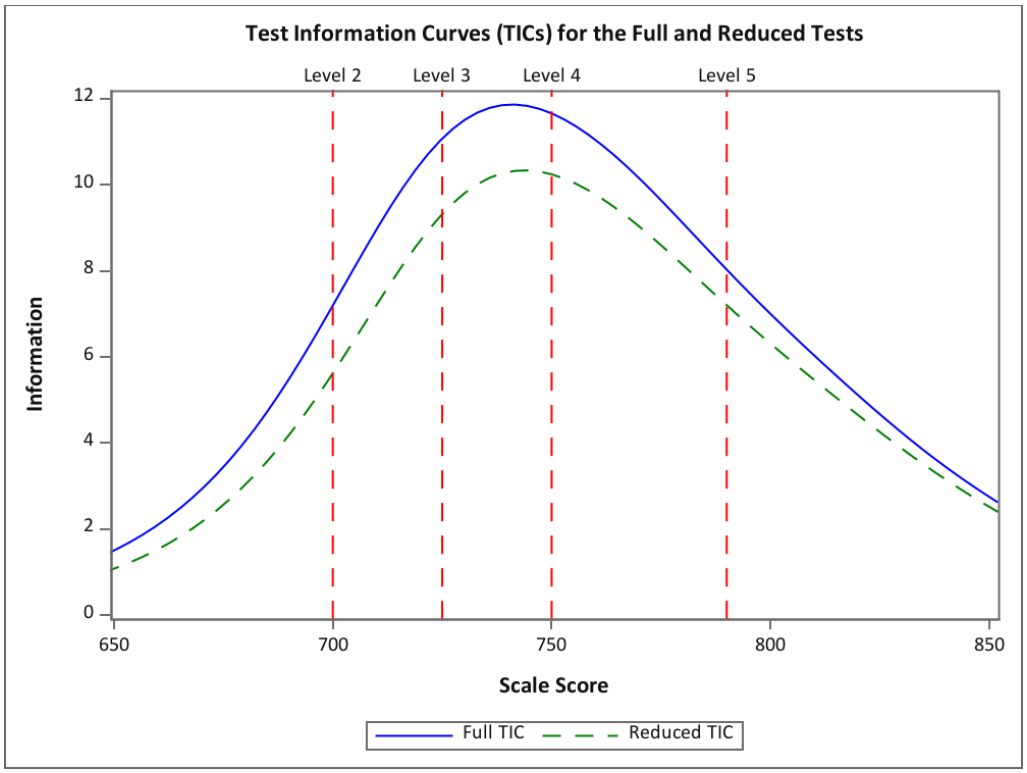


Figure 3.18. Overall CSEM Curves for Full and Reduced Length Tests—Grade 4 Math

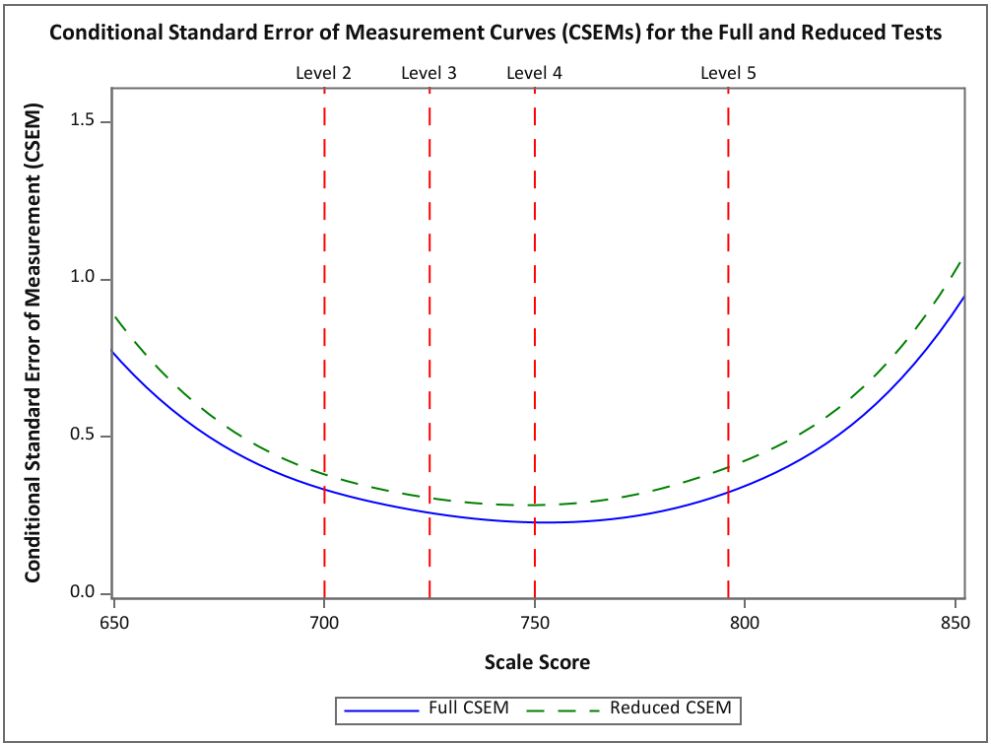


Figure 3.19. Overall CSEM Curves for Full and Reduced Length Tests—Grade 6 Math

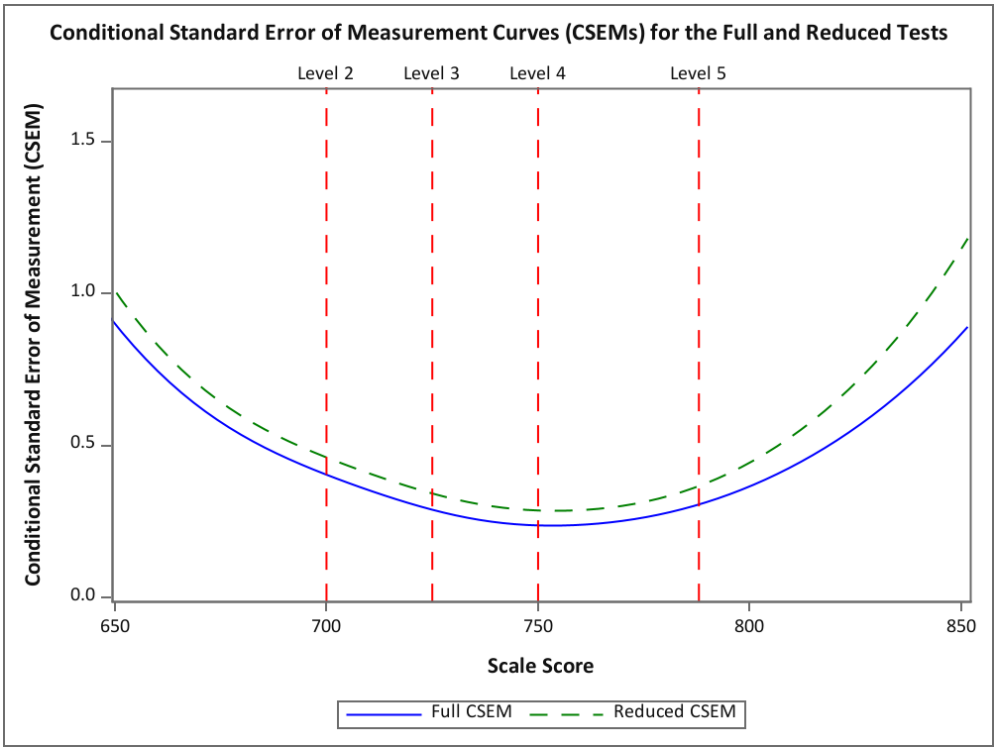


Figure 3.20. Overall CSEM Curves for Full and Reduced Length Tests—Grade 4 ELA

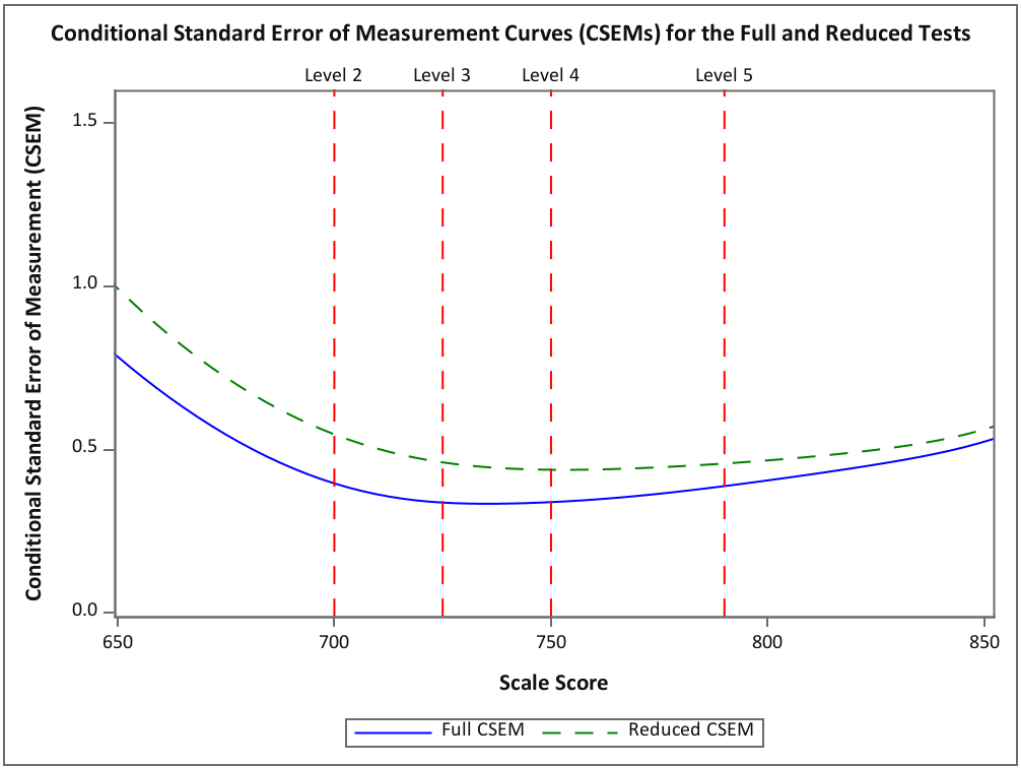
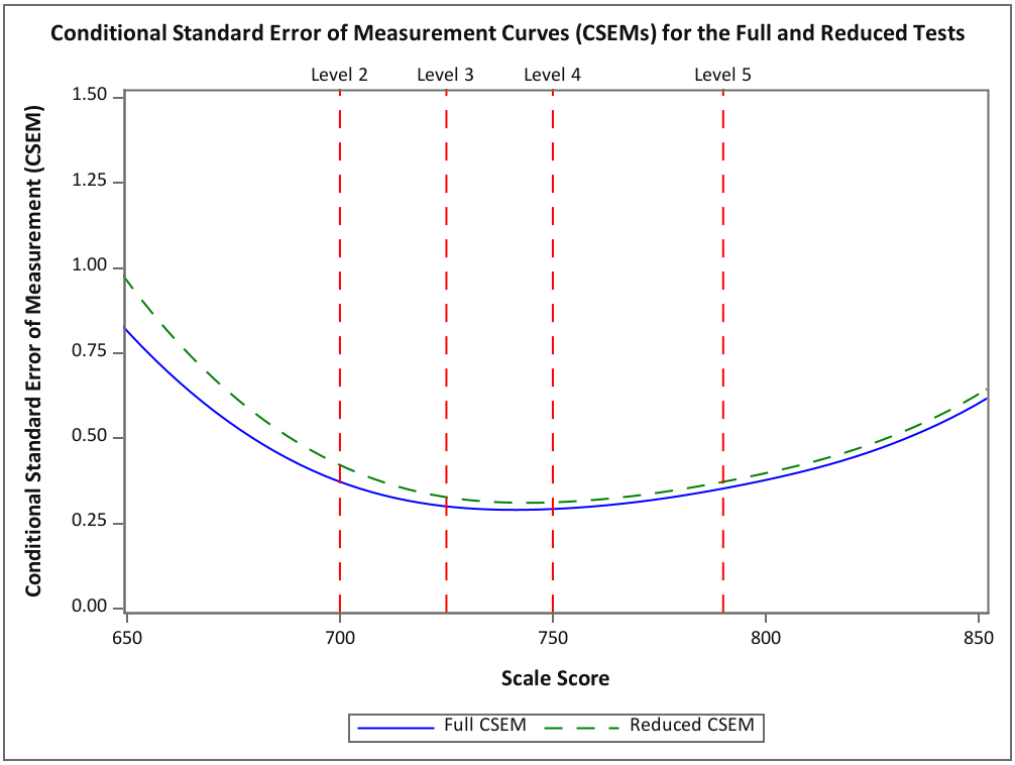


Figure 3.21. Overall CSEM Curves for Full and Reduced Length Tests—Grade 6 ELA



3.1.5. Reliability Indices

In addition to the standard error comparisons in the previous figures, the following tables present additional measurement precision indicators for the full and reduced length tests. Coefficient alpha is

referred to as a measure of internal consistency. Decision consistency and accuracy reflect reliability based on classification of students into the five respective performance levels and are presented based on the specific cut scores (scale scores that mark the transition from one performance level to the next as described previously). Cohen's Kappa and the probability of chance classifications are test-level indicators. For more detailed descriptions of these calculations, the reader is encouraged to consult the CMAS technical report (https://ed.cde.state.co.us/assessment/cmas-coalt-techreport).

Table 3.10 and Table 3.11 present these reliability indicators based on the full length and reduced tests. As with previous indicators, the results show lower reliability for the reduced length tests for both mathematics and ELA. Table 3.12 –Table 3.14 present the classification accuracy and consistency results.

Table 3.20. Coefficient Alpha for the Full and Reduced Length Tests—Mathematics

Assessment	Scenario	Overall	Subclaim A	Subclaim B	Subclaim C	Subclaim D
Mathematics 4	Full	0.93	0.90	0.70	0.71	0.57
	Reduced	0.89	0.85	0.70	0.57	*
Mathematics 6	Full	0.92	0.84	0.74	0.75	0.65
	Reduced	0.88	0.80	0.66	0.67	*

*Coefficient alpha for Subclaim D in mathematics could not be calculated because under the reduced blueprint, it consists of only one item.

Table 3.21. Coefficient Alpha for the Full and Reduced Length Tests—ELA

Assessment	Scenario	Overall	RL	RI	RV	WE	WKL
ELA 4	Full	0.88	0.70	0.72	0.61	0.65	0.76
	Reduced	0.83	0.56	0.55	0.38	0.65	0.76
ELA 6	Full	0.90	0.75	0.81	0.66	0.76	0.81
	Reduced	0.89	0.73	0.79	0.52	0.76	0.81

Table 3.22. Classification Accuracy and Consistency for the Full and Reduced Length Tests

Assessment	Scenario	Accuracy	Consistency	PChance	Kappa
Mathematics 4	Full	0.77	0.69	0.25	0.59
	Reduced	0.71	0.61	0.24	0.49
Mathematics 6	Full	0.75	0.65	0.23	0.55
	Reduced	0.69	0.59	0.23	0.46
ELA 4	Full	0.68	0.58	0.23	0.45
	Reduced	0.62	0.50	0.23	0.36
ELA 6	Full	0.73	0.63	0.25	0.51
	Reduced	0.71	0.61	0.25	0.48

Table 3.23. Classification Accuracy at the Cut Scores for the Full and Reduced Length Tests

Assessment	Scenario	*Partially Met Expectations* Cut	*Approached Expectations* Cut	*Met Expectations* Cut	*Exceeded Expectations* Cut
Mathematics 4	Full	0.95	0.92	0.92	0.95
	Reduced	0.94	0.90	0.90	0.94
Mathematics 6	Full	0.94	0.91	0.92	0.94
	Reduced	0.93	0.89	0.90	0.94
ELA 4	Full	0.94	0.90	0.89	0.88
	Reduced	0.93	0.89	0.86	0.87
ELA 6	Full	0.95	0.92	0.90	0.91
	Reduced	0.95	0.91	0.90	0.90

Table 3.24. Classification Consistency at the Cut Scores for the Full and Reduced Length Tests

Assessment	Scenario	*Partially Met Expectations* Cut	*Approached Expectations* Cut	*Met Expectations* Cut	*Exceeded Expectations* Cut
Mathematics 4	Full	0.93	0.89	0.89	0.94
	Reduced	0.91	0.86	0.86	0.93
Mathematics 6	Full	0.92	0.88	0.88	0.93
	Reduced	0.90	0.85	0.86	0.92
ELA 4	Full	0.92	0.86	0.85	0.86
	Reduced	0.93	0.89	0.86	0.87
ELA 6	Full	0.94	0.88	0.86	0.89
	Reduced	0.93	0.88	0.86	0.88

3.1.6. Timing Analyses

Lastly, timing analyses were carried out from actual student time-on-item data from the spring 2025 administrations. For the full-length test this reflects total testing time for each actual examinee for their testing experience. Since the full-length calculations included tryout (field test) items as depicting total actual testing time, these were also included for the reduced-length, to allow for direct comparisons.

Table 3.15 presents descriptive statistics for the full and reduced length tests. Testing time did not vary substantively comparing the full-length and reduced-length tests. Figure 3.21 – Figure 3.28 present histograms of the timing data for the full and reduced length tests. These break the examinee populations into thirds based on their scale scores. In each case the distributions are similar across the full and reduced length tests (lower performing third taking the shortest amount of time).

Table 3.25. Timing Descriptive Statistics for the Full and Reduced Length Tests

Assessment	Blueprint	10th Percentile	Median	Mean	90th Percentile
Mathematics 4	Full	83	142	139	191
	Reduced	57	100	98	135
Mathematics 6	Full	67	115	116	168
	Reduced	49	84	86	125
ELA 4	Full	99	180	179	263
	Reduced	80	152	152	226
ELA 6	Full	90	163	171	267
	Reduced	88	160	166	259

Figure 3.22. Histogram of Testing Time for The Full-Length Tests—Grade 4 Math

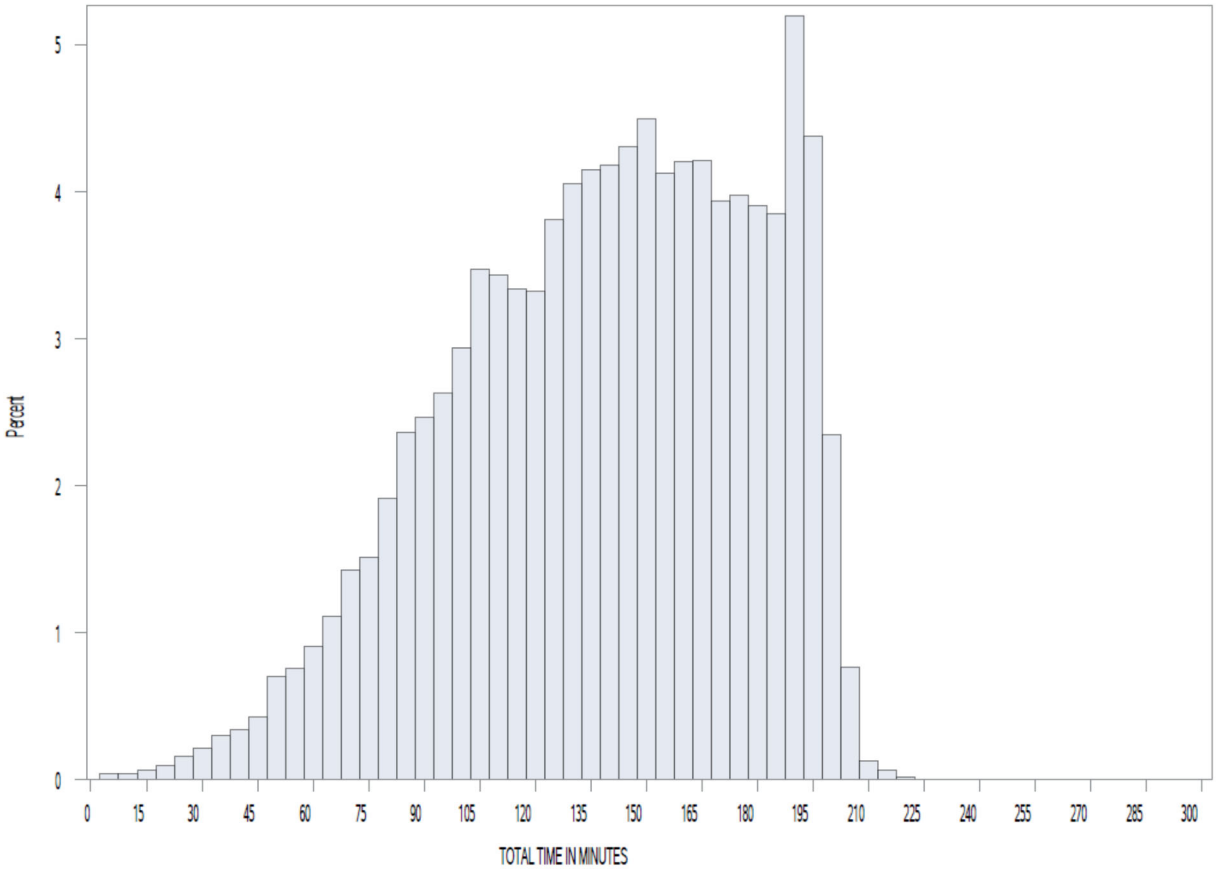


Figure 3.23. Histogram of Testing Time for The Reduced-Length Tests—Grade 4 Math

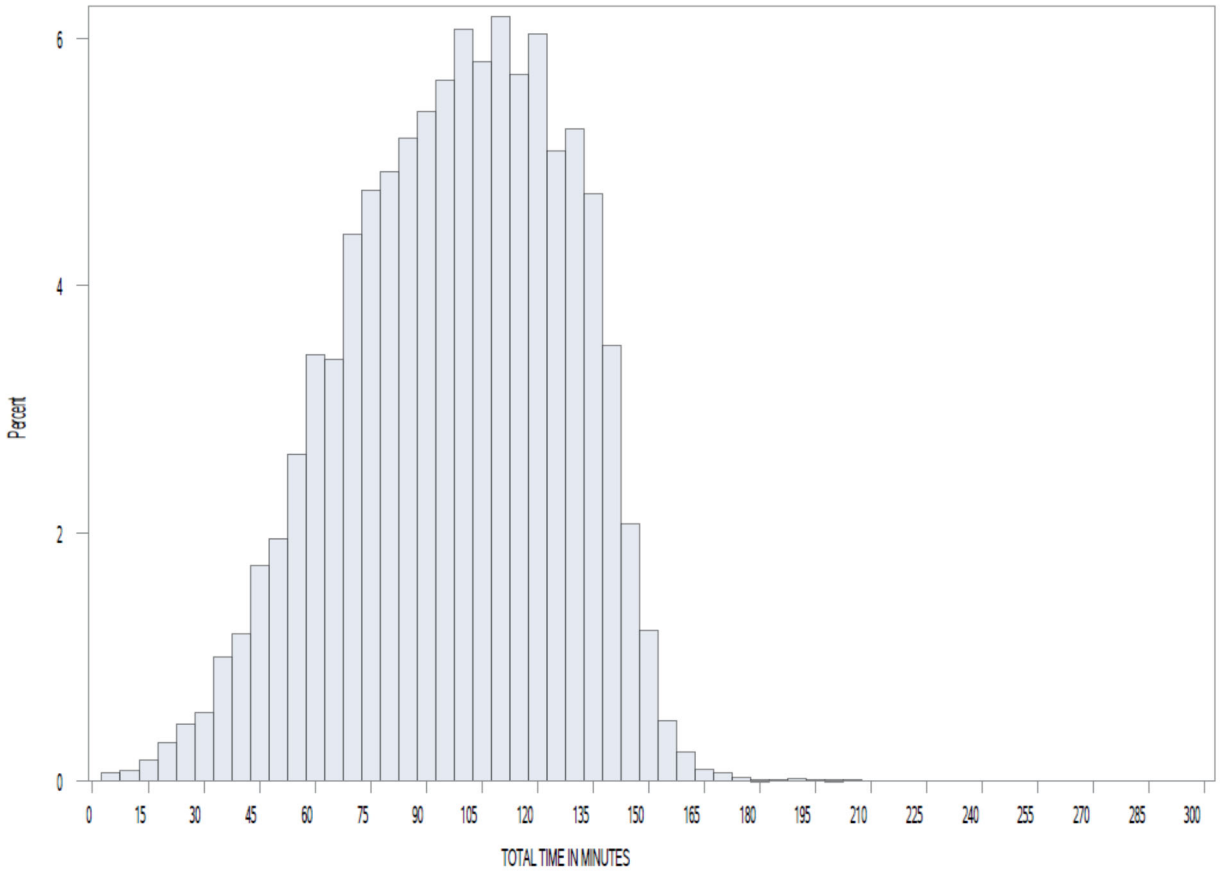


Figure 3.24. Histogram of Testing Time for The Full-Length Tests—Grade 6 Math

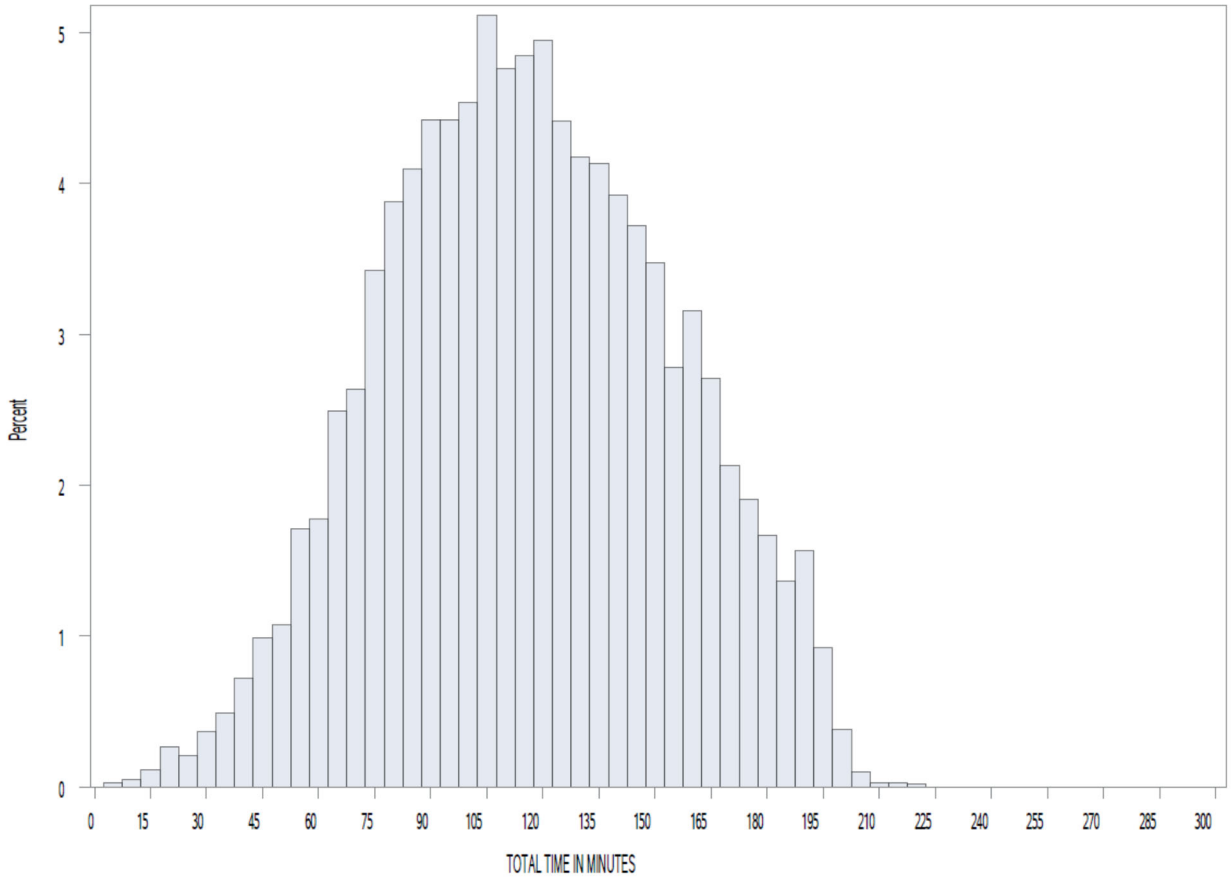


Figure 3.25. Histogram of Testing Time for The Reduced-Length Tests—Grade 6 Math

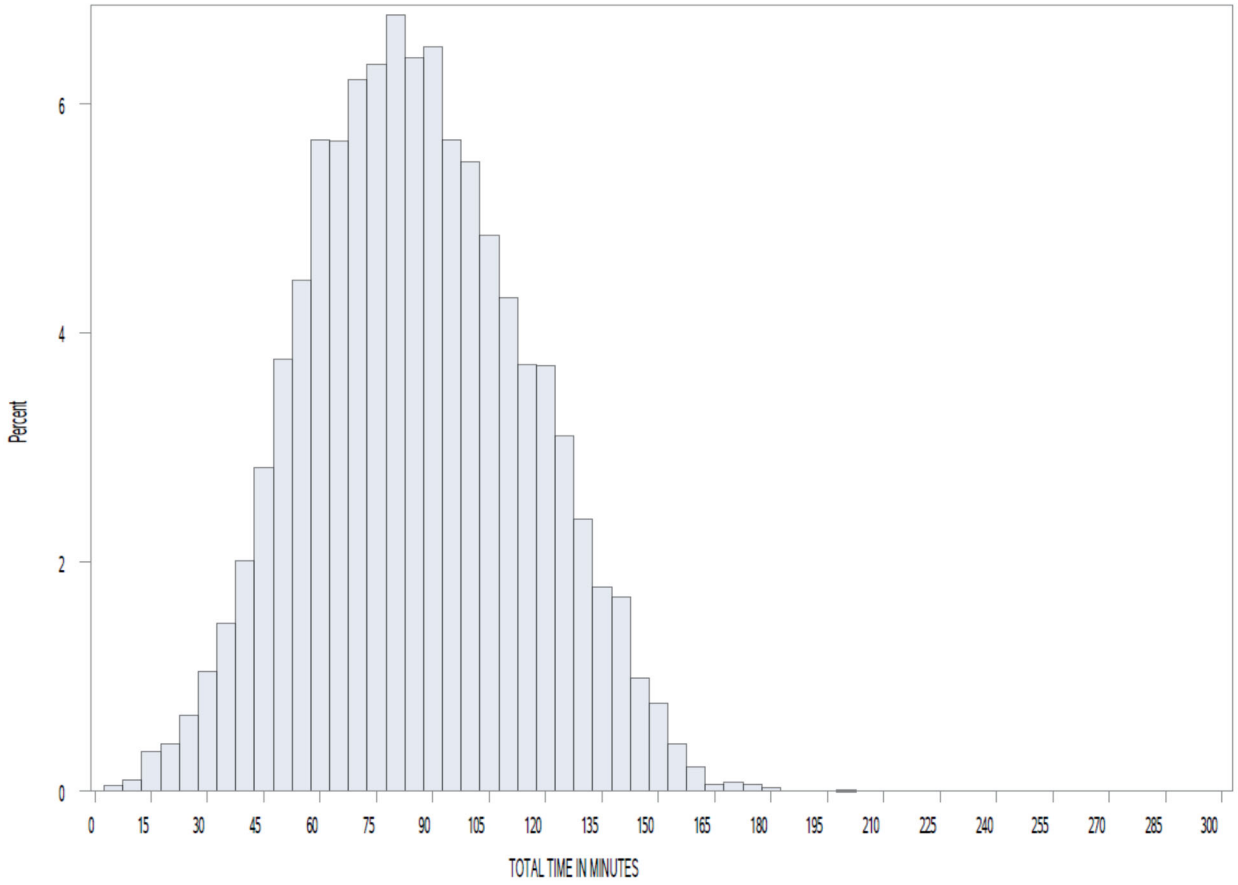


Figure 3.26. Histogram of Testing Time for The Full-Length Tests—Grade 4 ELA

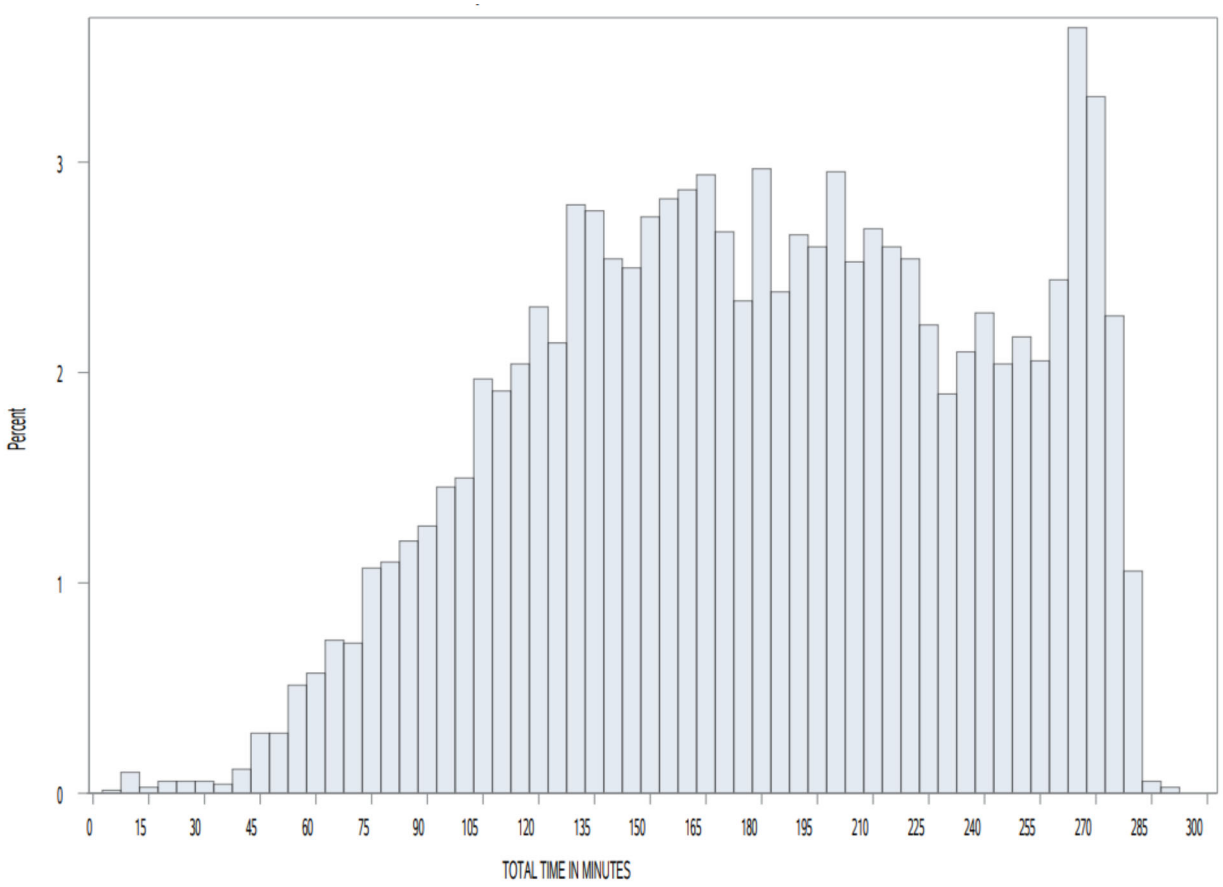


Figure 3.27. Histogram of Testing Time for The Reduced-Length Tests—Grade 4 ELA

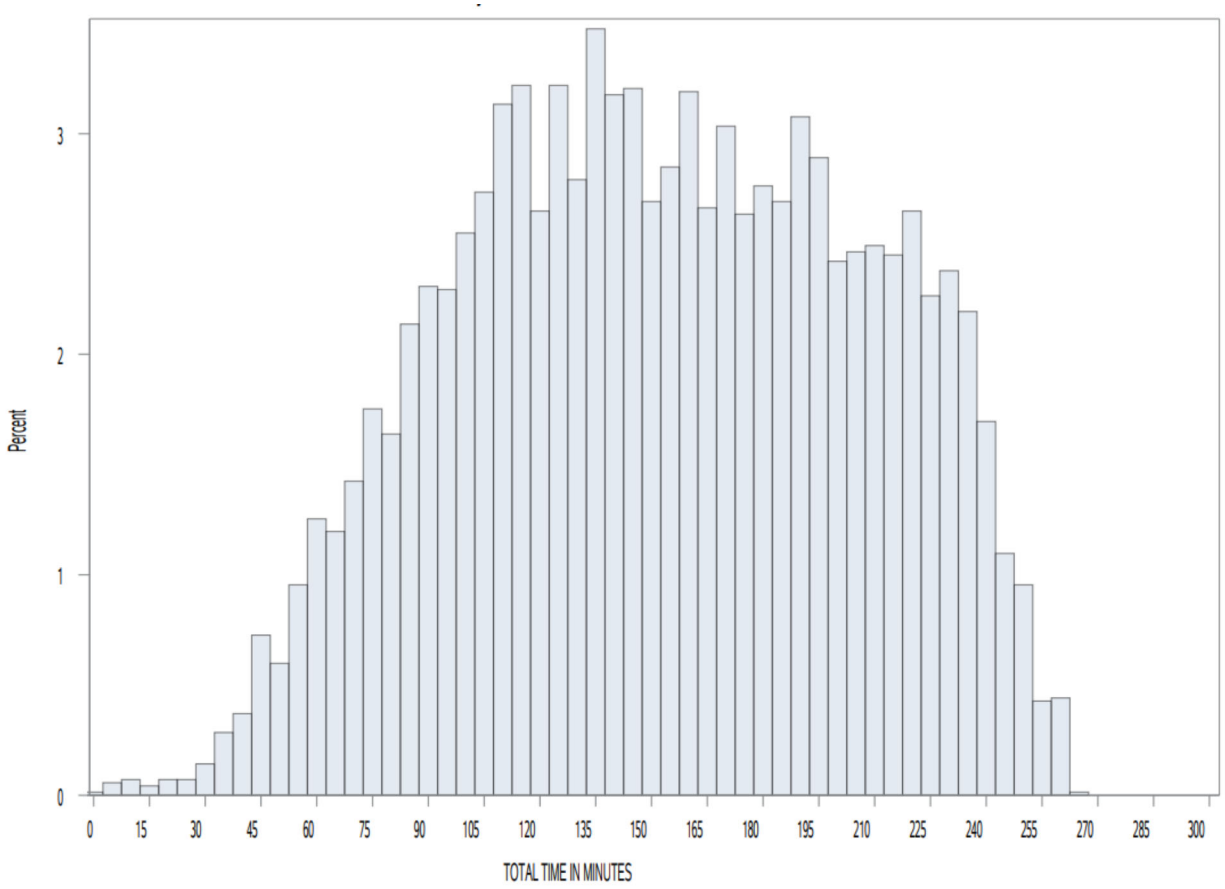


Figure 3.28. Histogram of Testing Time for The Full-Length Tests—Grade 6 ELA

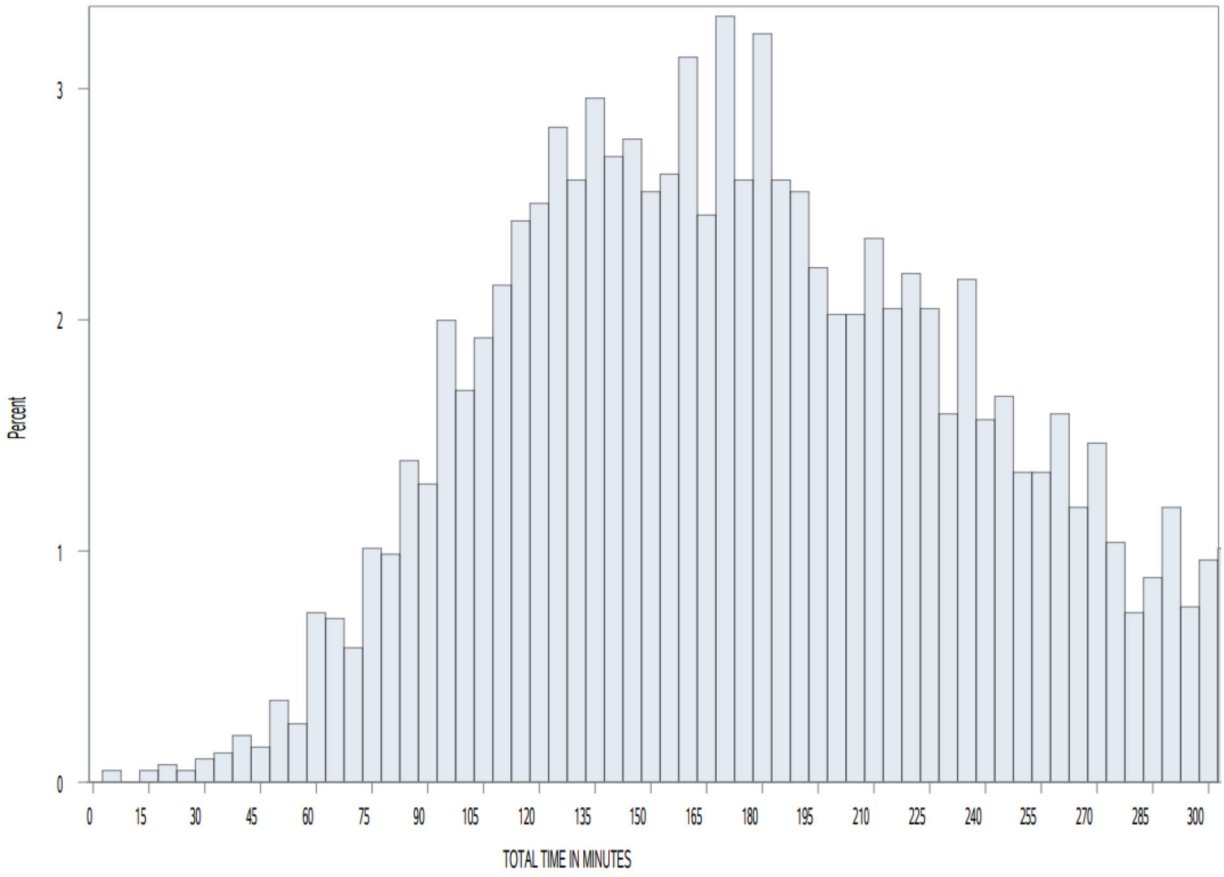
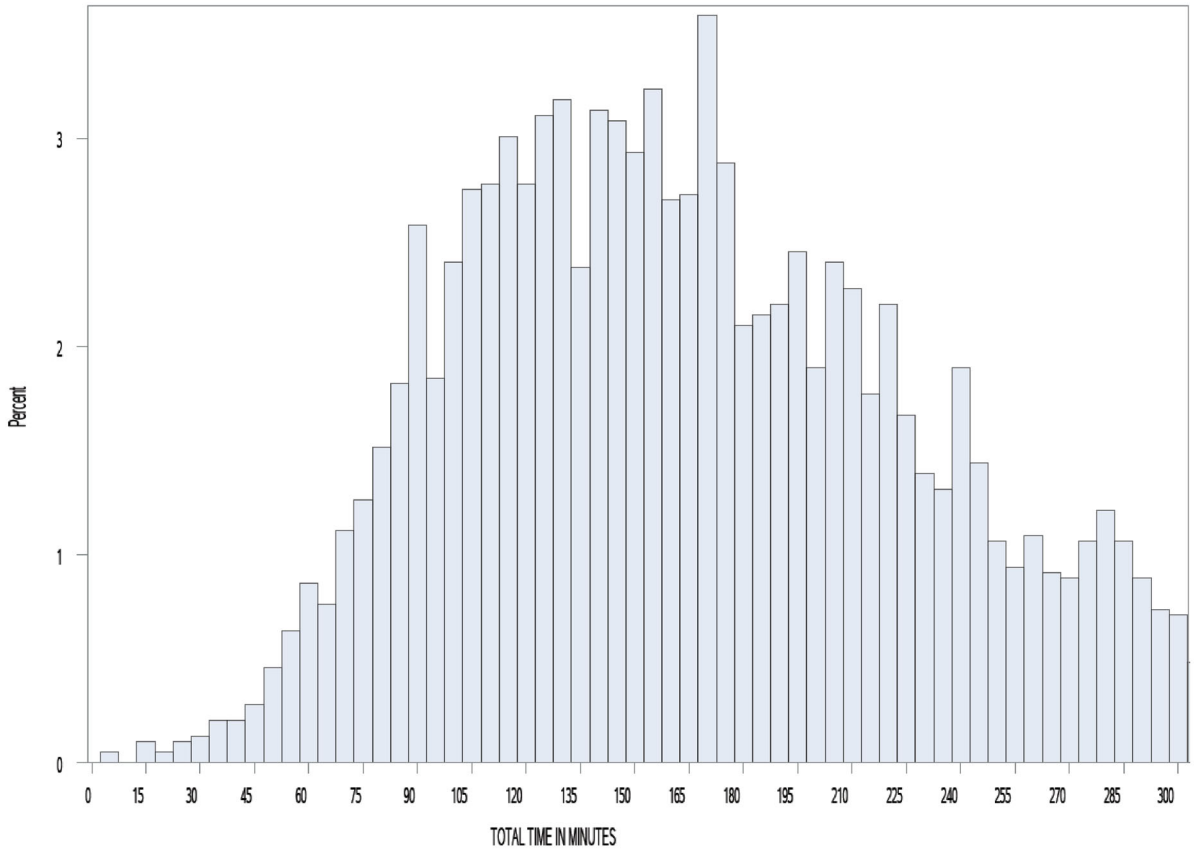


Figure 3.29. Histogram of Testing Time for The Reduced-Length Tests—Grade 6 ELA



3.2. CAT Simulation (Mathematics)

This section presents the results for the mathematics CAT simulation. These reflect full length and reduced fixed length simulations plus an unconstrained condition for grades 4 and 6. The results include overall performance level agreement, standard performance level agreement, simulator summary statistics, and scatterplots of the true versus simulated ability estimates.

Table 3.16 presents the percentage of students assigned the exact same performance level for both the full and the reduced length mathematics fixed CAT simulations plus the unconstrained CAT simulation. The percent of exact agreement was highest for the unconstrained condition, and lower for the reduced length fixed simulation than the full length.

Table 3.26. Summary Statistics for Full and Reduced Length Scores—Mathematics

Assessment	Scenario	Exact Agreement	Higher Level for Scenario	Lower Level for Scenario
Mathematics 4	Full	84.34%	7.58%	8.08%
	Reduced	80.79%	9.18%	10.03%
	Unconstrained	85.22%	6.95%	7.84%
Mathematics 6	Full	82.95%	8.56%	8.50%
	Reduced	79.32%	10.29%	10.39%
	Unconstrained	85.31%	7.03%	7.65%

Table 3.17 – Table 3.22 present the number and percentage of students comparing the full and reduced-length fixed form CAT simulation results and the unconstrained CAT simulation results to the "true" theta ability estimates. The values bolded in the tables represent exact agreement. For the full-length fixed form and unconstrained CAT conditions, if the performance level designation was not exact, the difference was always within an adjacent performance level. This pattern is similar to that presented in the previous table, where the reduced form had the lowest overall level of exact agreement.

Table 3.27. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 4 Math

Performance Level	Full Fixed L1	Full Fixed L2	Full Fixed L3	Full Fixed L4	Full Fixed L5	**Total**
2025 L1	**7,199 (13.71%)**	1,113 (2.12%)	0 (0%)	0 (0%)	0 (0%)	8,312 (15.83%)
2025 L2	1,254 (2.39%)	**9,450 (17.99%)**	1,205 (2.29%)	0 (0%)	0 (0%)	11,909 (22.68%)
2025 L3	0 (0%)	1,175 (2.24%)	**10,544 (20.08%)**	1,205 (2.29%)	0 (0%)	12,924 (24.61%)
2025 L4	0 (0%)	0 (0%)	1,176 (2.24%)	**15,389 (29.30%)**	723 (1.38%)	17,288 (32.92%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	374 (0.71%)	**1,711 (3.26%)**	2,085 (3.97%)
Total	8,453 (16.10%)	11,738 (22.35%)	12,925 (24.61%)	16,968 (32.31%)	2,434 (4.63%)	**52,518 (100%)**

Table 3.28. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 4 Math

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**6,990 (13.31%)**	1,322 (2.52%)	0 (0%)	0 (0%)	0 (0%)	8,312 (15.83%)
2025 L2	1,557 (2.96%)	**8,964 (17.07%)**	1,388 (2.64%)	0 (0%)	0 (0%)	11,909 (22.68%)
2025 L3	2 (0%)	1,388 (2.64%)	**9,976 (19.00%)**	1,558 (2.97%)	0 (0%)	12,924 (24.61%)
2025 L4	0 (0%)	0 (0%)	1,415 (2.69%)	**14,871 (28.32%)**	1,002 (1.91%)	17,288 (32.92%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	459 (0.87%)	**1,626 (3.10%)**	2,085 (3.97%)
Total	8,549 (16.28%)	11,674 (22.23%)	12,779 (24.33%)	16,888 (32.16%)	2,628 (5.00%)	**52,518 (100%)**

Table 3.29. Overall Performance Level Percent Agreement Comparing the Unconstrained CAT Simulation Results to Initial "True" Ability—Grade 4 Math

Performance Level	Unconstrained L1	Unconstrained L2	Unconstrained L3	Unconstrained L4	Unconstrained L5	**Total**
2025 L1	**7,267 (13.84%)**	1,045 (1.99%)	0 (0%)	0 (0%)	0 (0%)	8,312 (15.83%)
2025 L2	1,040 (1.98%)	**9,546 (18.18%)**	1,323 (2.52%)	0 (0%)	0 (0%)	11,909 (22.68%)
2025 L3	0 (0%)	1,056 (2.01%)	**10,553 (20.09%)**	1,315 (2.50%)	0 (0%)	129,24 (24.61%)
2025 L4	0 (0%)	0 (0%)	1,212 (2.31%)	**15,644 (29.79%)**	432 (0.82%)	17,288 (32.92%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	341 (0.65%)	**1,744 (3.32%)**	2,085 (3.97%)
Total	8,307 (15.82%)	11,647 (22.18%)	13,088 (24.92%)	17,300 (32.94%)	2,176 (4.14%)	**52,518 (100%)**

Table 3.30. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 6 Math

Performance Level	Full Fixed L1	Full Fixed L2	Full Fixed L3	Full Fixed L4	Full Fixed L5	**Total**
2025 L1	**8,065 (15.44%)**	1,238 (2.37%)	0 (0%)	0 (0%)	0 (0%)	9,303 (17.81%)
2025 L2	1,422 (2.72%)	**9,497 (18.18%)**	1,352 (2.59%)	0 (0%)	0 (0%)	12,271 (23.49%)
2025 L3	0 (0%)	1,515 (2.90%)	**11,565 (22.14%)**	1,305 (2.50%)	0 (0%)	14,385 (27.54%)
2025 L4	0 (0%)	0 (0%)	1,188 (2.27%)	**12,046 (23.06%)**	544 (1.04%)	13,778 (26.38%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	344 (0.66%)	**2,155 (4.13%)**	2,499 (4.78%)
Total	9,487 (18.16%)	12,250 (23.45%)	14,105 (27.00%)	13,695 (26.22%)	2,699 (5.17%)	**52,236 (100%)**

Table 3.31. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 6 Math

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**7,865 (15.06%)**	1,434 (2.75%)	4 (0.01%)	0 (0%)	0 (0%)	9,303 (17.81%)
2025 L2	1,693 (3.24%)	**8,952 (17.14%)**	1,624 (3.11%)	2 (0.00%)	0 (0%)	12,271 (23.49%)
2025 L3	2 (0.00%)	1,794 (3.43%)	**10,952 (20.97%)**	1,637 (3.13%)	0 (0%)	14,385 (27.54%)
2025 L4	0 (0%)	0 (0%)	1,491 (2.85%)	**11,562 (22.13%)**	725 (1.39%)	13,778 (26.38%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	396 (0.76%)	**2,103 (4.03%)**	2,499 (4.78%)
Total	9,560 (18.30%)	12,180 (23.32%)	14,071 (26.94%)	13,597 (26.03%)	2,828 (5.41%)	**52,236 (100%)**

Table 3.32. Overall Performance Level Percent Agreement Comparing the Unconstrained CAT Simulation Results to Initial "True" Ability—Grade 6 Math

Performance Level	Unconstrained L1	Unconstrained L2	Unconstrained L3	Unconstrained L4	Unconstrained L5	**Total**
2025 L1	**8,310 (15.91%)**	993 (1.90%)	0 (0%)	0 (0%)	0 (0%)	9,303 (17.81%)
2025 L2	954 (1.83%)	**10,087 (19.31%)**	1,230 (2.35%)	0 (0%)	0 (0%)	12,271 (23.49%)
2025 L3	0 (0%)	1,289 (2.47%)	**11,758 (22.51%)**	1,338 (2.56%)	0 (0%)	14,385 (27.54%)
2025 L4	0 (0%)	0 (0%)	1,115 (2.13%)	**12,227 (23.41%)**	436 (0.83%)	13,778 (26.38%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	316 (0.60%)	**2,183 (4.18%)**	2,499 (4.78%)
Total	9,264 (17.73%)	12,369 (23.68%)	14,103 (27.00%)	13,881 (26.57%)	2,619 (5.01%)	**52,236 (100%)**

Table 3.23 and Table 3.24 present summary statistics from the CAT simulations for the full and reduced-length fixed form and unconstrained conditions. Here the minimum, maximum, and average (mean) statistics are provided for number of items presented, item exposure rates, difference between "true" ability and estimated ability, and the conditional standard errors of measurement. The score difference and CSEM are on the scale score (reporting) metric.

Table 3.33. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form and Unconstrained Conditions—Grade 4 Math

Grade 4	**Full Length (MSE = 0.06)**			**Reduced (MSE = 0.11)**			**Unconstrained (MSE = 0.04)**		
	Min	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
Number of Items	30	30	30	20	20	20	13	24.37	245
Item Exposure Rate	<0.01%	13.51%	67.36%	<0.01%	10.70%	68.45%	0.12%	9.67%	98.95%
Score Difference	-104.77	0.06	107.61	-115.41	0.09	134.71	-38.83	0.13	30.54
CSEM	5.51	6.95	83.58	6.24	8.79	136.64	5.66	6.04	31.94

Table 3.34. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form and Unconstrained Conditions—Grade 6 Math

Grade 6	Full Length (MSE = 0.08)			Reduced (MSE = 0.14)			Unconstrained (MSE = 0.04)		
	Min	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
Number of Items	29	29	29	18	18	18	13	29.8	245
Item Exposure Rate	<0.01%	14.15%	59.86%	<0.01%	11.76%	66.26%	1.37%	12.12%	91.38%
Score Difference	-88.35	-0.38	66.15	-109.91	-0.44	107.46	-33.84	0.08	29.83
CSEM	5.27	7.31	74.55	6.47	9.3	124.58	5.37	5.73	28.13

Figure 3.30. Overall Full-Length Fixed Form Simulation vs. "True" Ability—Grade 4 Math

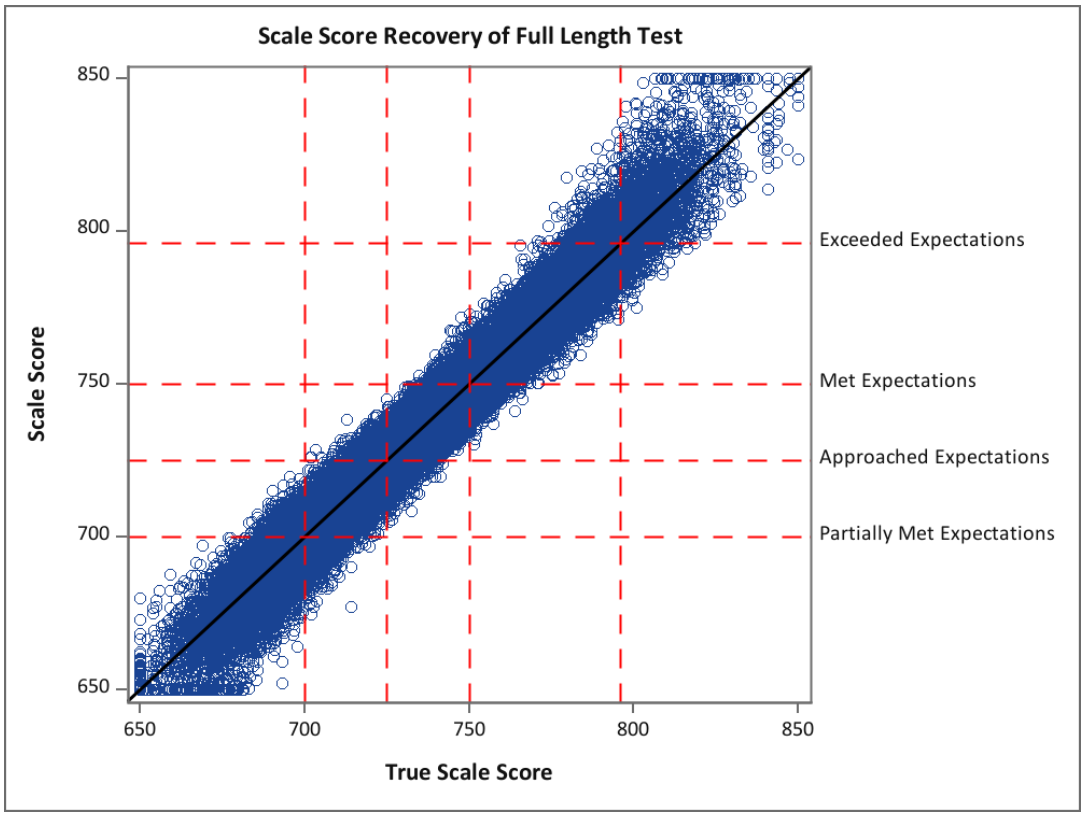


Figure 3.31. Overall Reduced-Length Fixed Form Simulation vs. "True" Ability—Grade 4 Math

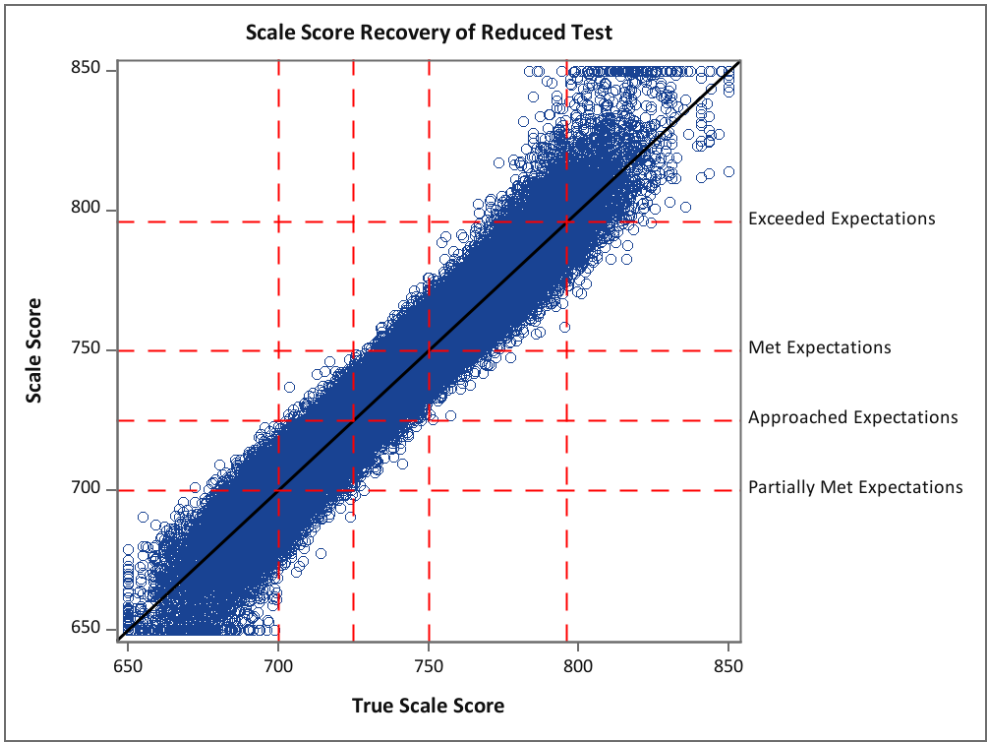


Figure 3.32. Unconstrained Simulation vs. "True" Ability—Grade 4 Math

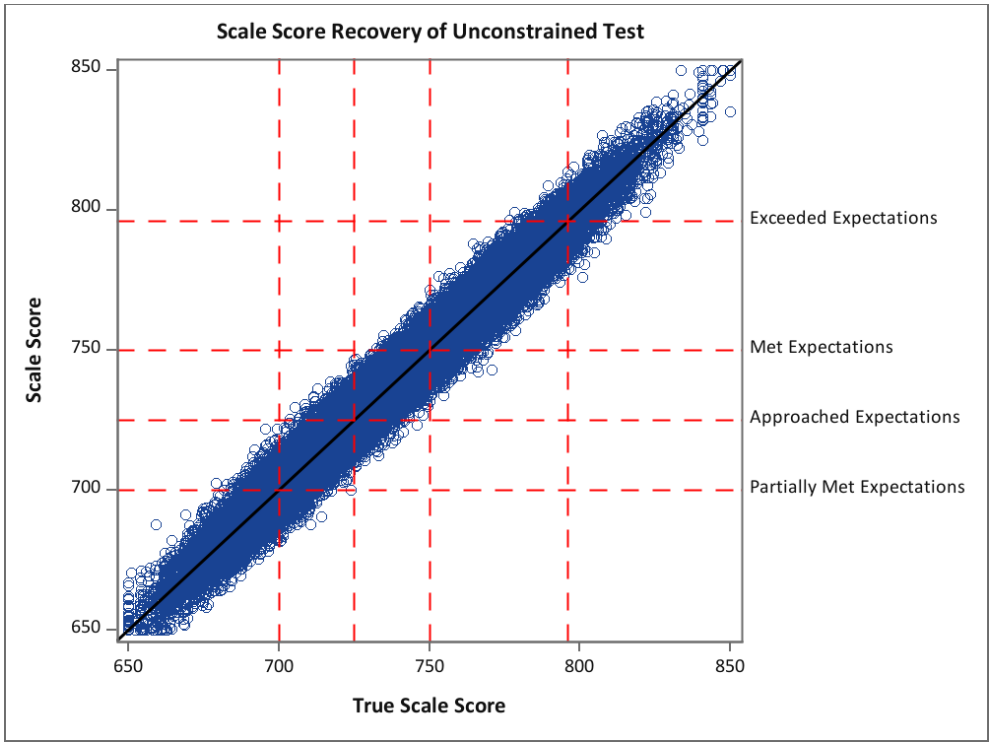


Figure 3.33. Overall Full-Length Fixed Form Simulation vs. "True" Ability—Grade 6 Math

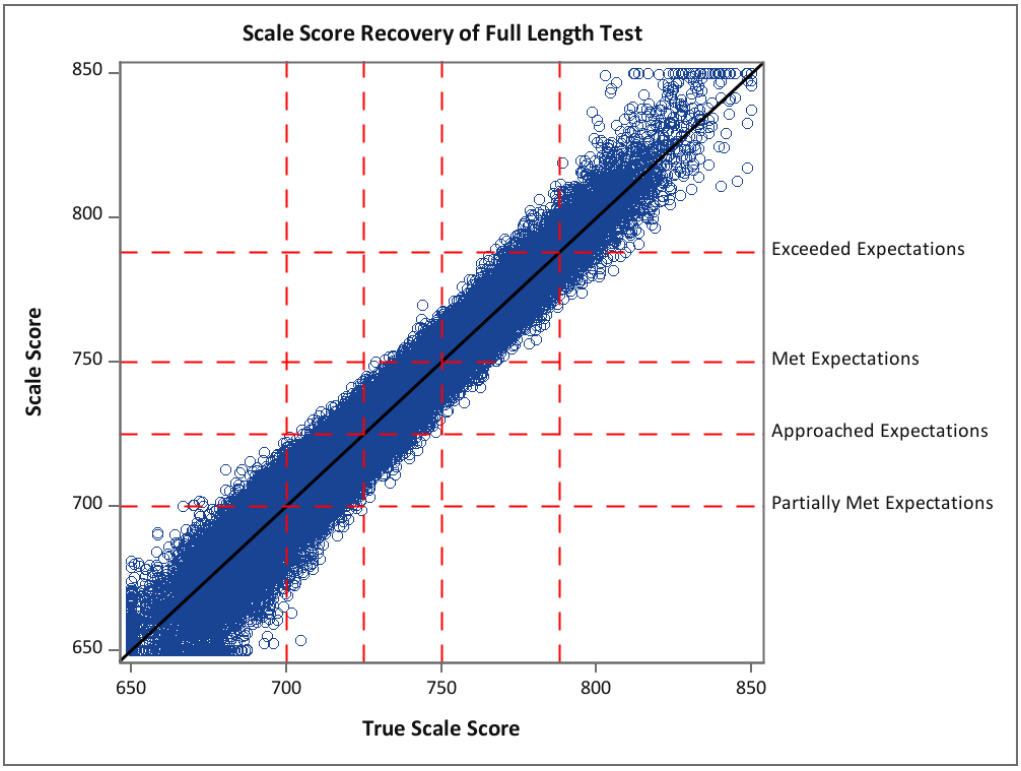


Figure 3.34. Overall Reduced-Length Fixed Form Simulation vs. "True" Ability—Grade 6 Math

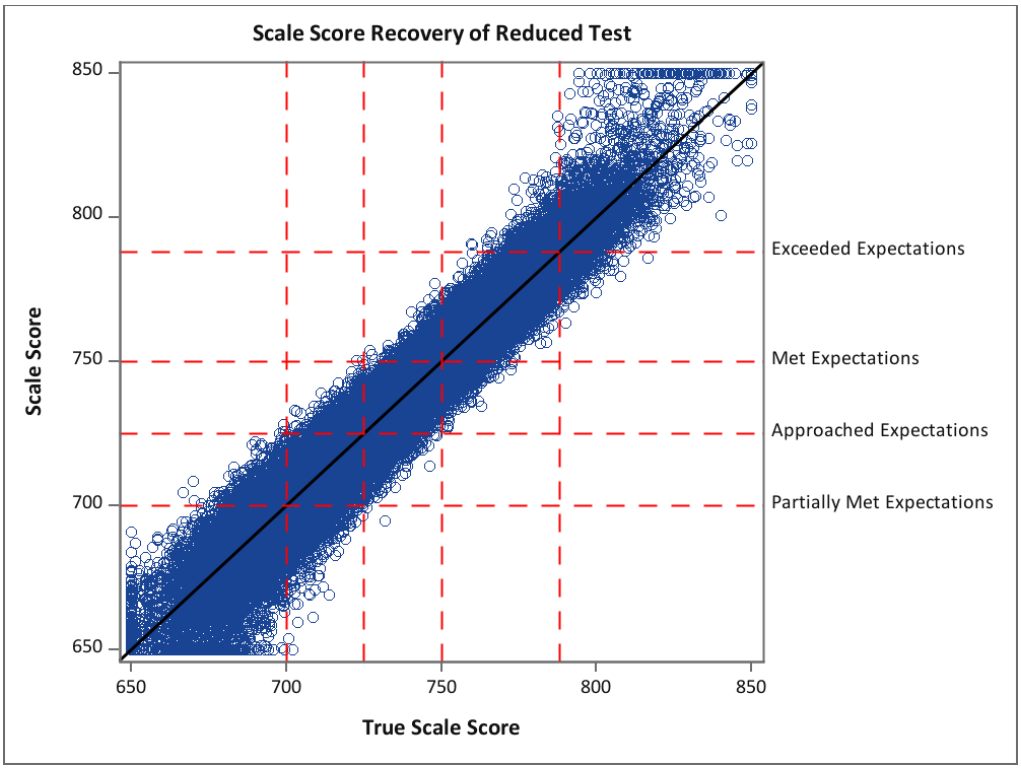
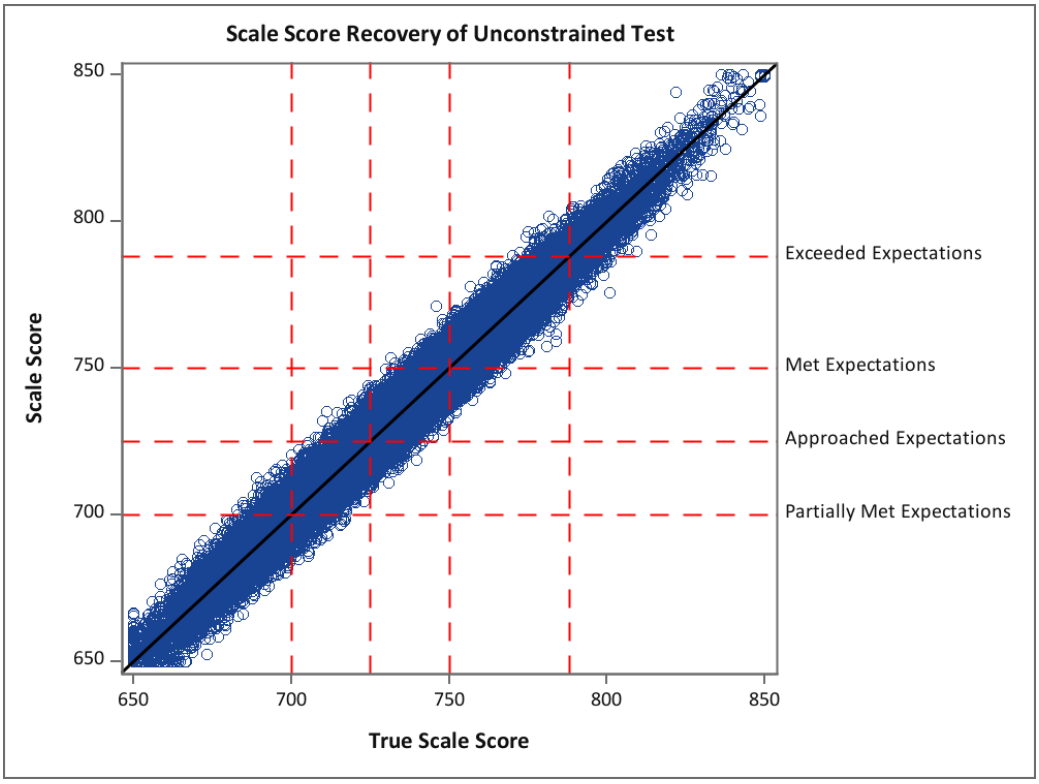


Figure 3.35. Unconstrained Simulation vs. "True" Ability—Grade 6 Math



3.3. Multistage Adaptive Testing Simulation (ELA)

This section presents the results for the ELA MST simulation. These reflect full length and reduced fixed length simulations conditions for grades 4 and 6. The results include overall performance level agreement, standard performance level agreement, simulator summary statistics, and scatterplots of the true versus simulated ability estimates for grades 4 and 6.

Table 3.25 presents the percentage of students assigned the exact same performance level for both the full and the reduced length ELA fixed CAT simulations. The percent of exact agreement was similar for the two conditions of grade 6. For grade 4, the percent of exact agreement was higher for the full length condition than that of the reduced length fixed condition.

Table 3.35. Summary Statistics for Full and Reduced Length Scores—ELA

Assessment	Scenario	Exact Agreement	Higher Level for Scenario	Lower Level for Scenario
ELA 4	Full	76.27%	11.74%	11.99%
	Reduced	71.48%	14.13%	14.38%
ELA 6	Full	78.90%	10.60%	10.50%
	Reduced	77.64%	11.18%	11.17%

Table 3.26 – Table 3.29 present the number and percentage of students comparing the full and reduced-length fixed form CAT simulation results to the "true" theta ability estimates. The values bolded in the tables represent exact agreement. In all the tables, most cases matched exactly or in adjacent levels. There were more instances of cases that differed by two levels. This pattern is similar to that presented in the previous table, where the reduced form had the lowest overall level of exact agreement.

Table 3.36. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 4 ELA

Performance Level	Full Fixed L1	Full Fixed L2	Full Fixed L3	Full Fixed L4	Full Fixed L5	**Total**
2025 L1	**6,416 (12.28%)**	1,108 (2.12%)	4 (0.01%)	0 (0%)	0 (0%)	7,528 (14.41%)
2025 L2	1,358 (2.60%)	**7,009 (13.41%)**	1,529 (2.93%)	14 (0.03%)	0 (0%)	9,910 (18.96%)
2025 L3	4 (0.01%)	1739 (3.33%)	**9,266 (17.73%)**	2,189 (4.19%)	0 (0%)	13,198 (25.26%)
2025 L4	0 (0%)	5 (0.01%)	2,211 (4.23%)	**13,042 (24.96%)**	1,420 (2.72%)	16,678 (31.92%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	820 (1.57%)	**4,122 (7.89%)**	4,942 (9.46%)
Total	7,778 (14.88%)	9,861 (18.87%)	13,010 (24.90%)	16,065 (30.74%)	5,542 (10.61%)	**52,256 (100%)**

Table 3.37. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 4 ELA

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**6,180 (11.83%)**	1,332 (2.55%)	16 (0.03%)	0 (0.00%)	0 (0.00%)	7,528 (14.41%)
2025 L2	1,725 (3.30%)	**6,279 (12.02%)**	1,843 (3.53%)	63 (0.12%)	0 (0.00%)	9,910 (18.96%)
2025 L3	25 (0.05%)	2,151 (4.12%)	**8,406 (16.09%)**	2,615 (5.00%)	1 (0.00%)	13,198 (25.26%)
2025 L4	0 (0.00%)	36 (0.07%)	2,513 (4.81%)	**12,482 (23.89%)**	1,647 (3.15%)	16,678 (31.92%)
2025 L5	0 (0.00%)	0 (0.00%)	3 (0.01%)	932 (1.78%)	**4,007 (7.67%)**	4,942 (9.46%)
Total	7,930 (15.18%)	9,798 (18.75%)	12,781 (24.46%)	16,092 (30.79%)	5,655 (10.82%)	**52,256 (100%)**

Table 3.38. Overall Performance Level Percent Agreement Comparing the Full-Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 6 ELA

Performance Level	Full Fixed L1	Full Fixed L2	Full Fixed L3	Full Fixed L4	Full Fixed L5	**Total**
2025 L1	**5,107 (9.74%)**	1,019 (1.94%)	1 (0.00%)	0 (0%)	0 (0%)	6,127 11.69%)
2025 L2	1,358 (2.59%)	**7,267 (13.86%)**	1,446 (2.76%)	4 (0.01%)	0 (0%)	10,075 19.22%)
2025 L3	0 (0%)	1,586 (3.03%)	**9,910 (18.90%)**	1,907 (3.64%)	0 (0%)	13,403 25.57%)
2025 L4	0 (0%)	1 (0.00%)	1,951 (3.72%)	**15,775 (30.09%)**	1,129 (2.15%)	18,856 35.97%)
2025 L5	0 (0%)	0 (0%)	0 (0%)	661 (1.26%)	**3,298 (6.29%)**	3,959 7.55%)
Total	6,465 (12.33%)	9,873 (18.83%)	13,308 (25.39%)	18,347 (35.00%)	4,427 (8.45%)	**52,420 (100%)**

Table 3.39. Overall Performance Level Percent Agreement Comparing the Reduced Length Fixed Form CAT Simulation Results to Initial "True" Ability—Grade 6 ELA

Performance Level	Reduced L1	Reduced L2	Reduced L3	Reduced L4	Reduced L5	**Total**
2025 L1	**5,020 (9.58%)**	1,103 (2.10%)	4 (0.01%)	0 (0.00%)	0 (0.00%)	6,127 (11.69%)
2025 L2	1,530 (2.92%)	**6,983 (13.32%)**	1,557 (2.97%)	5 (0.01%)	0 (0.00%)	10,075 (19.22%)
2025 L3	7 (0.01%)	1,659 (3.16%)	**9,721 (18.54%)**	2,016 (3.85%)	0 (0.00%)	13,403 (25.57%)
2025 L4	0 (0.00%)	4 (0.01%)	1,984 (3.78%)	**15,694 (29.94%)**	1,174 (2.24%)	18,856 (35.97%)
2025 L5	0 (0.00%)	0 (0.00%)	0 (0.00%)	677 (1.29%)	**3,282 (6.26%)**	3,959 (7.55%)
Total	6,557 (12.51%)	9,749 (18.60%)	13,266 (25.31%)	18,392 (35.09%)	4,456 (8.50%)	**52,420 (100%)**

Table 3.30 and Table 3.31 present summary statistics from the CAT simulations for the full and reduced-length fixed form and unconstrained conditions. The minimum, maximum, and average (mean) statistics are provided for number of items presented, item exposure rates, difference between "true" ability and estimated ability, and the conditional standard errors of measurement. The score difference and CSEM are on the scale score (reporting) metric.

Table 3.40. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form Conditions—Grade 4 ELA

	Full Length (MSE=0.14)			Reduced Length (MSE=0.24)		
	Min.	Mean	Max.	Min.	Mean	Max.
#Items	26	26	26	17	17	17
Scale Score Difference	-56.44	0.04	47.19	-57.54	-0.05	55.48
CSEM	8.70	10.86	82.42	10.83	13.58	117.27

Table 3.41. CAT Simulation Summary Statistics for Full and Reduced Length Fixed Form Conditions—Grade 6 ELA

	Full Length (MSE=0.11)			Reduced Length (MSE=0.14)		
	Min.	Mean	Max.	Min.	Mean	Max.
#Items	27	27	27	23	23	23
Scale Score Difference	-48.76	-0.13	53.10	-50.66	-0.20	46.66
CSEM	7.70	9.01	77.26	8.17	9.71	87.31

Figure 3.36. Overall Full-Length Fixed Form Simulation vs. "True" Ability—Grade 4 ELA

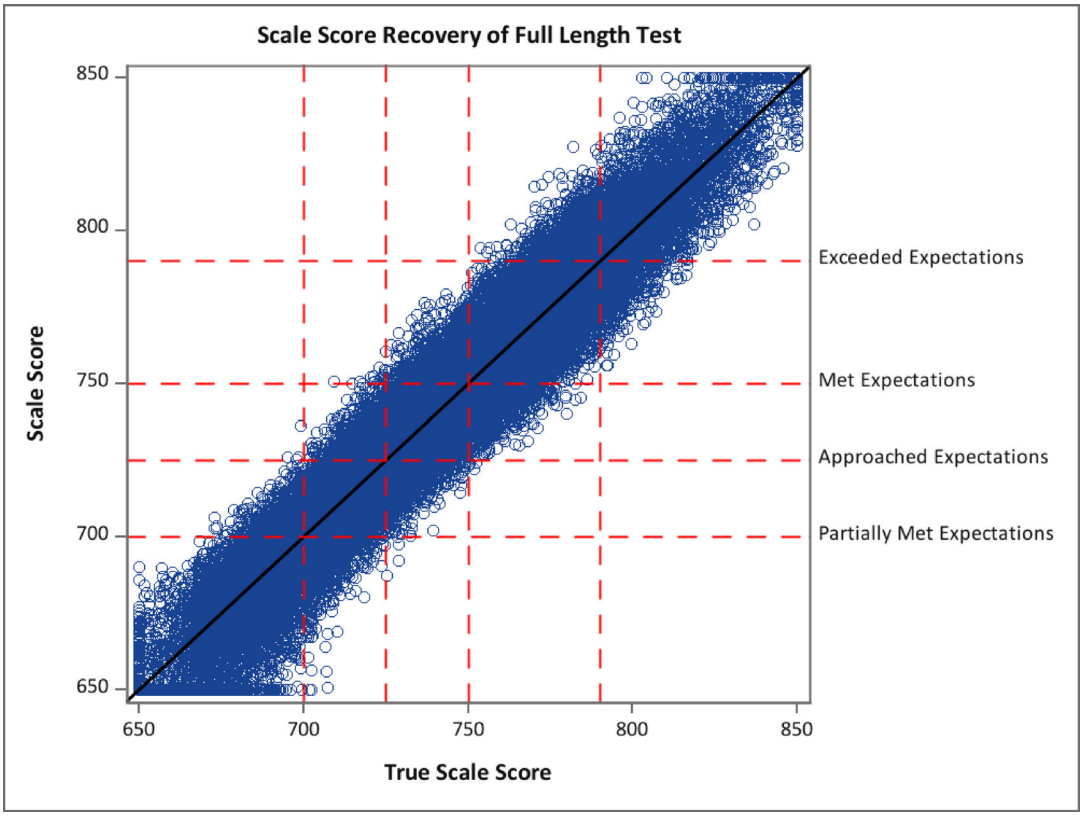


Figure 3.37. Overall Reduced-Length Fixed Form Simulation vs. "True" Ability—Grade 4 ELA

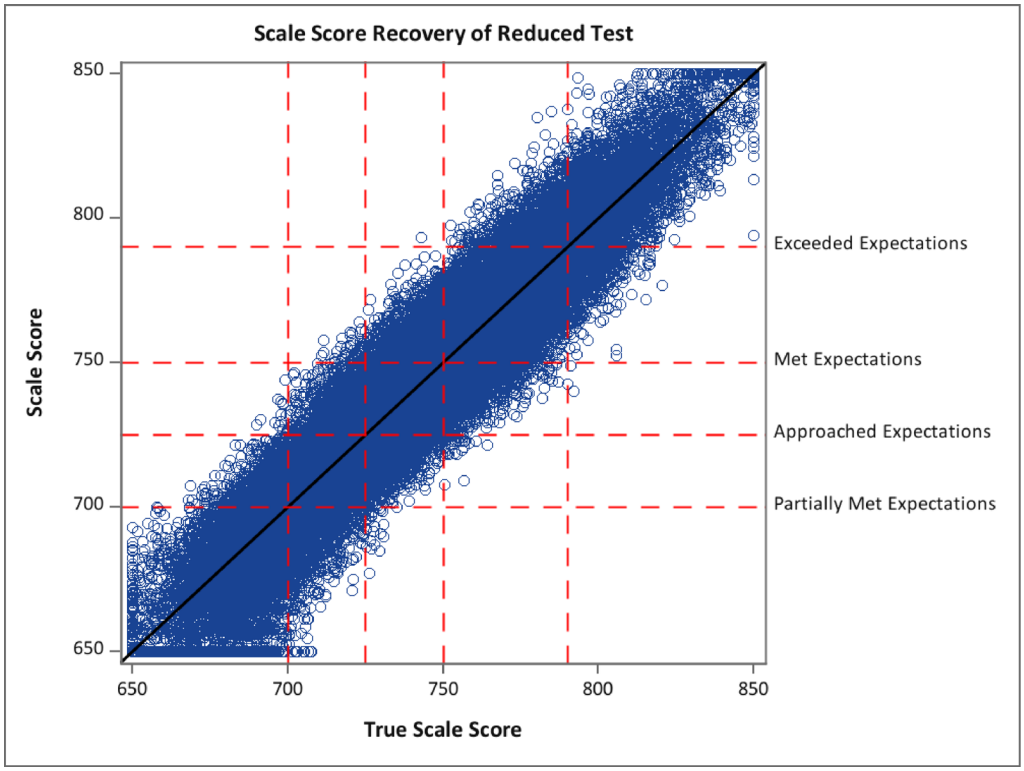


Figure 3.38. Overall Full-Length Fixed Form Simulation vs. "True" Ability—Grade 6 ELA

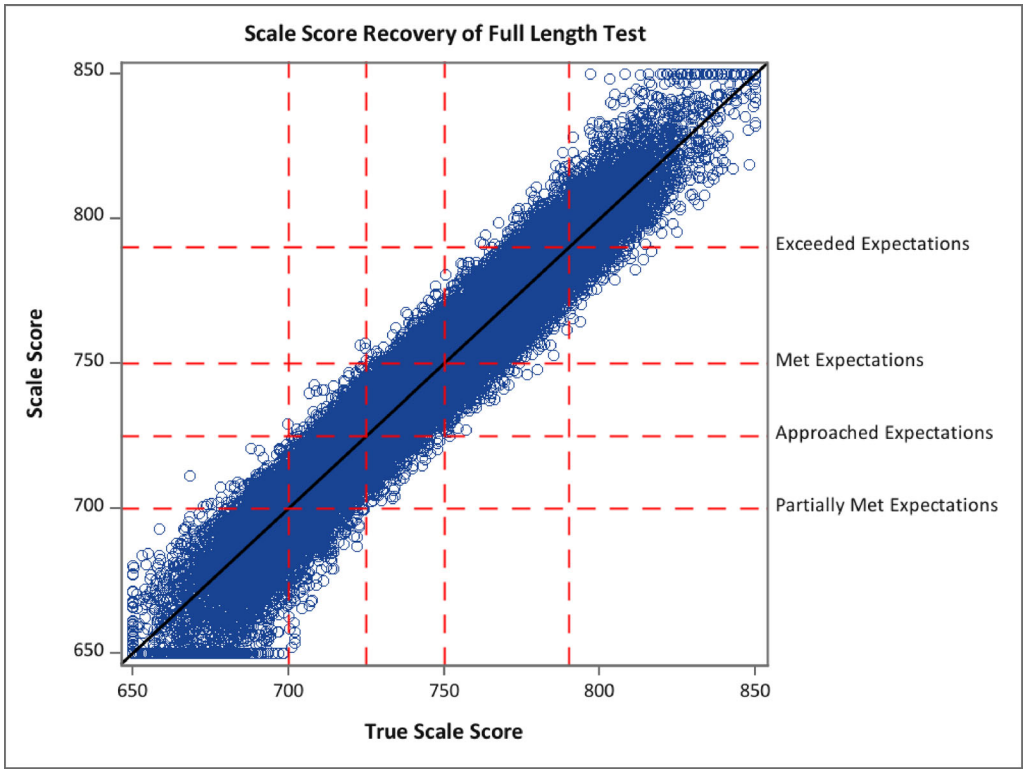
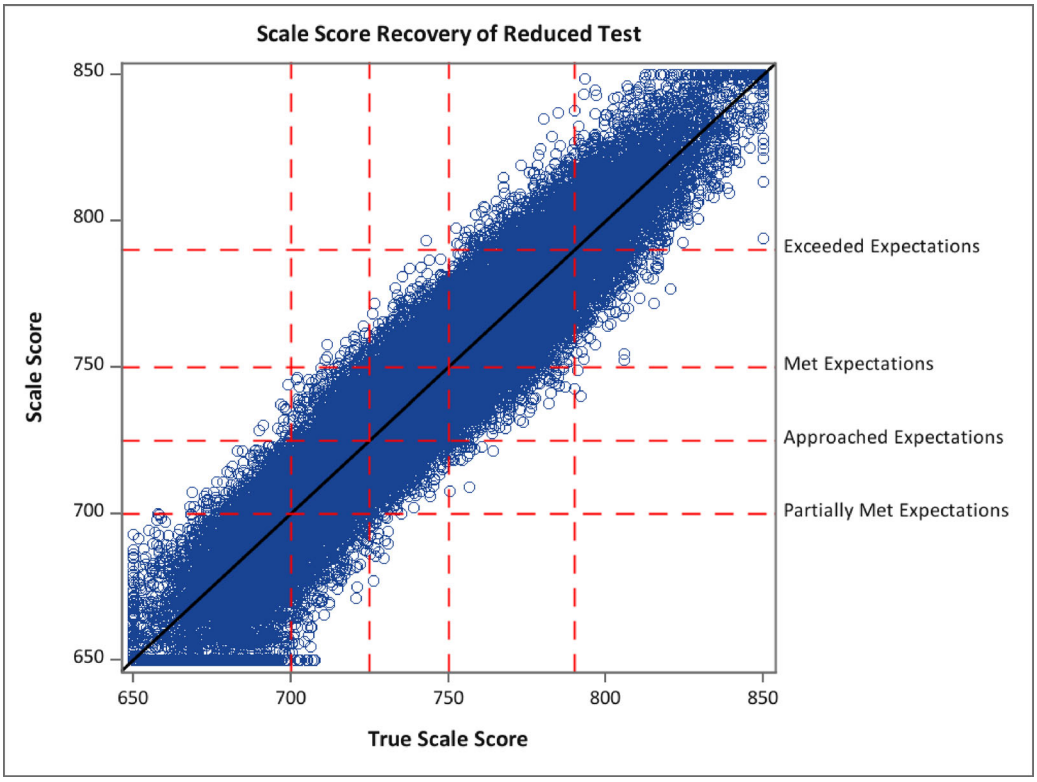


Figure 3.39. Overall Reduced-Length Fixed Form Simulation vs. "True" Ability—Grade 6 ELA



4. Summary

These analyses were intended to address H.B. 25-1278 study requirements and recommendations from the 1241 Accountability Task Force, which advised the state to investigate methods for shortening statewide assessments and to explore the implementation of adaptive assessment technologies. The analyses were designed to leverage existing student response data from the spring 2025 CMAS operational administration. They were carried out in two phases, where the first examined test shortening of CMAS Mathematics and ELA test options in line with federal accountability. The second phase explored computer adaptive testing within a CMAS framework by carrying out several simulations.

A main question coming into this research was articulated in the introduction regarding CAT and ESSA, where it was noted that a common misconception of adaptive testing is that it necessarily leads to reduced testing time. For CMAS, any transition would need to meet ESSA requirements. So that transition would necessarily have to involve meeting a blueprint that guarantees students are being assessed across the depth and breadth of the standards. For illustrative purposes, the condition of an unconstrained item-level adaptive condition for mathematics was carried out to examine the extent to which one might see reduced testing time. Keep in mind that the stopping rule for the unconstrained simulations was to meet a certain threshold measurement precision level.

The results were telling on several fronts. First, the number of items presented varied greatly depending on what ability level the simulated examinee was at. Keep in mind that the CMAS item banks are not built to support adaptive testing, but to support linear testing. Robust adaptive test banks cover the full range of standards across the entire ability range. Here we saw that the number of items administered ranged from 13 to 245. Second, the average number of items needed to meet the threshold level of precision was roughly the same as the full-length test. One additional observation with respect to the unrestricted condition is that without constraint, the selection of test questions is purely with respect to an examinee's ability level and the optimal statistical characteristics for the given item. As such, it's likely that the test a given examinee receives is vastly different in terms of content representativeness than other examinees. It's possible they could be administered a test comprised of items from a single standard. This too is a key reason why an unconstrained CAT is likely not a viable adaptive solution for Colorado within the federal accountability frame.

Despite this finding, the primary benefits of moving to an adaptive framework, even with blueprint constraints, is in terms of measurement precision. That is, since adaptive tests select items matched to an examinee's ability level, measurement precision is improved across the entire performance range, as opposed to being maximized around the cut scores (as with linear tests). This was observed as outcomes of each of the simulated item-level CAT conditions for CMAS Mathematics. It was less so for ELA, which differs from mathematics in that the adaptive nature occurs at only two points of the simulation (between stages 1 and 2, and between stages 2 and 3). Also, consistent with the test shortening analyses, the full-length fixed form simulation condition was more precise than the reduced length. One additional benefit often cited for adaptive testing is that examinees can be more engaged because the questions are selected to be at their level.

With respect to the test reduction analyses, we found predictable outcomes across the board. It is well known within the measurement community that reducing test length directly impacts the measurement properties of a given assessment. In effect, the reduced length measures offer less precision than longer length measures. And the impacts of the loss of precision manifests in terms of score reporting and interpretation. This was seen across all analyses with respect to the full and reduced length tests: greater standard errors, less precise classifications, lower reliabilities, etc.

However, these results of reduced test characteristics were still reasonably well aligned with the characteristics of the full-length versions. Most pronounced impacts were observed at the subclaim reporting level, where reductions would be magnified by those scores being based on only a handful of items per. But the overall test characteristics would generally be considered acceptable except for the grade 4 ELA reduced form, which resulted in a 0.83 reliability indicator.

Timing analyses were also presented. It is worth cautioning against overinterpreting the differences presented here. The full-length analyses include all the administered items (including tryout/field test items). For direct comparisons to the reduced-length tests, field test items were also included in those analyses. Neither scenario represents what might be implemented operationally. In other words, these results should only be interpreted within the context of this study and not a reflection of what the actual reduction in testing time would be in a real operational administration. However, it is noteworthy that these analyses highlight what was called out for the ELA reduction activity in that the limitations of the reductions made less of a difference where it came to actual reduction of testing time, in particular for Grade 6, where there was almost no difference to the full-length. What it does highlight is that there are likely other areas to explore in terms of test design with respect to reducing testing time.

One other area worth paying attention to has to do with the potential impact of shortening or transitioning to an adaptive scenario on score reporting and interpretation. As measurement precision decreases, so does the ability to make defensible finer grained score interpretations. And to the extent that important instructional or policy decisions are driven by test score comparisons, there is an increased chance of misinterpretation (i.e., concluding scores are different when unsupported, given the margin of error). The largest impacts from this study came at the subclaim level. At some point, shortening could limit the ability to report results like these out to stakeholders.

Currently CDE provides stakeholders with many different reports based on the results of the CMAS (and related) assessments. For reference, the reader in encouraged to consult the Interpretive Guide to Assessment Reports (https://www.cde.state.co.us/assessment/cmas_coalt_interpretiveguide). Throughout the guide are many instances where subclaim scores are reported and could be impacted. This would potentially undermine the availability of valuable information currently provided to stakeholders. Additionally, many scores that are provided in the current reports are based on a percent-correct metric and allow for score comparisons to be made across students within a given administration year, because students are tested on the same items. With a transition to CAT, especially for item-level, this would mean those scores could not be supported, since examinees would be tested on unique sets of items. Another kind of report that would also be impacted in this scenario is the Evidence Statement Analysis report. This report provides educators with comparative performance information at school, district, and state level on performance on items based on the assessed standards. This too is premised on the same scenario in which students currently take the same test forms and would be impacted with a transition to item-level CAT.

Lastly, there are several limitations to this study that are worth noting beyond the fact that only two CMAS grades and subjects were examined in this study. Given the similar methodologies for creating and

maintaining the CMAS assessments, we do expect these results to generalize to the other grade levels within subject. But we would recommend similar evaluations be carried out where other subjects are targeted.

Another limitation has to do with the fact that the current design of the CMAS tests is based on fixed linear forms built to maximize reliability at the cut scores. The banks of items that have been created to support these needs differ from the kinds of item banks necessary to support adaptive testing. In that scenario, robust CAT banks would ideally have items from all assessed standards reflected in the banks across the entire performance/ability range. In this study, the limitation manifested in the unconstrained CAT simulation where there were instances of simulated examinees being administered hundreds of items. Of special note is that the simulated examinees were drawn from the actual scores from the 2025 examinee populations, so reflect a real scenario of what would be faced if implementing CAT with the banks as they are now. The reality of transitioning to an item-level CAT for CMAS would mean an extensive item development effort would have to be implemented to attain sufficient coverage for all students.

The bank limitations also impacted the ELA MST simulations in that in an operational administration, the easier and harder item sets within a given stage would have more spread comparatively. That would contribute to more precise ability estimates of examinees routed to those respected sets. Here too, for CMAS ELA current linear tests are built to maximize precision at the cut scores. So while the spread of easier and harder item sets was possible, a more robust ELA bank established for MST would allow for improved measurement. It is also worth acknowledging that other MST design configurations could potentially yield more optimal measurement characteristics beyond the 1-2-2 MST simulation design implemented here.

One additional limitation was within ELA, in which the reduction activity attempted to balance content needs aligned to ESSA viability with the call to reduce testing time by reducing items. This approach could result in very little actual time savings, as was shown in the Grade 6 timing analysis. In that case, future attempts to shorten testing time will likely have to incorporate additional efforts towards modified test designs, which is no easy task for ELA in terms of adequately representing the depth and breadth of the standards within federal accountability.

5. Conclusion

This study evaluated two approaches for reducing testing burden while maintaining the technical quality required of CMAS under ESSA: shortening current fixed-length assessments and simulating adaptive designs aligned to the structures of the mathematics and ELA programs. Using spring 2025 operational data for grades 4 and 6, we examined comparability of results, classification agreement, measurement precision, timing, and the implications for reporting and federal peer review. The evidence clarifies what is feasible in the near term and what would require additional investment and design changes.

For mathematics, a 30% reduction in points (from 51 to 34) maintained strong comparability with full-length forms at the overall score level. Exact performance level agreement between full and reduced-length forms was about 86% in grades 4 and 6, with all disagreements adjacent, and overall score correlations were 0.985 in both grades. As anticipated, precision decreased: average conditional standard errors increased, overall coefficient alpha fell from 0.93 to 0.89 in grade 4 and from 0.92 to 0.88 in grade 6, and decision accuracy and consistency declined modestly. Precision losses were most evident for subclaim reporting. Notably, the shortened blueprints produced meaningful time savings in math, with median testing time decreasing from 142 to 100 minutes in grade 4 and from 115 to 84 minutes in grade 6.

For ELA, reductions were necessarily more modest due to the requirement to retain prose constructed responses to ensure adequate coverage of written expression and conventions. Grade 4 was reduced from 23 to 14 items and grade 6 from 24 to 19. Average overall scores remained comparable, but precision losses were larger than in math, particularly in grade 4. Exact performance level agreement was 75.8% in grade 4 (with a small number of non-adjacent shifts) and 90.1% in grade 6; overall score correlations were 0.959 and 0.991, respectively. Reliability declined across the board, with grade 4's overall coefficient alpha at 0.83 and notable reductions across reading subclaims, and decision accuracy and consistency likewise decreased. Critically, these ELA item reductions yielded little less reduction to testing time in Grade 4 and to no reduction in total testing time at Grade 6.

Adaptive simulations provide additional insights. For mathematics, constrained item-level CAT simulations aligned to CMAS blueprints produced overall classifications that matched "true" ability at rates similar to the linear tests, with exact performance level agreement around 84% for full-length and 81% for reduced-length simulations and predominantly adjacent disagreements. An unconstrained scenario, configured to stop at a fixed precision level without content balancing, delivered the highest exact agreement but underscored why such an approach is not viable for ESSA-aligned accountability. Item counts ranged from 13 to 245 depending on examinee ability and item availability, and test content could be unrepresentative; average item counts were roughly similar to full-length forms, demonstrating that, under ESSA, content balancing rather than pure measurement precision will govern test length. With appropriate constraints and adequately constructed item banks, item-level CAT can improve precision across the ability range, but the CMAS item banks would need to be more fully developed to support a transition to CAT.

For ELA, a multistage (1-2-2) design using passage-based sets yielded overall classification agreement comparable to linear testing, with modest precision benefits relative to reduced-length forms and less pronounced gains than item-level CAT in math. Exact agreement with "true" ability was 76.3% for full-length and 71.5% for reduced-length in grade 4, and 78.9% and 77.6% in grade 6; mean CSEMs increased under reduced forms in both grades. The magnitude of gains was constrained by current bank characteristics and by the requirement to maintain passage-based blueprints. More substantial benefits would require broader and deeper passage-based item banks with greater difficulty spread and potentially alternative MST configurations.

These findings carry important implications for accountability, and reporting. Any operational change must preserve blueprint coverage across the full breadth and depth of the Colorado Academic Standards for every student to satisfy ESSA and federal peer review. The unconstrained CAT scenario demonstrates that time savings achieved by ignoring content balancing are incompatible with these requirements. Subclaim reporting is the most sensitive to shortening; precision losses—especially in grade 4 ELA—may warrant limiting or qualifying such reporting to prevent misinterpretation. A transition to adaptive delivery would also necessitate rethinking reports that rely on students taking the same items within a year, including percent-correct comparisons and Evidence Statement Analyses; viable, meaningful alternatives would need to be defined and validated. As precision declines, the risk of overinterpreting small score differences increases, so changes must be accompanied by clear guidance on margins of error, appropriate inferences, and communication standards.

Realizing the potential benefits of adaptive testing under ESSA will require substantial investment in item bank development. For math item-level CAT, a multi-year plan would likely be needed to expand and rebalance banks to support content balancing across the full ability range, refine blueprints, exposure controls, and stopping rules suited to accountability, and conduct simulations with optimized banks to quantify precision gains and confirm lengths relative to different constraint settings. For ELA MST, banks would need additional passage sets to enhance stage separation; alternative MST configurations should be evaluated, and pilots implemented to study measurement benefits, student experience, and operational feasibility.

Overall, the study suggests that Colorado could consider the 30% reduced-length mathematics blueprint operationally. But careful consideration should be made with respect to the potential loss of information currently provided to stakeholders (e.g., subclaim reporting). For ELA, item count reductions alone are unlikely to substantively reduce testing time; any pursuit of shorter ELA assessments should focus on test design modifications, such as rebalancing passage lengths or counts and reconsidering task configurations while preserving ESSA-aligned coverage. Lastly, adaptive testing offers clear potential for more precise measurement across the performance continuum, but it is not a guaranteed path to shorter tests under ESSA's content-balancing requirements.

6. References

Colorado General Assembly. (2023). *Recommendations of the Technical Advisory Panel for the Colorado State Assessment Program*. Retrieved from https://leg.colorado.gov/sites/default/files/2025a_1278_signed.pdf

C.R.S. 22-11-101 et seq. (SB09-163); Colorado Department of Education Accountability resources (https://www.cde.state.co.us/accountability).

Elementary and Secondary Education Act of 1965, as amended by ESSA, 20 U.S.C. § 6311.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176.

Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009). Weighted Penalty Model for Content Balancing in CATs. Pearson Research Report. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/WeightedPenaltyModel.pdf?WT.mc_id=TMRS_Weighted_Penalty_Model_for_Content_Balancing

U.S. Department of Education, A State's Guide to the U.S. Department of Education's Peer Review of State Assessment Systems (2018).

Appendix B: Technical Advisory Panel (TAP) Feedback Summary

CMAS Testing Time Reduction Study

Technical Advisory Panel (TAP) Feedback Summary

Tradeoffs & Priorities

- **How should testing time be balanced with reliability, validity, and reporting?**
- **What must be preserved and where is there flexibility?**
- **What features should remain regardless of test length (e.g., constructed response/writing, growth)?**

- The subclaim reporting is very important to making the test data useful in targeting more specific areas of weakness/strength.
    - Given the expense of the state tests, maximizing their usage beyond solely end-of-year summative proficiency measures would be a better return on investment (e.g., maintain some instructional value with subclaim reporting).
- Writing responses need to remain in the test.
    - The data detail here is already fairly light, even cutting down on written responses could make the data even less valuable
- Reliability and validity are more important than reducing time. If test reliability is unacceptable then it ruins the value proposition.
    - High-stakes accountability requires valid and reliable measures to accredit schools and provide support through the SPF and UIP processes.
- The growth model must be able to continue functioning as normal in any test reduction scenario. There would need to be clear messaging and understanding of the need for a "hard reset" on results.

Approaches to Reducing Time

- **What are the key risks or limitations of proportional blueprint reduction, structural changes?**
- **Where do you see diminishing returns in reducing test length?**
- **What factors (technical, cost, operational) should guide decisions?**
- **Does CAT add value? Paper/pencil implications?**
- **Which approaches (design vs. operational) offer the greatest impact with the least risk?**

- Impacts to score interpretation/comparison amongst students and therefore risks to the growth model.
- As the study has been presented, cutting the tests by a third does not seem worth it. You don't gain a ton of reduction in ELA timing, and you lose substantial reporting in Math. In all cases reliability was lower, even though in some cases it was not as dramatic as others.

Approaches to Reducing Testing Time (Continued)

- Communicating the history of test reduction time may be important so stakeholders understand what has already been achieved in this area
- If the tests are shortened again, expectations for interpretation and use need to be adjusted to reflect the limited performance information and instructional applicability

Implementation and Impact

- ❖ **What changes most meaningfully reduce burden for schools?**
- ❖ **What scheduling/administration shifts matter most?**

- Moving to 2 math units, even if they had to be 90 minutes long (which looks like it would cover the 90th percentile student), I do think would make administration easier. However, you should get broader opinions because that naturally makes it harder to test a unit of ELA and math in the same day, which I suspect is common practice.
    - Adding to that, if middle school went to one non-calculator and one calculator unit that would help.
    - Conversely, if math were broken into smaller units with the same timing and most units could align 3-8, it could be more straightforward to schedule for calculator sections that our older students take.
- Keeping timing consistent grades 3-8 helps schools with all grade levels significantly.
- The less special situation groups for make-up testing, the better. (e.g. Unit 1 middle school math needs separate instructions so they cannot test with Unit 2 or Unit 3 students)
- Even with no changes to the test at all, allowing for more leniency in makeup testing groups would lessen the burden on schools. Some burden is enforced by requiring separate testing groups in the same content. (E.g. extended time students vs standard time maybe could be allowed when finishing make-up testing, since the timer is built into test now and make-up testing is already different than regularly scheduled testing)

There is no ideal scenario. What works for one grade level does not work as well for others and administration will always be a significant burden for schools, regardless. That's why clearly defining the purpose and use of the assessments and designing them accordingly makes the most sense. And then, communicating and sense-making that "why" to all stakeholders to explain the reasoning. Make an "all in" decision and stick to it, rather than piecemealing and watering down the tests so that no clear purpose is met effectively.

Appendix C: Accountability Work Group (AWG) Feedback Summary

CMAS Testing Time Reduction Study

Accountability Work Group (AWG) Feedback Summary

Tradeoffs & Priorities

- ❖ **How should testing time be balanced with reliability, validity, and reporting?**
- ❖ **What must be preserved and where is there flexibility?**
- ❖ **What features should remain regardless of test length (e.g., constructed response/writing, growth)?**

- Validity and reliability are the pieces that need to stay intact. We cannot lose that as we make shifts for testing time.
- If we shorten the timing of the tests but we are losing reliability and validity, what is the purpose? What are we doing? All of the time on assessment is wasted from that perspective.
- Losing growth data would be a detriment to the system. The tests can't be so short that growth data is lost.
- Growth as a measure from year to year must be preserved. Anything that would degrade the ability to demonstrate and prioritize growth should be avoided.
- Appreciate that we are getting more timely results, which makes it more tightly aligned to our instructional priority. Frameworks are a snapshot in the rearview mirror.
- Need better understanding of why the timeline is long between achievement and growth. Explain the timeline between achievement and growth.
- Agreement with TAP re: retaining ability to continue tracking trends.
- Reliability is a priority.
- If we were to lose trend and/or reliability, a reduction is not worth it.
- Agreement that growth is highly valued.

Approaches to Reducing Time

- ❖ **What are the key risks or limitations of proportional blueprint reduction, structural changes?**
- ❖ **Where do you see diminishing returns in reducing test length?**
- ❖ **What factors (technical, cost, operational) should guide decisions?**
- ❖ **Does CAT add value? Paper/pencil implications?**
- ❖ **Which approaches (design vs. operational) offer the greatest impact with the least risk?**

- In a school setting, shortening the time doesn't make as much of a difference as reducing the number of units.
- Tension between computer adaptive tech with costs and discussions about screen time
    - Some parents are upset about technology and screen time

Approaches to Reducing Testing Time (Continued)

- Could be a place for computer adaptive but need to balance the conversation with the broader conversation about technology.
- Can do more benchmarking with less time, increasing touch points with total time, gives a good picture about how students are growing throughout the year.
- We use this data to make decisions - only receiving aggregate scores masks the usefulness of the assessments. Lower-level information is critical for conversations with families.
- Needs to be able to talk about testing burden vs. testing accuracy.

Implementation and Impact

- ❖ **What changes most meaningfully reduce burden for schools?**
- ❖ **What scheduling/administration shifts matter most?**

- Reducing number of units would be the most helpful.
- Total time beyond the testing time (nice to see the actual testing time used)
    - Make-ups take up staff, interventionists to pull groups of 1-5 kids
    - The kids that are already absent and now they are being pulled
- Shortening could make it more manageable
- Tension between participation and make-ups and prep
- There are many misunderstandings about the assessments for parents that choose to opt out their kids.
- Testing windows - logistical burdens of testing windows.
- Aligning testing time with classroom expectation of time on task (3rd grade) - shortening which would potentially increase number of sessions
- Look at best practices in local administration ex amount of time.
- Amount of time for make-up, flexibility of make-up window, limitations for small system on technology and proctors.

Additional Considerations:

- Interesting misconception that CAT will shorten time. As a superintendent, the value of the itemized reports is huge. There is not a strong tradeoff. The reliability and validity and reporting - you may be saving 30 minutes but losing the value of the reporting and sacrificing reliability and comparability.
- Getting the message out there about reducing test time out there.
- It is unfortunate that most kids don't need the full time, but you must schedule it even though 90% of kids don't need it.
- Lean into the facts with messaging. There are a lot of misconceptions. A communication campaign about the importance of state assessments, purpose, and value. What is the actual amount of time being used for testing in comparison to a school calendar. Talk about actual instruction time and the amount of time we are taking to test.
- Recommendation on cutting time blocked for the test.
- Have a hard time getting buy-in from students and parents so less information/impact on instruction would perpetuate this challenge
- What are we solving for? Perception?
- Is the concern about testing time CMAS-specific or is it testing in general?
- TAP feedback was helpful to help with thinking
- Application of models for reductions in upper grades and applying to the younger grades
- Time concerns for younger grades
- Clarification about writing v. multiple choice, implications of typing- 6-point option reduction- are scaffolded, more to 3 points
- Information that describes the transition/relationship to P/SAT, length of test - acknowledging incoherence across these two assessments would be helpful.