

**California Department of Education Assessment Development & Administration Division**



# California Assessment of Student Performance and Progress Smarter Balanced Summative Assessments 2023–24 Technical Report

**Final Submitted May 20, 2025**

**By ETS**



**Contract No. CN220002**

[Chapter 1: Introduction 1](#_Toc192488454)

[1.1. Background 1](#_Toc192488455)

[1.2. Purposes of the Assessment 2](#_Toc192488456)

[1.3. Assessment Design 2](#_Toc192488457)

[1.3.1. Computer Adaptive Test 2](#_Toc192488458)

[1.3.2. Performance Tasks 3](#_Toc192488459)

[1.4. Intended Population 3](#_Toc192488460)

[1.5. Intended Use and Purpose of Test Scores 4](#_Toc192488461)

[1.6. Testing Window 5](#_Toc192488462)

[1.7. Significant CAASPP Developments in 2023–24 5](#_Toc192488463)

[1.7.1. Updated Statewide Testing Window 5](#_Toc192488464)

[1.7.2. Student Score Reports Redesign 5](#_Toc192488465)

[1.7.3. Composite Claims Reporting 5](#_Toc192488466)

[1.7.4. Test Delivery 6](#_Toc192488467)

[1.7.5. Accessibility Resources 6](#_Toc192488468)

[1.8. Groups and Organizations Involved with the CAASPP System 6](#_Toc192488469)

[1.8.1. California State Board of Education 6](#_Toc192488470)

[1.8.2. California Department of Education 7](#_Toc192488471)

[1.8.3. California Educators 7](#_Toc192488472)

[1.8.4. Smarter Balanced Assessment Consortium 7](#_Toc192488473)

[1.8.5. Contractors 7](#_Toc192488474)

[1.9. Systems Overview and Functionality 9](#_Toc192488475)

[1.9.1. Test Operations Management System 9](#_Toc192488476)

[1.9.2. Test Delivery System 9](#_Toc192488477)

[1.9.3. Practice and Training Tests 10](#_Toc192488478)

[1.9.4. California Educator Reporting System 10](#_Toc192488479)

[1.9.5. Test Results for California’s Assessments Website 10](#_Toc192488480)

[1.9.6. Constructed-Response Scoring Systems for ETS and Measurement   
Incorporated 11](#_Toc192488481)

[1.10. Overview of the Technical Report 11](#_Toc192488482)

[References 13](#_Toc192488483)

[Chapter 2: Overview of CAASPP Smarter Balanced Processes 14](#_Toc192488484)

[2.1. Item Development 14](#_Toc192488485)

[2.1.1. Item Format 14](#_Toc192488486)

[2.1.2. Item Specifications 14](#_Toc192488487)

[2.2. Test Assembly 15](#_Toc192488488)

[2.2.1. Test Length 15](#_Toc192488489)

[2.2.2. Test Blueprints 16](#_Toc192488490)

[2.2.3. Item Selection 16](#_Toc192488491)

[2.3. Test Administration 17](#_Toc192488492)

[2.3.1. Test Security and Confidentiality 17](#_Toc192488493)

[2.3.2. Procedures to Maintain Standardization 17](#_Toc192488494)

[2.4. Fairness and Accessibility 18](#_Toc192488495)

[2.4.1. Overview 18](#_Toc192488496)

[2.4.2. Student Accessibility Resources 18](#_Toc192488497)

[2.5. Scores 19](#_Toc192488498)

[2.5.1. Score Reporting 19](#_Toc192488499)

[2.5.2. Aggregation Procedures 19](#_Toc192488500)

[2.6. Calibration and Scaling 20](#_Toc192488501)

[2.6.1. Calibration 20](#_Toc192488502)

[2.6.2. Horizontal Scaling 21](#_Toc192488503)

[2.6.3. Vertical Scaling 21](#_Toc192488504)

[2.6.4. Vertical Scale Evaluation 22](#_Toc192488505)

[References 23](#_Toc192488506)

[Chapter 3: Item Development 25](#_Toc192488507)

[3.1. Background 25](#_Toc192488508)

[3.2. Additional Information 25](#_Toc192488509)

[Reference 26](#_Toc192488510)

[Chapter 4: Test Assembly 27](#_Toc192488511)

[4.1. Smarter Balanced Adaptive Item Selection Algorithm 27](#_Toc192488512)

[4.1.1. Content Match 28](#_Toc192488513)

[4.1.2. Information 28](#_Toc192488514)

[4.2. Simulation Study 29](#_Toc192488515)

[4.3. Test Blueprints 30](#_Toc192488516)

[4.4. Test Sections 30](#_Toc192488517)

[4.4.1. Computer Adaptive Testing Administration 30](#_Toc192488518)

[4.4.2. Performance Task Administration 31](#_Toc192488519)

[4.5. Special Version Forms 31](#_Toc192488520)

[4.5.1. Braille Form 31](#_Toc192488521)

[4.5.2. Paper–Pencil Form 32](#_Toc192488522)

[References 33](#_Toc192488523)

[Accessibility Information 34](#_Toc192488524)

[Alternative Text for Equation 4.1 34](#_Toc192488525)

[Appendix 4.A: Smarter Balanced Test Blueprints 35](#_Toc192488526)

[Appendix 4.B: Performance Task Test Length 36](#_Toc192488527)

[Appendix 4.C: Item Distribution 37](#_Toc192488528)

[Chapter 5: Test Administration 38](#_Toc192488529)

[5.1. Overview 38](#_Toc192488530)

[5.1.1. In-Person and Remote Testing 38](#_Toc192488531)

[5.2. User Roles and Standardization 39](#_Toc192488532)

[5.2.1. Local Educational Agency CAASPP Coordinator 39](#_Toc192488533)

[5.2.2. Site CAASPP Coordinator 39](#_Toc192488534)

[5.2.3. Test Administrators 40](#_Toc192488535)

[5.2.4. Instructions for Test Administration 41](#_Toc192488536)

[5.3. Local Educational Agency Training 42](#_Toc192488537)

[5.3.1. Synchronous and Asynchronous Training 43](#_Toc192488538)

[5.3.2. Videos and Guides 43](#_Toc192488539)

[5.3.3. Training for Proper Identification and Assignment of Designated Supports and Accommodations 43](#_Toc192488540)

[5.3.4. Feedback for Continuous Improvement Survey 44](#_Toc192488541)

[5.4. Accessibility Resources 45](#_Toc192488542)

[5.4.1. Accessibility Resource Categories 45](#_Toc192488543)

[5.4.2. Identification and Selection 50](#_Toc192488544)

[5.4.3. Assignment 51](#_Toc192488545)

[5.4.4. Delivery of Embedded and Non-Embedded Resources to Students 51](#_Toc192488546)

[5.4.5. Usage of Designated Supports and Accommodations 52](#_Toc192488547)

[5.5. Practice and Training Tests 63](#_Toc192488548)

[5.6. Test Security and Confidentiality 63](#_Toc192488549)

[5.6.1. The ETS Office of Testing Integrity 64](#_Toc192488550)

[5.6.2. Procedures to Maintain Standardization of Test Security 64](#_Toc192488551)

[5.6.3. Test Security Monitoring 65](#_Toc192488552)

[5.6.4. Security of Electronic Files Using a Firewall 66](#_Toc192488553)

[5.6.5. Transfer of Scores via Secure Data Exchange 66](#_Toc192488554)

[5.6.6. Data Management in the Secure Database 66](#_Toc192488555)

[5.6.7. Statistical Analysis on Secure Servers 67](#_Toc192488556)

[5.6.8. Student Confidentiality 67](#_Toc192488557)

[5.6.9. Student Test Results 67](#_Toc192488558)

[5.6.10. Security and Test Administration Incident Reporting System Process 68](#_Toc192488559)

[5.6.11. Appeals 70](#_Toc192488560)

[5.7. Processing and Scoring 73](#_Toc192488561)

[References 74](#_Toc192488562)

[Appendix 5.A: Accessibility Resource Assignment 75](#_Toc192488563)

[Chapter 6: Standard Setting 76](#_Toc192488564)

[6.1. Description 76](#_Toc192488565)

[Reference 77](#_Toc192488566)

[Chapter 7: Scoring and Reporting 78](#_Toc192488567)

[7.1. Scoring Process 78](#_Toc192488568)

[7.1.1. Structure of the Assessments 78](#_Toc192488569)

[7.1.2. Scoring Item Responses 79](#_Toc192488570)

[7.1.3. Certification of the Scoring System 81](#_Toc192488571)

[7.2. Scoring Constructed-Response Items and Extended Writing Tasks 82](#_Toc192488572)

[7.2.1. Rater Selection 82](#_Toc192488573)

[7.2.2. Training Overview 83](#_Toc192488574)

[7.2.3. Training Processes 84](#_Toc192488575)

[7.2.4. Score Summary for Condition Codes 90](#_Toc192488576)

[7.2.5. Rater Monitoring 92](#_Toc192488577)

[7.2.6. Automated Scoring 97](#_Toc192488578)

[7.3. Student Test Scores 112](#_Toc192488579)

[7.3.1. Total Test Scores 113](#_Toc192488580)

[7.3.2. Claim and Composite Claim Scores and Claim Levels (Subscores) 120](#_Toc192488581)

[7.3.3. Assessment Target Reports 123](#_Toc192488582)

[7.3.4. Theta Score Standard Errors 124](#_Toc192488583)

[7.3.5. Scale Score Standard Errors 125](#_Toc192488584)

[7.3.6. Error Band 126](#_Toc192488585)

[7.3.7. Lexile® and Quantile® Scores 126](#_Toc192488586)

[7.4. Overview of Score Aggregation Procedures 127](#_Toc192488587)

[7.4.1. Score Distributions and Summary Statistics 127](#_Toc192488588)

[7.4.2. Demographic Student Group Summaries 134](#_Toc192488589)

[7.5. Reports Produced and Scores for Each Report 136](#_Toc192488590)

[7.5.1. Online Reporting 136](#_Toc192488591)

[7.5.2. Special Cases 137](#_Toc192488592)

[7.5.3. Types of Score Reports 137](#_Toc192488593)

[7.5.4. Score Report Applications 139](#_Toc192488594)

[7.5.5. Criteria for Interpreting Test Scores 140](#_Toc192488595)

[7.5.6. Criteria for Interpreting Score Reports 140](#_Toc192488596)

[References 142](#_Toc192488597)

[Accessibility Information 145](#_Toc192488598)

[Alternative Text for Figure 7.2 145](#_Toc192488599)

[Alternative Text for Figure 7.3 145](#_Toc192488600)

[Alternative Text for Equation 7.1 146](#_Toc192488601)

[Alternative Text for Equation 7.2 146](#_Toc192488602)

[Alternative Text for Equation 7.3 146](#_Toc192488603)

[Alternative Text for Equation 7.4 146](#_Toc192488604)

[Alternative Text for Equation 7.5 146](#_Toc192488605)

[Alternative Text for Equation 7.6 146](#_Toc192488606)

[Alternative Text for Equation 7.7 146](#_Toc192488607)

[Alternative Text for Equation 7.8 147](#_Toc192488608)

[Alternative Text for Equation 7.9 147](#_Toc192488609)

[Alternative Text for Equation 7.10 147](#_Toc192488610)

[Alternative Text for Equation 7.11 147](#_Toc192488611)

[Alternative Text for Equation 7.12 147](#_Toc192488612)

[Alternative Text for Equation 7.13 147](#_Toc192488613)

[Alternative Text for Equation 7.14 147](#_Toc192488614)

[Alternative Text for Equation 7.15 147](#_Toc192488615)

[Alternative Text for Equation 7.16 147](#_Toc192488616)

[Appendix 7.A: Overall Theta Score Distribution 148](#_Toc192488617)

[Appendix 7.B: Scale Scores of Assessments 149](#_Toc192488618)

[Appendix 7.C: Summary Statistics and Performance Levels of Composite Claims 150](#_Toc192488619)

[Appendix 7.D: Demographic Student Group Summaries 151](#_Toc192488620)

[Appendix 7.E: Student Completion Conditions 152](#_Toc192488621)

[Chapter 8: Psychometric Analyses 153](#_Toc192488622)

[8.1. Overview 153](#_Toc192488623)

[8.1.1. Summary of the Analyses 153](#_Toc192488624)

[8.1.2. Samples Used for Analyses 154](#_Toc192488625)

[8.1.3. Test-Taking Rates 154](#_Toc192488626)

[8.2. Omission and Completion Analyses 155](#_Toc192488627)

[8.2.1. Omit Rates 155](#_Toc192488628)

[8.2.2. Completion Rates 155](#_Toc192488629)

[8.3. Conditional Exposure Rates of Items 156](#_Toc192488630)

[8.4. Item Parameter Estimation 157](#_Toc192488631)

[8.4.1. All Items 157](#_Toc192488632)

[8.4.2. Computer Adaptive Test Items 157](#_Toc192488633)

[8.4.3. Performance Task Items 157](#_Toc192488634)

[8.5. Response Time Analyses 158](#_Toc192488635)

[8.6. Reliability Analyses 159](#_Toc192488636)

[8.6.1. Sample for Reliability Analyses 160](#_Toc192488637)

[8.6.2. Reliability Measures 160](#_Toc192488638)

[8.6.3. Standard Error of Measurement 161](#_Toc192488639)

[8.6.4. Intercorrelations, Reliabilities, and Standard Errors of Measurement for Composite Claim Scores 162](#_Toc192488640)

[8.6.5. Student Group Reliabilities and Standard Errors of Measurement 163](#_Toc192488641)

[8.6.6. Conditional Standard Errors of Measurement 164](#_Toc192488642)

[8.6.7. Decision Classification Analyses 166](#_Toc192488643)

[8.6.8. Interrater Agreement (Constructed-Response Scoring Reliability) 168](#_Toc192488644)

[8.6.9. Agreement Between Artificial Intelligence and Human Scoring 170](#_Toc192488645)

[8.7. Validity Evidence 171](#_Toc192488646)

[8.7.1. Design of the CAASPP Smarter Balanced Summative Assessments 171](#_Toc192488647)

[8.7.2. Content 173](#_Toc192488648)

[8.7.3. Response Processes 175](#_Toc192488649)

[8.7.4. Internal Structure 176](#_Toc192488650)

[8.7.5. Relations to Other Variables 179](#_Toc192488651)

[8.7.6. Consequences of Testing 180](#_Toc192488652)

[References 181](#_Toc192488653)

[Accessibility Information 186](#_Toc192488654)

[Alternative Text for Equation 8.1 186](#_Toc192488655)

[Alternative Text for Equation 8.2 186](#_Toc192488656)

[Alternative Text for Equation 8.3 186](#_Toc192488657)

[Alternative Text for Equation 8.4 186](#_Toc192488658)

[Alternative Text for Equation 8.5 186](#_Toc192488659)

[Alternative Text for Equation 8.6 186](#_Toc192488660)

[Alternative Text for Equation 8.7 186](#_Toc192488661)

[Appendix 8.A Test-Taking Rates 187](#_Toc192488662)

[Appendix 8.B: Omission and Completion Analyses 188](#_Toc192488663)

[Appendix 8.C: Item Exposure 189](#_Toc192488664)

[Appendix 8.D: Item Response Theory Parameter Estimates 190](#_Toc192488665)

[Appendix 8.E: Response Time Analyses 191](#_Toc192488666)

[Appendix 8.F: Reliability Analyses 192](#_Toc192488667)

[Appendix 8.G: Scale Score Conditional Standard Error of Measurement Distribution 193](#_Toc192488668)

[Appendix 8.H: Analyses of Classification 194](#_Toc192488669)

[Appendix 8.I: Interrater Reliability 195](#_Toc192488670)

[Appendix 8.J: Correlations Between Content Areas 196](#_Toc192488671)

[Chapter 9: Quality-Control Procedures 197](#_Toc192488672)

[9.1. Quality Control of Test Materials 197](#_Toc192488673)

[9.1.1. Developing Assessments 197](#_Toc192488674)

[9.1.2. Test Administration Manuals 197](#_Toc192488675)

[9.1.3. Collecting Test Materials 197](#_Toc192488676)

[9.1.4. Processing Test Materials 198](#_Toc192488677)

[9.2. Quality Control of Test Administration 198](#_Toc192488678)

[9.3. Quality Control of Scoring 199](#_Toc192488679)

[9.3.1. Human Scoring 199](#_Toc192488680)

[9.3.2. Quality Control of Automated Scoring 202](#_Toc192488681)

[9.3.3. Machine-Scoring Procedures 203](#_Toc192488682)

[9.3.4. Rater Agreement Results 203](#_Toc192488683)

[9.3.5. Development of Scoring Specifications 203](#_Toc192488684)

[9.3.6. Paper–Pencil Scoring 204](#_Toc192488685)

[9.4. Quality Control of Psychometric Processes 204](#_Toc192488686)

[9.4.1. Score Verification 204](#_Toc192488687)

[9.4.2. Psychometric Analyses 204](#_Toc192488688)

[9.5. Quality Control of Reporting 205](#_Toc192488689)

[9.5.1. Exclusion of Student Scores from Summary Reports 206](#_Toc192488690)

[9.6. Quality Control of End-to-End Testing 206](#_Toc192488691)

[9.6.1. Computer-based Assessments 206](#_Toc192488692)

[9.6.2. Paper–Pencil Tests 206](#_Toc192488693)

[References 207](#_Toc192488694)

[Chapter 10: Historical Comparisons 208](#_Toc192488695)

[10.1. Student Performance 208](#_Toc192488696)

[10.1.1. Cross-Sectional Comparisons on the Overall Assessments 208](#_Toc192488697)

[10.1.2. Longitudinal Comparisons on the Overall Groups 209](#_Toc192488698)

[10.2. Test Characteristics 211](#_Toc192488699)

[Appendix 10.A: Cross-Sectional Comparisons of the Overall Group and Student Groups on the Overall Assessments 212](#_Toc192488700)

[Appendix 10.B: Longitudinal Comparison of the Overall Group and Student Groups on the Overall Assessments 213](#_Toc192488701)

[Appendix 10.C: Comparison of Test Characteristics 214](#_Toc192488702)

[Chapter 11: Paper–Pencil Forms 215](#_Toc192488703)

[11.1. Overview 215](#_Toc192488704)

[11.2. Universal Tools, Designated Supports, and Accommodations 216](#_Toc192488705)

[11.3. Calibration and Scaling 216](#_Toc192488706)

[11.3.1. Initial Calibration 217](#_Toc192488707)

[11.3.2. Anchor Item Evaluation 217](#_Toc192488708)

[11.3.3. Final Calibration 217](#_Toc192488709)

[11.4. Scoring 218](#_Toc192488710)

[11.4.1. Number of Students Who Took the 2023–24 Smarter Balanced Paper–‍Pencil Assessments 218](#_Toc192488711)

[11.4.2. Item Response Theory Parameter Values 218](#_Toc192488712)

[References 219](#_Toc192488713)

[Appendix 11.A: Item Response Theory Parameter Estimates 220](#_Toc192488714)

[Chapter 12: Continuous and Systematic Improvement 221](#_Toc192488715)

[12.1. 2023–24 Feedback for Continuous Improvement Survey 221](#_Toc192488716)

[12.2. Communications 221](#_Toc192488717)

[12.2.1. Updated Program Website 221](#_Toc192488718)

[12.3. Reporting 221](#_Toc192488719)

[12.4. Accessibility Resources 222](#_Toc192488720)

[Reference 223](#_Toc192488721)

List of Tables

[Acronyms and Initialisms Used in the *CAASPP Smarter Balanced Technical Report* x](#_Toc192152399)

[Table 5.1 Number of Students by Assessment Location 38](#_Toc192152400)

[Table 5.2 Summary of Accommodations and Designated Supports Used by Students—  
ELA 54](#_Toc192152401)

[Table 5.3 Summary of Accommodations and Designated Supports Used by Students—Mathematics 59](#_Toc192152402)

[Table 5.4 Types of Appeals 69](#_Toc192152403)

[Table 5.5 Number and Types of Incidents Submitted in STAIRS—ELA 71](#_Toc192152404)

[Table 5.6 Number and Types of Incidents Submitted in STAIRS—Mathematics 72](#_Toc192152405)

[Table 5.7 Number of Appeals Approved in STAIRS 73](#_Toc192152406)

[Table 5.8 Number of Appeals Rejected in STAIRS 73](#_Toc192152407)

[Table 7.1 Machine-Scored Computer-based Item Types 80](#_Toc192152408)

[Table 7.2 Scoring Condition Codes 91](#_Toc192152409)

[Table 7.3 Number of Human-Scored CR Items Flagged, by Content Area and Grade   
Level 96](#_Toc192152410)

[Table 7.4 Number of Items Eligible for Model Training, by Grade Level and Content Area 99](#_Toc192152411)

[Table 7.5 Initial Model Evaluation Criteria 101](#_Toc192152412)

[Table 7.6 Demographic Variables and Categories for Student Groups 102](#_Toc192152413)

[Table 7.7 Secondary Validation Criteria 102](#_Toc192152414)

[Table 7.8 Summary of Secondary Validation Results by Grade Level and Content Area 103](#_Toc192152415)

[Table 7.9 Summary of Live Training and Validation Results by Grade Level and Content Area 104](#_Toc192152416)

[Table 7.10 Flags for Alert Responses and Nonscorable Cases 106](#_Toc192152417)

[Table 7.11 Human–Automated Scoring Agreement for ELA Short-Answer Items on Initial and Secondary Validation Samples by Grade Level 108](#_Toc192152418)

[Table 7.12 Human–Automated Scoring Agreement for ELA Essay Items on Initial and Secondary Validation Samples by Grade Level 109](#_Toc192152419)

[Table 7.13 Human–Automated Scoring Agreement for Mathematics Items on Initial and Secondary Validation Samples by Grade Level 110](#_Toc192152420)

[Table 7.14 Human–Automated Scoring Agreement for ELA Short-Answer Items on Live Validation Sample by Grade Level 111](#_Toc192152421)

[Table 7.15 Human–Automated Scoring Agreement for ELA Essay Items on Live Validation Sample by Grade Level 111](#_Toc192152422)

[Table 7.16 Human–Automated Scoring Agreement for Mathematics Items on Live Validation Samples by Grade Level 112](#_Toc192152423)

[Table 7.17 Theta of Lowest and Highest Obtainable Scores 115](#_Toc192152424)

[Table 7.18 Treatment of Incomplete Assessments 116](#_Toc192152425)

[Table 7.19 Lowest and Highest Obtainable Scale Scores 118](#_Toc192152426)

[Table 7.20 Theta Thresholds for Achievement Levels 119](#_Toc192152427)

[Table 7.21 Scale Score Ranges for Achievement Levels 120](#_Toc192152428)

[Table 7.22 Claims Identified for ELA 121](#_Toc192152429)

[Table 7.23 Claims Identified for Mathematics 121](#_Toc192152430)

[Table 7.24 Operational Mean and SD of Theta and Scale Scores for the Standard Assessment 127](#_Toc192152431)

[Table 7.25 Operational Mean and SD of Theta and Scale Scores for Students Taking the Embedded Field Test PT Form Versions 128](#_Toc192152432)

[Table 7.26 Achievement Levels for Students Who Took Only Operational Items for the Smarter Balanced for ELA 129](#_Toc192152433)

[Table 7.27 Achievement Levels for Students Who Took Only Operational Items for the Smarter Balanced for Mathematics 131](#_Toc192152434)

[Table 7.28 Achievement Levels for Students Who Took Embedded Field Test–Only PTs for the Smarter Balanced for Mathematics 133](#_Toc192152435)

[Table 7.29 Demographic Student Groups to Be Reported 135](#_Toc192152436)

[Table 8.1 Minimum Number of Items for a Complete Composite Claim Score in the Adjusted, Shortened-Form Blueprints 156](#_Toc192152437)

[Table 8.2 PT Field Test Minimum Number of Items for a Complete Composite Claim Score If Assessment Includes Field Test PT Items 156](#_Toc192152438)

[Table 8.3 Summary Statistics for Scale Scores and Theta Scores, Reliability, and SEMs 162](#_Toc192152439)

[Table 8.4 Scale Score CSEMs at Achievement-Level Thresholds 164](#_Toc192152440)

[Table 8.5 Mean CSEMs for Each Achievement Level 165](#_Toc192152441)

[Table 8.6 Decision Accuracy for Reaching an Achievement Level 167](#_Toc192152442)

[Table 8.7 Decision Consistency for Reaching an Achievement Level 167](#_Toc192152443)

[Table 8.8 Frequencies of Ratings 169](#_Toc192152444)

[Table 8.9 Correlations for All Students 178](#_Toc192152445)

[Table 9.1 Summary of Characteristics of MI Human Raters Scoring CAASPP Smarter Balanced Assessments 200](#_Toc192152446)

[Table 9.2 Rater Qualification Standard for Agreement with Correct Scores 201](#_Toc192152447)

[Table 11.1 Number of Students with Valid Scores from CAASPP Smarter Balanced Summative PPTs 215](#_Toc192152448)

List of Figures

[Figure 2.1 Vertical scaling 22](#_Toc182989103)

[Figure 7.1 Writing traits 87](#_Toc182989104)

[Figure 7.2 PEG engine overview 98](#_Toc182989105)

[Figure 7.3 Hybrid scoring response routing rules 105](#_Toc182989106)

[Figure 7.4 Percentage of achievement levels in ELA, operational assessments 130](#_Toc182989107)

[Figure 7.5 Percentage of achievement levels in mathematics, operational assessments 132](#_Toc182989108)

[Figure 7.6 Percentage of achievement levels in mathematics, embedded field test PTs 134](#_Toc182989109)

Acronyms and Initialisms Used in the *CAASPP Smarter Balanced Technical Report*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| 2PL | two-parameter logistic |
| AD | Assessment Development |
| ADEL | adult English learner |
| AERA | American Educational Research Association |
| AI | artificial intelligence |
| AIR | American Institutes for Research |
| APA | American Psychological Association |
| ASL | American Sign Language |
| CAA | California Alternate Assessment |
| CAASPP | California Assessment of Student Performance and Progress |
| CAI | Cambium Assessment, Inc. |
| CALPADS | California Longitudinal Pupil Achievement Data System |
| CalTAC | California Technical Assistance Center |
| CAT | computer adaptive test |
| *CCR* | *California Code of Regulations* |
| CCSS | Common Core State Standards |
| CDE | California Department of Education |
| CDS | county/district/school |
| CERS | California Educator Reporting System |
| CI | confidence interval |
| CR | constructed response |
| CRESST | Center for Research on Evaluation, Standards, & Student Testing |
| CSEM | conditional standard error of measurement |
| DEI | Data Entry Interface |
| *DFA* | *Directions for Administration* |
| DIF | differential item functioning |
| EAP | Early Assessment Program |
| *EC* | *Education Code* |
| EL | English learner |
| ELA | English language arts/literacy |
| ELPAC | English Language Proficiency Assessments for California |
| EO | English only |
| eSKM | Enterprise Score Key Management |
| FT | field test |
| GPCM | generalized partial credit model |
| HAT | hybrid adaptive test |
| HOSS | highest obtainable scale score |
| HOT | highest obtainable theta |
| IEP | individualized education program |
| IFEP | initial fluent English proficient |
| IRT | item response theory |
| ISAAP | Individual Student Assessment Accessibility Profile |
| LDA | Latent Dirichlet Allocation |
| LEA | local educational agency |
| LOSS | lowest obtainable scale score |
| LOT | lowest obtainable theta |
| LSA | Latent Semantic Analysis |
| MC | multiple choice |
| MI | Measurement Incorporated |
| MLE | maximum likelihood estimation |
| NAEP | National Assessment of Educational Progress |
| NCME | National Council on Measurement in Education |
| NEL | not tested English learner |
| OTI | Office of Testing Integrity |
| PAR | Psychometric Analysis & Research |
| PEG | Project Essay Grade |
| PIN | problem item notification |
| PISA | Program for International Student Assessment |
| PPT | paper–pencil test |
| PT | performance task |
| QA | quality assurance |
| QWK | quadratic-weighted kappa |
| RFEP | reclassified fluent English proficient |
| SBE | State Board of Education |
| SCOE | Sacramento County Office of Education |
| SD | standard deviation |
| SEM | standard error of measurement |
| SFTP | secure file transfer protocol |
| SRC | Scoring Resource Center |
| SS | scale score |
| SSID | Statewide Student Identifier |
| SSR | Student Score Report |
| STAIRS | Security and Test Administration Incident Reporting System |
| TBD | to be determined |
| TCC | test characteristic curve |
| TDS | Test Delivery System |
| TIF | test information function |
| TIPS | Technology and Information Processing Services |
| TOMS | Test Operations Management System |
| UAT | user acceptance testing |
| *USC* | *United States Code* |
| VSC | Virtual Scoring Center |
| wABC | weighted Area Between the Curves |
| WER | writing extended response |

## Introduction

This chapter provides an overview of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced testing program, including background information, the purpose of the assessment, the intended population, and organizations and systems involved.

### Background

In October 2013, Assembly Bill 484 established CAASPP as the new student assessment system that replaced the Standardized Testing and Reporting program. The primary purpose of the CAASPP System of assessments is to assist teachers, administrators, and students and their parents/guardians by promoting high-quality teaching and learning through the use of a variety of item types and assessment approaches. These assessments provide the foundation for the state’s school accountability system.

The Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics were administered during the 2023–24 CAASPP administration as a result of California’s participation in the Smarter Balanced Assessment Consortium. This technical report describes the results of that administration.

In 2023–24, the CAASPP System comprised the following assessments:

* Smarter Balanced assessments and tools
* Summative Assessments—Computer-based assessments for ELA and mathematics in grades three through eight and grade eleven
* Interim Assessments—Optional resources developed for grades three through eight and grade eleven designed to inform and promote teaching and learning by providing information that can be used to monitor student progress toward mastery of the Common Core State Standards (CCSS) that may be administered to students at any grade level
* Tools for Teachers—Professional development materials and instructional resources designed to help teachers use formative assessment processes for improved teaching and learning in all grade levels
* California Alternate Assessments (CAAs) for ELA and mathematics in grades three through eight and grade eleven for students with significant cognitive disabilities
* Science assessments in grades five and eight and the high school grade band (that is, grade ten, eleven, or twelve, if the student is not repeating grade twelve; these are the California Science Test and the CAA for Science)
* The California Spanish Assessment, optional for eligible students in grades three through eight and the high school grade band and designed to measure a student’s literacy in Spanish language arts inclusive of reading, writing mechanics, and listening, as well as to serve as a high school measure suitable to be used in part for the California Seal of Biliteracy

The CAASPP Smarter Balanced assessments are presented as computer-based assessments. Braille, large-print, and standard paper–pencil test (PPT) versions of the Smarter Balanced assessments are made available to individual students within a local educational agency (LEA) whose need to take a PPT is documented in a student’s individualized education program (IEP) or Section 504 plan. Students who repeatedly experience difficulty accessing the computer-based assessments because of technical issues that cannot be resolved within two weeks may be allowed to take a standard PPT, upon approval by the California Department of Education (CDE). The PPT versions are fixed forms (that is, an assessment where students are given a fixed set of items irrespective of the student’s responses or ability) that also include the components of the computer-based assessment such as constructed-response (CR) items and performance tasks (PTs).

More background information about the CAASPP System can be found on the CAASPP Description – *CalEdFacts* web page on the CDE website.

### Purposes of the Assessment

The purposes of the Smarter Balanced assessment system are to provide teachers with information and the tools they need to improve teaching and learning; and to prepare students for college and career. The Smarter Balanced Summative Assessments are comprehensive, end-of-year assessments of grade-level learning that measure students’ progress toward college and career readiness. These assessments, which are aligned with the California CCSS for ELA and mathematics, form one component of the Smarter Balanced assessment system.

### Assessment Design

Smarter Balanced Summative Assessments are composed of two required components: a computer adaptive test (CAT) and a PT. A student’s final scale score is calculated by combining the student’s responses to both components.

#### Computer Adaptive Test

The computer-adaptive portion of the assessment is designed to present items of difficulty to match the ability of each student, as indicated by the responses the student provided to previous assessment items. By adapting to the student’s ability as the assessment is being taken, the CAT presents an individually tailored set of items that is appropriate for each student. As a result, it provides more accurate scores for all students across the full range of the achievement continuum. Compared with a fixed-form assessment—that is, an assessment where all students are given the same items, regardless of their responses or ability—a CAT requires fewer items to obtain an equally precise estimate of a student’s ability.

At the beginning of the assessment, the Test Delivery System (TDS) assumes that the student is of average ability and presents an item that is appropriate for an average student. During the assessment, if a student gives an incorrect answer, the TDS will follow up with an easier item or a group of items; if the student answers correctly, the next item or next group of items will be slightly more difficult. As the adaptive assessment continues, the next item or group of items are based on the student’s answers to all previous items—the student’s responses to the current item and previous items determine the pathway to a subsequent item.

Because the answers on items used to estimate the student’s ability are machine-scored, the student’s performance on the items already administered is known immediately, and the successive items are selected to adapt to the estimated ability of the student. The CAT selects items based on a student’s responses, scores the responses, and revises its estimate of the student’s ability. This process continues until the test content outlined in the assessment’s blueprint is covered.

The CAT requires a large pool of assessment items statistically calibrated on a common scale to cover the ability range. For the Smarter Balanced Summative Assessments, the test item statistics were obtained mainly from the spring 2013–14 field test. Each year, new items are field-tested and added to the Smarter Balanced item pools.

#### Performance Tasks

The PT is a nonadaptive portion of a Smarter Balanced content-area assessment designed to provide students with an opportunity to demonstrate their ability to apply knowledge and higher-order thinking skills to explore and analyze a complex, real-world scenario. PTs are assigned to students randomly. (Refer to section [*4.1 Smarter Balanced Adaptive Item Selection Algorithm*](#_Smarter_Balanced_Adaptive) for additional information about how items are assigned to students.)

Some PT responses are machine-scored and others are human-scored. Scores are combined with CAT results for the student’s final score.

### Intended Population

All students enrolled in grades three through eight and grade eleven are required to take part in the Smarter Balanced Summative Assessments unless they are eligible to participate in the alternate assessments (*California Code of Regulations*, Title 5 [5*CCR*] Education, Division 1, Chapter 2, Subchapter 3.75, Article 1, Section 851.5). English learner (EL) students who are in their first 12 months of attending school in the United States are exempt from taking the ELA portion of the assessment. EL students are defined as follows:

English learner students are those students for whom there is a report of a primary language other than English on the state-approved Home Language Survey **and** who, on the basis of the state approved oral language (grades kindergarten through grade twelve) assessment procedures and literacy (grades three through twelve only), have been determined to lack the clearly defined English language skills of listening comprehension, speaking, reading, and writing necessary to succeed in the school’s regular instructional programs.[[1]](#footnote-2)

EL students within their first 12 months of enrollment in a US school who choose to participate in taking the ELA assessment are included in the calculation of the percentage of students tested, but their scores are excluded from all aggregated calculations so long as the student record has an include indicator of “E” and the condition code “NEL” (not tested English learner). The condition code “NEL” is automatically assigned on the basis of the following criteria when the student’s English language acquisition status is “EL,” the student has been enrolled in a US school for less than twelve months, and one of the following criteria is true:[[2]](#footnote-3)

* Student was enrolled and did not test.
* *or* Student tested but did not meet the criteria for attemptedness of “Y” (“Yes”).
* *or* Assessment is a force-complete.

For students with significant cognitive disabilities, the decision to administer the Smarter Balanced Summative Assessments or the CAAs is made by their IEP team. Parents/‌Guardians may submit a written request to have their child exempted from taking any or all parts of the Smarter Balanced Summative Assessments or CAAs.

Students whose parents/guardians submit a written request are exempted from taking the assessments (*Education Code* [*EC*] Section 60615). Additionally, students who were not tested because of a medical emergency are also exempt.

### Intended Use and Purpose of Test Scores

The results of assessments within the CAASPP System are used for two primary purposes as described in *EC* sections 60602.5(a) and (a)(4). (Excerpted from the *EC* Section 60602 web page.)

60602.5(a) It is the intent of the Legislature in enacting this chapter to provide a system of assessments of pupils that has the primary purposes of assisting teachers, administrators, and pupils and their parents; improving teaching and learning; and promoting high-quality teaching and learning using a variety of assessment approaches and item types. The assessments, where applicable and valid, will produce scores that can be aggregated and disaggregated for the purpose of holding schools and local educational agencies accountable for the achievement of all their pupils in learning the California academic content standards.

60602.5(a)(4) Provide information to pupils, parents and guardians, teachers, schools, and local educational agencies on a timely basis so that the information can be used to further the development of the pupil and to improve the educational program.

Therefore, the two primary purposes of an assessment within the CAASPP System are the following:

1. To communicate students’ progress in achieving the state’s academic standards to students, parents/guardians, and teachers
2. To inform decisions that teachers and administrators make about improving the educational program

Sections 60602.5(c) and (d) provide additional information regarding use and purpose of test scores for the system of assessments:

60602.5(c) It is the intent of the Legislature that parents, classroom teachers, other educators, pupil representatives, institutions of higher education, business community members, and the public be involved, in an active and ongoing basis, in the design and implementation of the statewide pupil assessment system and the development of assessment instruments.

60602.5(d) It is the intent of the Legislature, insofar as is practically feasible and following the completion of annual testing, that the content, test structure, and test items in the assessments that are part of the statewide pupil assessment system become open and transparent to teachers, parents, and pupils, to assist stakeholders in working together to demonstrate improvement in pupil academic achievement. A planned change in annual test content, format, or design should be made available to educators and the public well before the beginning of the school year in which the change will be implemented.

### Testing Window

The Smarter Balanced Summative Assessments for grades three through eight and grade eleven are administered within a testing window pursuant to 5 *CCR*, sections 855(a)(1), 855(a)(2), 855(b), and 855(c). The state testing window for the 2023–24 CAASPP Smarter Balanced Summative Assessments started on January 9, 2024, and ended on June 28, 2024.

A student could take the CAASPP Smarter Balanced within the testing window over as many days as required to meet the student’s needs (5 *CCR* Section 855[a]). Similar to other CAASPP assessments, the Smarter Balanced assessments were untimed for test takers. Refer to subsection [*2.2.1 Test Length*](#_Test_Length) for more information about test length.

### Significant CAASPP Developments in 2023–24

#### Updated Statewide Testing Window

5 *CCR* Section 855(a)(2) was amended to shift the end of the statewide testing window to June 30 (from July 15). For the 2023–24 test administration window, because June 30, 2024, occurred on a weekend, the end of the testing window was June 28, 2024.

#### Student Score Reports Redesign

Redesigned Student Score Reports (SSRs) were made available for the 2023–24 test administration. Changes included the following:

* + - 1. SSR formats are PDF and HTML. For an HTML SSR, an LEA or parent or student portal vendor provided a link to a parent/guardian.
      2. Where applicable, results of a science assessment were included in the same PDF SSR as the results of the ELA and mathematics assessments.
      3. All SSRs included comparisons to average student performance.
      4. Writing extended-response scores were added to the ELA SSR.
      5. Lexile and Quantile measures were added to the reporting of ELA and mathematics results.

Additionally, Arabic was added as an available language.

#### Composite Claims Reporting

Because Smarter Balanced suggested states may report composite claim scores for ELA and mathematics (Smarter Balanced, 2023), California has started reporting these composite claim scores as follows:

* ELA Composite Claim 1: Reading and listening
* ELA Composite Claim 2: Writing and research
* Mathematics Composite Claim 1: Concepts and procedures
* Mathematics Composite Claim 2: Problem solving; communicating reasoning; and modeling and data analysis

#### Test Delivery

##### Changes to the Test Administrator Interface

The Test Administrator Interface was updated to a cleaner, more user-friendly appearance. This included a new functionality that allowed the test administrator to pin information for specific students to the top of the screen for monitoring.

##### Changes to Ending the Assessment in the Test Delivery System

The process for ending the assessment was streamlined. After the last question was presented, students selected [**Next**] (instead of [**End Test**]) to reach the review screen, which included the [**Submit Test**] button.

#### Accessibility Resources

The following accessibility resource–related updates were made:

* The definition of the non-embedded medical supports designated support was updated to allow Bluetooth hearing aids.
* The definition of the non-embedded amplification designated support was amended to remove noise buffers and white noise machines.
* Word prediction was added as an embedded accommodation for the Smarter Balanced for ELA and mathematics.
* A non-embedded printed copy of the Smarter Balanced for ELA and mathematics oral test directions in English, created by Smarter Balanced, was added as a designated support that may be provided to a student.
* Non-embedded translated test directions in American Sign Language for the Smarter Balanced for ELA and mathematics were added as designated supports that may be provided to a student.
* The embedded Spanish stacked–dual language translation designated support was updated to be activated using a toggle mode.

### Groups and Organizations Involved with the CAASPP System

#### California State Board of Education

The California State Board of Education (SBE) is the state agency that establishes educational policy for kindergarten through grade twelve in the areas of standards, instructional materials, assessment, and accountability. The SBE adopts textbooks for kindergarten through grade eight, adopts regulations to implement legislation, and has the authority to grant waivers of the *EC*.

In addition to adopting the rules and regulations for itself, its appointees, and California’s public schools, the SBE is also the state educational agency responsible for overseeing California’s compliance with programs that meet the requirements of the federal Every Student Succeeds Act that provides multiple measures of the academic performance and progress of schools on a variety of academic metrics (CDE, 2024d).

#### California Department of Education

The CDE oversees California’s public school system, which is responsible for the education of more than 5,800,000 children and young adults in more than 11,000 schools.[[3]](#footnote-4) California aims to provide a world-class education for all students, from early childhood to adulthood. The CDE serves the state by innovating and collaborating with educators, school staff, parents/guardians, and community partners which together, as a team, prepare students to live, work, and thrive in a highly connected world.

Within the CDE, it is the Instruction, Measurement, & Administration Branch that oversees programs promoting improved student achievement. Programs include oversight of statewide assessments and the collection and reporting of educational data (CDE, 2024c).

#### California Educators

A variety of California educators, including teachers and school administrators—who were selected on the basis of their qualifications, experiences, demographics, and geographic locations—were invited to participate in various aspects of the assessment process. This included defining the purpose and scope, test design, item development, and standard setting. In 2023–24, California educators were involved in scoring of the Smarter Balanced Summative Assessment CR items.

#### Smarter Balanced Assessment Consortium

The Smarter Balanced Assessment Consortium is a public agency governed by a consortium of states, of which California is a member. The Consortium created a computer-based assessment system aligned to the CCSS that is composed of year-end summative assessments and optional interim assessments (Smarter Balanced, n.d.). Smarter Balanced provided the collection of test items in the item bank as well as provided access to Tools for Teachers, which provided an online collection of resources to help teachers improve classroom-based assessment practices.

Finally, Smarter Balanced developed and hosted the California Educator Reporting System (CERS), which was used to report both summative and interim test results to educators.

#### Contractors

A number of organizations contribute to the success of the CAASPP Smarter Balanced Summative Assessments.

##### Primary Testing Contractor—ETS

The CDE and the SBE contract with ETS to administer, and report the CAASPP Smarter Balanced for ELA and mathematics. As the primary testing contractor, ETS has overall responsibility for working with the CDE to implement and maintain an effective assessment system and coordinating the ETS work with its subcontractors.

Activities conducted directly by ETS include, but are not limited to, the following:

* Providing management of the program activities
* Supporting and training county offices of education, LEAs, and direct funded charter schools
* Constructing, producing, and controlling the quality of PPT booklets and related test materials
* Developing processes and scripts associated with remote testing
* Hosting and maintaining a website with resources for LEA CAASPP coordinators
* Developing, hosting, and providing support for the Test Operations Management System (TOMS)
* Supporting CERS
* Processing student test assignments
* Processing orders and shipment of test materials
* Producing and distributing score reports electronically
* Developing a summary score reporting website that can be viewed by the public
* Completing all psychometric procedures
* Providing a tiered help desk support system for LEAs

##### Subcontractor—Cambium Assessment, Inc.

ETS also monitors and manages the work of Cambium Assessment, Inc. (CAI), subcontractor to ETS for the CAASPP System of computer-based assessments. Activities conducted by CAI include

* providing the CAI proprietary TDS, including the Student Testing Interface, Test Administrator Interface, secure browser, and practice and training tests;
* hosting and providing support for its TDS, a component of the overall CAASPP Assessment Delivery System;
* hosting and providing support for the Data Entry Interface, the web browser–based application that, for the operational administration of the Smarter Balanced Summative Assessments, allows users to enter student responses;
* scoring machine-scorable items; and
* providing high-level technology help desk support to LEAs for technology issues directly related to the TDS.

##### Subcontractor—Sacramento County Office of Education

ETS contracted with the Sacramento County Office of Education to manage all activities associated with educator recruitment, training, and outreach, including the following:

* Supporting and training county offices of education, LEAs, and charter schools
* Developing informational materials
* Recruiting and providing logistics for educator meetings
* Producing test administration scripts

##### Subcontractor—Measurement Incorporated

ETS monitors and manages the work of Measurement Incorporated (MI), a subcontractor to ETS for the CAASPP System. MI uses its automated scoring system to score the CR items for the CAASPP Smarter Balanced Online Summative Assessments.

### Systems Overview and Functionality

#### Test Operations Management System

TOMS is the password-protected, web-based system used by LEAs to manage all aspects of CAASPP testing. TOMS serves various functions, including, but not limited to, the following:

* Managing test administration windows
* Assigning and managing CAASPP online user roles
* Managing student test assignments and accessibility resources
* Ordering test materials
* Viewing and downloading reports
* Reporting security incidents
* Providing a platform for authorized user access to secure materials, such as CAASPP *Directions for Administration*, student data and results, CAASPP user information, and access to the CAASPP Security and Test Administration Incident Reporting System/Appeals process

TOMS receives student enrollment data and LEA and school hierarchy data from the California Longitudinal Pupil Achievement Data System (CALPADS) via daily feed. CALPADS is “a longitudinal data system used to maintain individual-level data including student demographics, course data, discipline, assessments, staff assignments, and other data for state and federal reporting.”[[4]](#footnote-5)

LEA staff involved in the administration of the CAASPP—such as LEA CAASPP coordinators, site CAASPP coordinators, test administrators, and test examiners—are assigned varying levels of access to TOMS. For example, only an LEA CAASPP coordinator is given permission to assign and manage user roles; a test administrator or test examiner cannot download student reports. A description of user roles is explained more extensively in the *2023–24 CAASPP Online Test Administration Manual* (CDE, 2024b).

#### Test Delivery System

The TDS is the means by which the statewide computer-based assessments are delivered to students. Components of the TDS include

* the Test Administrator Interface, the web browser–based application that allows test administrators to activate student assessments and monitor student testing;
* the Student Testing Interface, on which students take the assessment using the secure browser; and
* the secure browser, the computer-based application through which the Student Testing Interface may be accessed. (The secure browser prevents students from accessing other applications during testing.)

#### Practice and Training Tests

All California testing programs have practice and training tests to inform educators, parents/‌guardians, and students about the individual assessments. The practice and training tests were provided to LEAs to prepare students and LEA staff for administration of the Smarter Balanced Summative Assessments. These tests simulated the experience of the Smarter Balanced Summative Assessments computer-based assessments. Unlike the summative assessments, the practice and training tests did not gauge student success on the operational assessment, or produce scores. Students, teachers, and the public could access the practice tests and training tests using a web browser, although accessing them through the secure browser permitted students to test with different embedded accommodations, such as text-to-speech; and to try out different assistive technology. When remote testing was added as an optional means of test administration, the practice and training tests permitted test administrators and students to practice using the remote monitoring and communication features.

The purpose of the training tests is to allow students and test administrators to quickly become familiar with interacting with the user interface, different item types, and components of the TDS as well as with the process of starting and completing a testing session.

The purpose of the practice tests is to allow students and test administrators to experience a grade-level assessment, grade-specific items, and difficulty levels; and become familiar with the format and structure of an operational assessment.

A purpose of both the practice and training tests is to provide an opportunity for educators to assign embedded designated supports and accommodations and determine how they worked for their students prior to using the resources in an operational test setting.

#### California Educator Reporting System

CERS is the system used by LEAs to view student results from CAASPP testing as they became available. The primary purpose of CERS is to provide educators and administrators with access to timely assessment results for individual students and groups of students.

CERS allows educators to view their students’ test results at the individual student level and at the aggregate level using grouping and other features. For example, educators can create customized groups from assigned student groups based on student demographic information or other characteristics of their choosing. The student results sent to CERS are appropriate for analysis of assessment results for use in informing instruction and local planning.

#### Test Results for California’s Assessments Website

The Test Results for California’s Assessments website is used by educators, families, researchers, and interested members of the public to view aggregate results from the Smarter Balanced Summative Assessments. The primary purpose of the Test Results for California’s Assessments website is to provide users with access to results data for groups of students and to allow comparison of test result data for various student groups. Test scores for a given grade level are aggregated at the school, LEA or direct funded charter school, county, and state levels. The aggregate scores are generated for selected student groups of interest (for example, gender, ethnicity, economic status, migrant status, and disability status) and for the total population.

#### Constructed-Response Scoring Systems for ETS and Measurement Incorporated

CR items from the TDS were routed to the MI CR scoring systems based on the agreement between ETS and MI that MI conduct all CR item scoring tasks for the 2023–24 test administration. CR items were scored by certified raters. A small percentage of CR items were deemed appropriate to be scored by the artificial intelligence (AI) system and were routed for both AI scoring and human scoring for the purpose of producing agreement samples. More information regarding scoring of CR items is available in [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting).

Targeted efforts were made to hire California educators for human-scoring opportunities. Hired raters were provided in-depth training and certified before starting the human-scoring process. Human raters were organized under a scoring leader and provided Smarter Balanced scoring materials such as anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes for unscorable responses within the interface. The quality-control processes for CR scoring are explained further in [*Chapter 9: Quality-Control Procedures*](#_Quality_Control_Procedures).

### Overview of the Technical Report

This technical report addresses the characteristics of the CAASPP Smarter Balanced Summative Assessments administered in spring 2024. The technical report contains 11 additional chapters as follows:

* [Chapter 2](#_Overview_of_CAASPP_1) presents an overview of the processes involved in a testing cycle for a Smarter Balanced Summative Assessment. This includes test administration, generation of test scores, and dissemination of score reports.
* [Chapter 3](#_Item_Development_1) discusses the procedures followed during the development of Smarter Balanced items to help ensure valid interpretation of test scores.
* [Chapter 4](#_Test_Assembly) discusses the content and psychometric criteria that guide the construction of the Smarter Balanced Summative Assessments.
* [Chapter 5](#_Test_Administration) details the processes involved in the administration of the 2023–24 Smarter Balanced Summative Assessments. It also describes the procedures followed by ETS to maintain test security throughout the test administration process.
* [Chapter 6](#_Standard_Setting) discusses the standard setting process outlined by Smarter Balanced.
* [Chapter 7](#_Scoring_and_Reporting) summarizes the types of scores and score reports that are produced at the end of each administration of the Smarter Balanced Summative Assessments.
* [Chapter 8](#_Psychometric_Analyses) summarizes the results of the analyses performed on the data resulting from the 2023–24 administration. These include
* item response theory parameters,
* omission and completion analyses,
* conditional exposure analyses,
* reliability analyses that include assessments of the reliability of test scores and composite claim scores for the population as a whole and for selected student groups,
* consistency and accuracy of the achievement-level classifications,
* interrater reliability statistics for the human-scored items and statistics showing the agreement of AI scoring with human scoring, and
* procedures designed to ensure the validity of score uses and interpretations.
* [Chapter 9](#_Quality_Control_Procedures) highlights the quality-control processes used at various stages of administration of the Smarter Balanced assessments.
* [Chapter 10](#_Historical_Comparisons) presents cross-sectional and longitudinal historical comparisons of the overall assessments for all students and selected student groups. Descriptions and data are provided on the basis of student performance and test characteristics.
* [Chapter 11](#_Paper–Pencil_Forms) provides a summary of test assembly, test administration, calibration, and scaling procedures that are specifically applied to the PPTs; and the number of students who took PPTs instead of the computer-based assessments. Note, however, that psychometric analyses were not conducted because of the small sample size—on average, about 12 students took a PPT, with a minimum of 7 students and a maximum 17 students per assessment across grade levels and content areas.
* [Chapter 12](#_Continuous_and_Systematic) discusses the various procedures used to gather information to improve the CAASPP program and the Smarter Balanced assessments, as well as strategies to implement possible improvements.

### References

*California* *Code of Regulations*,Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2, Section 851.5. (n.d.).

California Department of Education. (2024a). *2023–24 CAASPP student data file layout*. Sacramento, CA: California Department of Education.

California Department of Education. (2024b). *CAASPP online test administration manual*. Sacramento, CA: California Department of Education.

California Department of Education. (2024c, August). *Organization*. California Department of Education website.

California Department of Education. (2024d, October). *State Board of Education responsibilities*. California Department of Education website.

California *Education Code*, Title 2, Elementary and Secondary Education, Division 4, Part 33, Chapter 5, sections 60602.5(1) and 60602.5(a)(4). (n.d.).

Every Student Succeeds Act of 2015, 20 U.S.C. 20 § 6301 (2015).

Smarter Balanced Assessment Consortium. (n.d.). *Smarter assessments*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2023). *Smarter Balanced scoring specifications for summative and interim assessments*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

## Overview of CAASPP Smarter Balanced Processes

This chapter provides an overview of the processes implemented by ETS during a typical, full testing cycle for the Smarter Balanced Summative Assessments, including test design, test administration, and scoring. The details on each step in the process will be presented in the subsequent chapters.

### Item Development

All items in the Smarter Balanced operational item bank for the first year of testing were developed and revised during the pilot and field test periods. Thereafter, Smarter Balanced items are developed dynamically. New items are developed and field-tested by being embedded in the operational assessments. Each year, some new items are added into the Smarter Balanced operational item banks, and some poorly performing items are removed from the item banks.

During item development, item and performance task (PT) specifications provide guidance on how to translate the Smarter Balanced content specifications into actual assessment items (Smarter Balanced, 2016, 2017, 2018b). Guidelines for bias and sensitivity, accessibility and accommodations, and style help item developers and reviewers ensure consistency and fairness across the item development process. These specifications and guidelines from Smarter Balanced were reviewed by member states, local educational agencies (LEAs), higher education professionals, and other educational partners (Smarter Balanced, 2016). For more information regarding the item response theory (IRT) methodology used by Smarter Balanced to form the basis for new item development, test equating, and computer adaptive testing, refer to chapter 8 of the *2018–19 Smarter Balanced Summative Technical Report* (Smarter Balanced, 2020)*.*

#### Item Format

The Smarter Balanced assessments include the following computer-based item formats:

* Selected response
* Constructed response
* Technology enhanced

Formats for these item types are described in more detail in subsection [*7.1.2 Scoring Item Responses*](#_Scoring_Item_Responses)*.*

#### Item Specifications

The item specifications describe the characteristics of the items that should be written to measure each content standard. Items of the same type should consistently measure the content standards in the same way. The *Smarter Balanced Item and Task Specifications*, hosted in the Smarter Content Explorer, were given to item developers to help ensure that the assessments are measuring the intended constructs without influence from extraneous factors. These documents contain item specification tables and provide item writers with definitions of the constructs that are intended to support the claims of measurement and clear direction regarding the types of evidence needed for students to demonstrate their knowledge and skills (Smarter Balanced, 2016, 2017, 2018b); note that because these specifications were reorganized following the initial development, their publication dates were updated.

### Test Assembly

#### Test Length

The California Assessment of Student Performance and Progress (CAASPP) computer-based summative assessments for English language arts/literacy (ELA) and mathematics are composed of two portions: the computer adaptive test (CAT) and the PT.

##### Operational Testing

The number of PT items that a student is administered depends on the particular PT a student is assigned randomly at the student level. Refer to table 4.B.1 in [appendix 4.B](#_Appendix_4.B:_Performance) for the number of items in each PT. Refer to table 4.C.1 through table 4.C.3 in [appendix 4.C](#_Appendix_4.C:_Item) for the distributions of the number of items presented to students in the total assessment, CAT, and PT components, respectively.

The number of CAT items encountered in an individual testing session may vary from student to student. The length of the CAT portion is determined by the termination rule of the CAT engine, which includes the following conditions:

1. Administer at least a specified minimum number of items in each reporting category and overall
2. Achieve a target level of precision on the overall test score
3. Achieve a target level of precision on all reporting categories

The termination rule of CAASPP assessments is discussed in more detail in the *Smarter Balanced Adaptive Item Selection Algorithm Design Report* (American Institutes for Research [AIR], 2014).

##### Field Testing

Field test PTs have been embedded into the Smarter Balanced operational assessments since the 2016–17 administration. Students who were assigned an embedded field test PT were not assigned an operational PT. Instead, they were assigned a CAT version with additional items for the purpose of reporting aggregated claim results. For ELA, these students received three additional items. For mathematics, these students received two additional items. Refer to *Appendix 2.A: Smarter Balanced Test Blueprints* in *Enhanced CAT Blueprints for Students Participating in the 2019–20 Smarter Balanced Embedded Field Test of PTs* for the number of CAT items with embedded field test PTs in the blueprints (Smarter Balanced, 2018a). (Note that in the 2023–24 administration, the adjusted, shortened-form blueprint for the field test was used, which is included in [appendix 4.A](#_Appendix_4.A:_Smarter).)

#### Test Blueprints

Blueprints represent a set of constraints and specifications to which each test form must conform. Each grade band—grades three through five, grades six through eight, and grade eleven—of the Smarter Balanced assessments includes a separate blueprint ([appendix 4.A](#_Appendix_4.A:_Smarter)) with criteria including, but not limited to,

* whether the assessment is adaptive or fixed form,
* termination conditions for the segment,
* content constraints such as minimum or maximum number of items administered in each content category, and
* nonnested content constraints such as priority weights for a group of items.

##### Operational Items

In November 2020, the California Department of Education (CDE) adopted the Smarter Balanced Summative Assessment adjusted, shortened-form blueprints for ELA and mathematics computer-based assessments. The adjusted blueprint reduces the number of items in the CAT. As of the 2022–23 test administration, the braille hybrid adaptive test used the adjusted blueprint as well. The paper–pencil test (PPT) used the Smarter Balanced Summative Assessment full-form blueprint.

Both the adjusted-form blueprint and the full-form blueprint are included in [appendix 4.A](#_Appendix_4.A:_Smarter). Table 4.C.4 in [appendix 4.C](#_Appendix_4.C:_Item) presents the percentages of students who met the operational blueprint overall and by claim.

##### Field Test Items

In the 2023–24 administration, the adjusted, shortened-form blueprint for the embedded field test PTs was used. The field test PTs do not contribute to score reporting. Instead, the additional operational CAT items as shown in the field test blueprints are counted into score reporting. The adjusted-form blueprint for the field test PTs is included in [appendix 4.A](#_Appendix_4.A:_Smarter). Table 4.C.5 in [appendix 4.C](#_Appendix_4.C:_Item) presents the percentages of students who met the embedded field test blueprint overall and by composite claim for mathematics.

#### Item Selection

In the CAT portion of each assessment, items are presented to a student according to the adaptive algorithm mapped onto the test blueprint (AIR, 2015). Use of the adaptive algorithm in 2015–16 testing and simulation studies in the following years is discussed in the report *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm* (AIR, 2015; Smarter Balanced, 2019).

For more information regarding test length, refer to [*2.2.1 Test Length*](#_Test_Length); the test blueprints (including adjusted, shortened-form blueprints for regular computer-based testing and braille hybrid and full blueprints for the PPTs) are provided in [appendix 4.A](#_Appendix_4.A:_Smarter). Refer also to the *ELA/Literacy Adjusted Form Summative Assessment Blueprint* and the *Mathematics Adjusted Form Summative Assessment Blueprint* on the Smarter Balanced website (Smarter Balanced, 2021a, 2021b).

### Test Administration

The Smarter Balanced Summative Assessments were administered using the secure browser and Test Delivery System (TDS), ensuring a secure, confidential, standardized, consistent, and appropriate administration for students. Additional information about the administration of the Smarter Balanced Summative Assessments can be found in [*Chapter 5: Test Administration*](#_Test_Administration).

Testing could occur in person and remotely. Students receiving in-person instruction were tested in person, at a school site. Remote administration, which is intended as an option for an LEA only when its students are receiving remote instruction, occurred when either the students, test administrator, or both were located at different physical locations. In remote testing, the test administrator monitors students’ progress throughout the assessment by using remote monitoring tools connected to the TDS.

#### Test Security and Confidentiality

All operational assessments within the CAASPP System are secure. For the Smarter Balanced Summative Assessment administration, every person having access to test materials maintained the security and confidentiality of the assessments. The ETS internal Code of Ethics requires that all test information, including tangible materials (such as test booklets, test items and test results), confidential files, processes, and activities were kept secure. To ensure security for all assessments that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI). A detailed description of the OTI and its mission is presented in subsection[*5.6.1 The ETS Office of Testing Integrity*](#_The_ETS_Office) in [*Chapter 5: Test Administration*](#_Test_Administration).

In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes involved in an assessment development and administration cycle. Those processes are listed next. The practices related to each of the following security processes are discussed in detail in section [*5.6 Test Security and Confidentiality*](#_Test_Security_and_3):

* Procedures to maintain standardization of test security
* Test security monitoring
* Security of electronic files using a firewall
* Transfer of scores via secure data exchange
* Data management in the secure database
* Statistical analysis on secure servers
* Student confidentiality
* Student test results

#### Procedures to Maintain Standardization

ETS takes all necessary measures to ensure the standardization of administration of the Smarter Balanced Summative Assessments.

The Smarter Balanced Summative Assessments are administered in conjunction with the other assessments that compose the CAASPP System. ETS employs processes to ensure the standardization of an administration cycle; these processes are discussed in more detail in section [*5.2 User Roles and Standardization*](#_User_Roles_and).

Staff at LEAs involved in the CAASPP administration include LEA CAASPP coordinators, site CAASPP coordinators, and test administrators. The responsibilities of each of the staff members are described in the *Test Operations Management System* *(TOMS) User Guide* CDE, 2024b).

Several series of instructions regarding the CAASPP administration are compiled in detailed manuals and provided to the LEA staff. Such documents include, but are not limited to, the following:

* ***CAASPP Online Test Administration Manual*—**This web-based manual provides test administration procedures and guidelines for LEA CAASPP coordinators and site CAASPP coordinators, as well as the script and *Directions for Administration* to be followed exactly by test administrators during a testing session (CDE, 2024a). (Refer to [*5.2.4.2 CAASPP Online Test Administration Manual*](#_CAASPP_Online_Test) in [chapter 5](#_Test_Administration) for more information.)
* ***TOMS User Guide*—**This web-based manual provides instructions for TOMS, allowing LEA staff, including LEA CAASPP coordinators and site CAASPP coordinators, to perform several tasks, including setting up test administrations, adding and managing users, assigning assessments, and configuring computer-based student test settings (CDE, 2024b). (Refer to [*5.2.4.3 Test Operations Management System User Guide*](#_Test_Operations_Management) in [chapter 5](#_Test_Administration) for more information.)

### Fairness and Accessibility

Several procedures are in place to ensure that the Smarter Balanced Summative Assessments are fair and accessible to all students. This section provides information on the available accessibility resources.

#### Overview

All eligible students enrolled in a California public school participate in the CAASPP System of assessments, including students with disabilities and English learner (EL) students. Additional resources are sometimes needed for these students. The CDE provides a full range of assessment resources for all students, including those who are EL students and students with disabilities.

#### Student Accessibility Resources

There are four different categories of student accessibility resources in the California assessment accessibility system, including universal tools, designated supports, accommodations, and unlisted resources that are permitted for use in CAASPP computer-based assessments. These are listed in the CDE California Assessment Accessibility Resources Matrix (Accessibility Matrix) (CDE, 2023).

**Universal tools** are available to all students. These resources may be turned on and off when embedded as part of the technology platform for the computer-based CAASPP on the basis of student preference and selection.

**Designated supports** are available to all students when determined as needed by an educator or team of educators, with parent/guardian and student input as appropriate, or when specified in the student’s individualized education program (IEP) or Section 504 plan.

**Accommodations** must be permitted on the CAASPP for all eligible students when specified in the student’s IEP or Section 504 plan.

**Unlisted resources** are non-embedded and made available if specified in the eligible student’s IEP or Section 504 plan and do not jeopardize test security, and only on approval by the CDE. An unlisted resource may change the construct being measured.

[Appendix 5.A](#_Appendix_5.A:_Accessibility_1) presents counts and percentages of students assigned designated supports, accommodations, and unlisted resources for the 2023–24 Smarter Balanced Summative Assessment administration. The tables in [appendix 5.A](#_Appendix_5.A:_Accessibility_1) were created using student demographic data in version 3 of the production data file (“P3”) updated on October 7, 2024.

The majority of students did not use any designated supports, accommodations, or unlisted resources.

### Scores

Individual student scores were reported for the 2023–24 Smarter Balanced Summative Assessment administration. Student performance on the reporting scale was designated into one of the four achievement levels described in subsection [*7.3.1.5 Achievement Levels*](#_Achievement_Levels). For information regarding score specifications and score reports, refer to [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting).

#### Score Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage aspects of CAASPP test administration such as test assignment and the assignment of test settings. TOMS also provides a secure means for LEA CAASPP coordinators to download Student Score Reports as PDF files.

Smarter Balanced Summative Assessment scores can also be viewed through the California Educator Reporting System (CERS), a secure website that provides authorized users with interactive and cumulative online reports for content area at the student, school, and LEA levels. CERS also provides individual score reports. Refer to subsection [*7.5.1 Online Reporting*](#_Online_Reporting_3) for details about TOMS and CERS and subsection [*7.5.3 Types of Score Reports*](#_Types_of_Score) for the content of each type of score report.

#### Aggregation Procedures

To provide meaningful results to interested educators, Smarter Balanced Summative Assessment scores for a given grade-level assessment were aggregated at the school, LEA or direct funded charter school, county, and state levels. State-level results are available on the Test Results for California’s Assessments website. The aggregate scores were presented for all students or selected demographic student groups.

Aggregate scores were generated by combining student scores at the state, LEA or direct funded charter school, or school level; combining student scores for all students; or by combining student scores for students who represent selected demographic student groups.

The aggregation procedures used to present Smarter Balanced Summative Assessment results are described in section [*7.4 Overview of Score Aggregation Procedures*](#_Overview_of_Score). Aggregate results by demographic variables are presented in [appendix 7.D](#_Appendix_7.D:_Demographic). In table 7.D.1 through table 7.D.6, students are reported by demographic groups, including gender, ethnicity, English language fluency, disability status, and economic status, as well as crosstab analysis for ethnicity and economic status. The tables show the numbers of students with valid scores in each group, scale score means and standard deviations, and the percentage of students in each achievement level. To protect student privacy, statistics are presented in the tables as “N/A” when the number of students in the sample is 10 or fewer. Definitions for the demographic student groups included in these tables are provided in table 7.29.

### Calibration and Scaling

IRT methods are ideally suited to the assessments and measurement goals of Smarter Balanced in both establishing a common scale and ongoing maintenance of the program. The purpose of calibration, equating, and scaling using IRT methods is to place item difficulty and student ability estimates at all grade levels in each content area onto a common theta scale. As a result, scores on different versions of the same assessment are statistically adjusted to compensate for any differences in difficulty between the test versions.

The Common Core State Standards were developed with the intent of supporting inferences concerning a student’s change in achievement (that is, progress) as demonstrated by performance on the corresponding assessments. *Vertical scaling* is an approach that places test scores across grade levels onto a common scale. A vertical scale is a single scale for scores on assessments at different grade levels of the same content area. Reporting scores on a vertical scale allows student progress to be tracked for a particular content area across grade levels; it is expected that students’ proficiency increases across different levels of the assessment. An advantage of vertical scaling is that progress expectations concerning the establishment of achievement levels across grade levels can be inspected and ordered by standard setting panelists.

All items used on the Smarter Balanced Online Summative Assessments were calibrated within grade levels and vertically scaled during the 2013–14 Smarter Balanced field test phase (Smarter Balanced, 2016). These activities supported the creation of scale scores.

The basic steps in the process of scaling the scores in each content area—ELA or mathematics—are as follows:

1. Calibrate the items at each grade level
2. Transform the ability scales at the different grade levels onto a common ability scale
3. Transform the common ability scale onto the reported score scale by applying a single linear transformation for all grade levels

The reported test scores for the 2023–24 administration of the Smarter Balanced assessments were expressed on the baseline scale. The baseline scale was defined following the 2013–14 Smarter Balanced field test administration. Items developed in later years were linked to the baseline scale after being field-tested.

#### Calibration

Unidimensional IRT models were used for calibration. Based on the psychometric research conducted during the pilot and field test phases by the Smarter Balanced Assessment Consortium, the two-parameter logistic (2PL) model (Birnbaum,1968) and the generalized partial credit model (GPCM) (Muraki, 1992) were chosen for calibration. Refer to [equation 7.1](#EQ7_1) in subsection [*7.3.1.1 Theta Scores*](#_Theta_Scores) for the 2PL model and GPCM formulas.

Item parameter calibration software, model-to-data fit, and evaluation of vertical scale anchor items are described in more detail in chapter 6 of the *2013–2014 Smarter Balanced Technical Report* (Smarter Balanced, 2016). The summary statistics describing the distribution of item difficulty and discrimination parameter estimates at each grade level for the 2023–24 administration item pool are available in [appendix 8.D](#_Appendix_8.D:_Item).

#### Horizontal Scaling

Item parameters for the Smarter Balanced assessments were linked during the Smarter Balanced field test administration by concurrently calibrating items within each grade level for each content area. The calibration approach relied on a hybrid of the “common items” approach and the “randomly equivalent groups” linking approach. The common items approach requires that items and tasks partially overlap and be administered to different student samples. For the randomly equivalent groups approach, the assessment items presented to different student samples are considered as comparably “on scale” by virtue of the random equivalence of the groups. The horizontal linking design incorporated both types of approaches and was accomplished by assembling test versions with partially overlapping test content and randomly assigning the test versions to students.

#### Vertical Scaling

After the grade-specific horizontal scaling was conducted for a content area, a separate, cross-grade, vertical scaling was conducted by the Smarter Balanced Assessment Consortium using common items (vertical linking items). To implement the vertical scaling, representative sets of off-grade items were administered to some students in the next lower adjacent grade—for example, a set of grade five items was administered to some students in grade four.

Vertical linking item sets were intended to sample the construct that included both the CAT and PT components and associated item types as well as claims that conformed to the test blueprint. Linking items from the lower grade were administered to the upper-adjacent-grade–level students. Content experts designated a target grade for each item and a minimum and maximum grade designation. A set of PTs was given on-grade; the same set was administered off-grade for vertical linking.

The vertical scaling was undertaken separately for ELA and for mathematics, using grade six as the base grade. Grade seven was linked to grade six, and then grade eight was linked to grade seven, and so forth, until grade eleven was placed onto the vertical scale. Likewise, grade five was linked to grade six, grade four was linked to grade five, and so forth, until grade three was placed onto the vertical scale (refer to figure 2.1).

Grade 3

Grade 4

Grade 5

Grade 6

Grade 7

Grade 8

Grade 11

Figure 2.1 Vertical scaling

Once the Smarter Balanced horizontal and vertical scales were established, the remaining items (that is, the entire calibration item pool, including the noncommon items) were linked onto this final scale in each grade level and content area.

#### Vertical Scale Evaluation

The results of vertical scaling were evaluated using a number of methods. Refer to the section *Vertical Scale Evaluation* in *Chapter 9 Field Test Design, Sampling, and* *Administration* in the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016). This source includes the following information:

* Correlation of difficulties of common items across grade levels
* Changes in test difficulty across grade levels
* Comparison of mean scale scores across grade levels
* Comparison of scale scores associated with achievement levels across grade levels
* Comparison of overlap and separation of scale score distributions across grade levels
* Comparison of variability in scale scores within and across grade levels

### References

American Institutes for Research. (2014). *Smarter Balanced adaptive item-selection algorithm design report*. Washington, DC: Jon Cohen and Larry Albright.

American Institutes for Research. (2015). *Smarter Balanced summative assessments testing procedures for adaptive item selection algorithm, 2014–2015 test administrations, English language arts/literacy (ELA), grades three–eight and grade eleven, and mathematics, grades three–eight and grade eleven*. Washington, DC: American Institutes for Research.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, PA: Addison-Wesley.

California Department of Education. (2023). *California assessment accessibility resources matrix*. California Department of Education website.

California Department of Education. (2024a). *CAASPP online test administration manual*. Sacramento, CA: California Department of Education.

California Department of Education. (2024b). *Test Operations Management System user guide*. Sacramento, CA: California Department of Education.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–76.

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–2014 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017). *ELA computer adaptive test (CAT) and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018a). *Enhanced CAT blueprints for students participating in the 2017-18 Smarter Balanced embedded field test of performance tasks*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT and PT item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: 2017–18 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: 2018–19 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2021a). *ELA/Literacy adjusted form summative assessment blueprint*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2021b). *Mathematics adjusted form summative assessment blueprint*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

## Item Development

This chapter discusses the procedures followed during the development of Smarter Balanced items to help ensure valid interpretation of test scores.

### Background

The Smarter Balanced Assessment Consortium, in coordination with its member states, developed innovative item types and authored items based on the Common Core State Standards. The Consortium used an iterative process involving higher education and kindergarten through grade twelve educators who were trained in item development, as well as state partners, professional item writers, and assessment vendors at various stages in the item development process.

### Additional Information

More information regarding the item development process (including the qualifications of those involved), item development specifications, and content alignment studies undertaken by Smarter Balanced to produce item types and items for the assessment can be found in chapter 3 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016).

### Reference

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–2014 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

## Test Assembly

The Smarter Balanced Summative Assessments were administered operationally as part of the California Assessment of Student Performance and Progress (CAASPP) for the first time during the 2014–15 school year. The summative assessments each consist of two parts: a computer adaptive test (CAT) and performance tasks (PTs).

The Smarter Balanced Summative Assessments are constructed to measure students’ performance relative to Common Core State Standards (CCSS). The assessments also are constructed to produce scores that meet professional standards for reliability and validity of test score interpretation. The content standards and desired psychometric attributes are used as the basis for assembling the test forms. This chapter discusses the content and psychometric criteria that guided the construction of the Smarter Balanced Summative Assessments.

### Smarter Balanced Adaptive Item Selection Algorithm

This section describes the algorithm and the design for implementation of adaptive item selection of CAT items by the Smarter Balanced Test Delivery System (TDS). The implementation builds extensively on the algorithm implemented in the Cambium Assessment, Inc. (CAI) TDS. It should be noted that a student is assigned PTs randomly at the student level, regardless of performance on the CAT portion of a content-area assessment.

The general item selection approach is that the next item to be administered to a specific student is chosen on the basis of a function of three variables. The first variable is an index of the importance of the item for meeting the content requirements of the assessment. The other two variables are values of the item response theory (IRT) item information functions in the region of the student’s current ability estimate. One of these information functions is for the student’s total score; the other is for the student’s claim score (refer to table 7.22 and table 7.23 for the list and descriptions of English language arts/literacy [ELA] and mathematics claims, respectively).

More information about how each of these three measures is defined can be found in the *Smarter Balanced Adaptive Item-Selection Algorithm Design Report* (American Institutes for Research [AIR], 2015).

Values for these three measures are calculated to guide and support item selection. A value is computed for whether the item or group of items will be selected on the basis of how well that item matches the target content, contributes to overall score information, and contributes to claim score information. *Refer to the* [*Alternative Text for Equation 4.1*](#_Alternative_Text_for) *for a description of equation 4.1.*

Equation 4.1; a link to the long description for this equation is found in the preceding paragraph. (4.1)

This objective function is used to measure an item’s contribution to each of these objectives. A higher value for “Content Match” means that an item is more important for meeting the content requirements. A higher value for “Overall Information” means that an item contributes more information to the estimation of the student’s current overall ability. A higher value for “Claim Information” means that an item contributes more information for estimating the student’s current claim ability. Weights of these objectives can be adjusted to achieve the desired balance and optimize performance for a given item pool. This algorithm enables users to maximize information subject to the constraint that the blueprint is almost always met, with minimal exceptions.

#### Content Match

Each item or item group is characterized by its contribution to meeting the blueprint, given the items that have already been administered at any point. The contribution is based on the presence or absence of features specified in the blueprint.

The Smarter Balanced summative test blueprints describe the content of the ELA and mathematics summative assessments for all grade levels tested and the means by which that content is assessed. The summative computer-based test blueprints reflect the depth and breadth of the performance expectations of the CCSS.

The test blueprints have information about the number of items and depth of knowledge for items associated with each claim and assessment target. Each assessment is described by a blueprint for both the overall assessment and each claim within the assessment.

Each blueprint has features referred to as *constraints*. Constraints are features such as the minimum and maximum number of items required in a specific content area. For example, a constraint might require a minimum of four and a maximum of six algebra items. The value of content match

* is highest for items with content that has not met its minimum constraint,
* decreases for items representing content for which the minimum number of items has been reached but the maximum has not, and
* becomes negative for items representing content that has met the maximum.

Refer to the blueprints for the Smarter Balanced ELA and mathematics assessments provided in [appendix 4.A](#_Appendix_4.A:_Smarter) for additional details.

#### Information

Every item has an overall information value within the CAT algorithm and an information value for each claim. Details on how information is calculated are provided in [equation 7.8](#EQ7_8) through [equation 7.12](#EQ7_12) in [*7.3.4 Theta Score Standard Errors*](#_Theta_Score_Standard).

Items with higher discrimination parameters offer more information and therefore are generally given preference in item selection. Because these highly discriminating items offer more information than items that do not discriminate as well, they are more likely to be selected by an item selection algorithm. Given the choice between two items that differ only on their discrimination, the higher-discriminating item will always result in a better subsequent estimation of the student’s proficiency compared to the lesser-discriminating item. Since these items are more likely to be selected, they are also more likely to be overexposed; therefore, measures must be taken to control for this potential overexposure. Because this is a test security risk, the item selection algorithm includes additional rules to control the exposure of the items that provide the highest measurement information (AIR, 2014).

### Simulation Study

Before the operational 2023–24 testing window opened for the CAT, simulation studies were conducted to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm. The purposes of the CAT simulations were to configure the adaptive algorithm, thus optimizing item selection to meet the adjusted test blueprint specifications; while also targeting test information to student ability and controlling item exposure rates. The simulation tool allowed manipulation of key blueprint and configuration settings to match the assessment blueprint and minimize measurement error.

In this simulation study, the adaptive assessments were administered in one segment (section) in ELA for grades three through eight, grade ten[[5]](#footnote-6), and grade eleven as well as in mathematics for grades three through five. It was administered in two segments in mathematics grades six through eight and grade eleven, including calculator and no-calculator segments. Each segment was simulated separately.

The study results (CAI, 2024) examine the robustness of the item-selection algorithm of the Smarter Balanced CAT administrations in ELA and mathematics for grades three through eight and grade eleven. The information provided by the simulations includes

* the process of the item-selection algorithm;
* the percentage of assessment content aligned with the test blueprints for all constraints (blueprint match rates);
* the number of targets covered in the simulated forms;
* accuracy of ability estimates indicated by bias and precision of ability estimates indicated by standard error and confidence intervals;
* item exposure rates; and
* selection of off-grade items and corresponding psychometric properties.

Results from the simulation study showed that the Smarter Balanced CAT delivery system administers assessments with items representing the breadth and depth of the assessment blueprints. The blueprint match results demonstrated that all test forms adhered to the exact assessment blueprint requirements and, therefore, provided evidence of content comparability. While each form is unique with respect to the items contained, all forms aligned with the curricular expectations outlined in the assessment blueprints.

The summary statistics of the estimated abilities show that, for all students in all grade levels, the item-selection algorithm selects optimized items on the basis of each student’s ability—to the extent that the item pool allowed—in meeting the blueprint constraints. These results demonstrated that the student ability estimates generated on the basis of the selected items were optimal, an indication that the algorithm is working as expected, given the item pool. The final score on the CAT component for each student recovers the true score across ability ranges except in the extremely low or high ability ranges.

Overall, the simulation results on the item-selection algorithm support the following interpretations:

* Scores are comparable with respect to the targeted content.
* Scores are measured with good precision across the range of proficiency, given the distribution of the items’ contents and level of difficulty in the pool. However, larger biases were observed at the lower and upper ends of the ability distribution.
* Item exposure is minimized on the basis of the configuration parameters chosen, with most items appearing in 10 percent or fewer of student test administrations. Off-grade items are administered according to the criteria.

### Test Blueprints

As stated in subsection [*2.2.2 Test Blueprints*](#_Test_Blueprints), the adjusted, shortened-form blueprints for ELA and mathematics computer-based assessments and the field test PTs were used for the Smarter Balanced computer-based assessments. The adjusted blueprint reduced the number of items in the CAT.

The adjusted, shortened-form blueprints were used for braille hybrid adaptive tests (HATs) as well. The full-form blueprints were used for the paper–pencil tests. Both the adjusted-form blueprint and the full-form blueprint are included in [appendix 4.A](#_Appendix_4.A:_Smarter).

### Test Sections

The test delivery sections correspond to the CATs and PT portions of the assessments.

The distributions of the number of items presented to students for the total assessment, as well as the CAT and the PT components, are presented in table 4.C.1 through table 4.C.3 in [appendix 4.C](#_Appendix_4.C:_Item). Table 4.C.4 presents the percentage of students administered items who meet the criteria specified in the operational blueprints, students who do not meet the criteria, and students who exceed the criteria by claim. Results show that more than 99 percent of the individual assessments delivered to students met the requirements of the blueprints for overall assessment and claims across grade levels and content areas. Table 4.C.5 presents the same information for the embedded field test blueprint. Results are similar to the operational blueprints; more than 99 percent of the individual assessments delivered to students met the requirements of the blueprints for the overall assessment and claims across grade levels and content areas.

Table 4.B.1 in [appendix 4.B](#_Appendix_4.B:_Performance) lists the number of ELA PTs given to students and the number of items in each PT by genre. [Appendix 4.C](#_Appendix_4.C:_Item) presents item distributions, including a summary of items presented for the total assessment—in the CAT portion and in the PT portion—as well as the percentage of students meeting the Smarter Balanced adjusted, shortened-form blueprint.

#### Computer Adaptive Testing Administration

CAT assessments are assembled and delivered dynamically to obtain a unique assessment for each student from a defined item pool so that each student is given a unique, content-conforming test form. CAT items are delivered on the basis of the student’s performance on the previous items; students typically are presented with many different items, and items presented to any two students may appear in different locations within the assessment.

Item statistics based on IRT are used to determine the administration and adaptation of test items based on student responses and ability; this information is incorporated into the delivery algorithm. The item selection algorithm is described in more detail in [*4.1 Smarter Balanced Adaptive Item Selection Algorithm*](#_Smarter_Balanced_Adaptive), along with item exposure rates.

Item exposure control (for example, Sympson & Hetter, 1985) can be used to ensure that uniform rates of item administration are achieved because it is not desirable to have some items presented to many students while other items are presented to relatively few students.

#### Performance Task Administration

For a given PT, students are presented with the same items in the same order of presentation and associated test length (refer to table 4.B.1 in [appendix 4.B](#_Appendix_4.B:_Performance) for the number of items in each operational PT). PTs are assigned randomly at the student level.

Smarter Balanced Assessment Consortium item and task specifications assume computer-based delivery of the items and tasks. Most tasks are long enough to warrant several administration sessions. Such sessions could be same-day, back-to-back sessions with short breaks between sessions.

Estimated testing times for completing PTs and administration time are provided in the *CAASPP Online Test Administration Manual* (California Department of Education [CDE], 2024).

Student directions for all tasks begin with an overview of the entire task that briefly describes the necessary steps. The overview gives students advanced knowledge of the scorable products or performances to be created (Khattri, Reeve, & Kane, 1998). Allowable teacher–student interactions for a task are standardized (that is, carefully scripted or described in task directions for purposes of comparability, fairness, and security). Teachers are directed not to assist students in the production of their scorable products or presentations.

### Special Version Forms

#### Braille Form

There are three braille forms available for the Smarter Balanced Summative Assessments. One is the emergency paper form, which can be ordered in braille; the second is the computer-based general form in braille, where the CAT and PTs can be taken in braille; and the third is the braille HAT. The HAT is for mathematics only and is a mix of CAT and fixed forms with an accompanying tactile graphics package. The braille HAT allows local educational agencies or schools to order a set of preprinted embossed graphics for a student’s mathematics assessment in cases where the student would not otherwise have access to an embossing printer that works with the Smarter Balanced mathematics assessment items.

Smarter Balanced braille provides students with visual impairments the opportunity to demonstrate what they know and can do, equitably. Offering braille as a built-in accommodation is a complex process and starts at the creation of a new item for the assessment. Multiple reviewers perform a review of all the items created each year to ensure that the items will not disadvantage any particular group of students. Careful thought is put into the development of items; any images used are checked to ensure that the items can be translated into braille. From there, the items are sent to certified braille transcribers who create multiple forms of braille for each item.

The digital braille files stay with the items and are delivered directly to the student via the computer on testing day. The student accesses the braille directly using a voice output program (screen reader) with a refreshable braille display that delivers braille to the student’s fingertips. If there are images within the items, the student can send those braille files directly to a braille embosser—a printer for braille—and have a paper copy of the brailled item available. Smarter Balanced also offers paper braille versions of the summative assessment.

#### Paper–Pencil Form

The 2023–24 CAASPP Smarter Balanced Summative Assessment administration included a paper-based form for students whose individualized education program or Section 504 plan specified testing on a paper–pencil form or when a school did not have the necessary computer network infrastructure to administer the computer-based assessments. ETS also provided the braille, large-print, and standard-print versions to schools experiencing unexpected, temporary technology issues beyond the schools’ control, on the basis of CDE approval of the request. The Smarter Balanced Assessment Consortium selects a paper form each year among developed paper forms. The paper form used in 2023–24 was Form 6.

### References

American Institutes for Research. (2014). *Smarter Balanced adaptive item selection algorithm design report*. Washington, DC: American Institutes for Research.

American Institutes for Research. (2015). *Smarter Balanced Summative Assessments testing procedures for adaptive item-selection algorithm*. Washington, DC: American Institutes for Research.

California Department of Education. (2024). *CAASPP online test administration manual*. Sacramento, CA: California Department of Education.

Cambium Assessment, Inc. (2024). *Smarter Balanced Summative Assessments simulation results for computer-adaptive test component with adjusted blueprints*. [Unpublished manuscript]. Washington, DC: Cambium Assessment, Inc.

Khattri, N., Reeve, A., & Kane, M. (1998). *Principles and practices of performance assessment*. Mahwah, NJ: Routledge.

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings from the 27th Annual Meeting of the Military Testing Association* (pp. 973–77). San Diego, CA: Navy Personnel Research and Development Center.

### Accessibility Information

#### Alternative Text for Equation 4.1

Item Selection equals w sub 1 multiplied by Content Match plus w sub 2 multiplied by Overall Information plus w sub 3 multiplied by Claim Information. *(Return to* [*equation 4.1*](#EQ4_1)*.)*

### Appendix 4.A: Smarter Balanced Test Blueprints

*This content is located in a separate file.*

### Appendix 4.B: Performance Task Test Length

*This content is located in a separate file.*

### Appendix 4.C: Item Distribution

*This content is located in a separate file.*

## Test Administration

This chapter details the processes involved in the administration of the 2023–24 Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics. It also describes the procedures followed by ETS to maintain test security throughout the test administration process.

### Overview

The Smarter Balanced Summative Assessments were administered to students in grades three through eight and grade eleven in 2023–24 in conjunction with the other assessments that compose the California Assessment of Student Performance and Progress (CAASPP) System.

In accordance with the procedures for the computer-based CAASPP, local educational agencies (LEAs) identified test administrators and entered the test administrators as users into the Test Operations Management System (TOMS). ETS provided LEA staff with the appropriate training materials, such as test administration manuals, videos, and webinars, to ensure that the LEA staff and test administrators understood how to administer the computer-based Smarter Balanced Summative Assessment content-area assessments.

The testing window for the 2023–24 administration of the Smarter Balanced Summative Assessments was January 9, 2024, through June 28, 2024. Specific test administration schedules within that window were determined locally pursuant to *California Code of Regulations*,Title 5 (5*CCR*), Section 855(a).

#### In-Person and Remote Testing

Table 5.1 shows the number of students who took the Smarter Balanced assessments for ELA and mathematics in each grade level, broken down by test administration type (in person, remote, or hybrid). The large majority of test takers at every grade level tested in person.

Table 5.1 Number of Students by Assessment Location

|  |  |  |  |
| --- | --- | --- | --- |
| **Content Area and Grade Level** | **In-Person Assessment** | **Remote Assessment** | **Hybrid Assessment** |
| ELA 3 | 394,559 | 8,800 | 216 |
| ELA 4 | 404,367 | 8,928 | 184 |
| ELA 5 | 409,305 | 8,915 | 193 |
| ELA 6 | 410,495 | 9,625 | 269 |
| ELA 7 | 413,362 | 10,663 | 287 |
| ELA 8 | 413,328 | 10,897 | 470 |
| ELA 11 | 423,203 | 11,671 | 415 |
| Mathematics 3 | 399,564 | 8,834 | 171 |
| Mathematics 4 | 409,036 | 8,910 | 195 |
| Mathematics 5 | 413,435 | 8,899 | 178 |
| Mathematics 6 | 413,925 | 9,695 | 152 |
| Mathematics 7 | 415,986 | 10,759 | 166 |
| Mathematics 8 | 415,555 | 10,998 | 298 |
| Mathematics 11 | 422,687 | 11,724 | 292 |

### User Roles and Standardization

The test administration procedures were designed so that the assessments are administered in a standardized manner. ETS took all necessary measures to ensure the standardization of test administration, as described in this section.

#### Local Educational Agency CAASPP Coordinator

An LEA CAASPP coordinator was designated by the district superintendent or charter school administrator at the beginning of the 2023–24 school year. LEAs include public school districts, California State Board of Education–authorized charter schools, county office of education programs, and direct funded charter schools.

LEA CAASPP coordinators were responsible for ensuring the proper and consistent administration of the CAASPP. In addition to the responsibilities set forth in 5*CCR* Section 857, their responsibilities included

* adding site CAASPP coordinators and test administrators into TOMS;
* training site CAASPP coordinators and test administrators regarding the state requirements and CAASPP administration as well as security policies and procedures;
* providing checklists for site CAASPP coordinators and test administrators to review in preparation for administering the summative assessments;
* overseeing test administration activities;
* reporting test security incidents (including testing irregularities) to the California Department of Education (CDE) using the online Security and Test Administration Incident Reporting System (STAIRS)/Appeals process; and
* requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeals process).

#### Site CAASPP Coordinator

A site CAASPP coordinator is trained by the LEA CAASPP coordinator for each test site (5*CCR* Section 857[f]). A site CAASPP coordinator must be an employee of the LEA and must sign a security agreement (5 *CCR* Section 859[a]).

A test site coordinator was responsible for identifying test administrators and ensuring that they have signed CAASPP *Test Security Affidavits* (5 *CCR* Section 859[d]). A site CAASPP coordinator’s duties may have included

* adding test administrators into TOMS;
* entering test settings for students;
* creating testing schedules and procedures for a school consistent with state and LEA policies;
* working with technology staff to ensure secure browsers are installed and any technical issues are resolved;
* monitoring testing progress during the testing window and ensuring students take the Smarter Balanced Summative Assessments, as appropriate;
* coordinating and verifying the correction of student data errors in the California Longitudinal Pupil Achievement Data System;
* ensuring a student’s test session is rescheduled, if necessary;
* addressing testing problems;
* reporting test security incidents (including testing irregularities) to the CDE using the online STAIRS/Appeals process;
* overseeing administration activities at a school site; and
* requesting an Appeal (if indicated by TOMS prompts while reporting an incident using the STAIRS/Appeals process).

#### Test Administrators

Test administrators were identified by site CAASPP coordinators as individuals who would administer the Smarter Balanced Summative Assessments.

A test administrator must have signed a security affidavit (5 *CCR* Section 859[d]).

A test administrator’s duties may have included

* ensuring the physical conditions of the testing room meet the criteria for a secure test environment;
* administering the CAASPP, including the Smarter Balanced Summative Assessments;
* reporting all test security incidents to the site CAASPP coordinator and LEA CAASPP coordinator in a manner consistent with state and LEA policies;
* viewing student information prior to testing to ensure that the correct student receives the proper assessment with appropriate resources and reporting potential data errors to site CAASPP coordinators and LEA CAASPP coordinators;
* monitoring student progress throughout the test session using the Test Administrator Interface; and
* fully complying with all directions provided in the *Directions for Administration* (*DFAs*) for the Smarter Balanced Summative Assessments (CDE, 2024d).

#### Instructions for Test Administration

##### *Directions for Administration*

The *DFAs* of the Smarter Balanced Summative Assessmentsused by test administrators to administer the Smarter Balanced assessments to students are included in the *CAASPP Online Test Administration Manual* (CDE, 2024d). Test administrators must follow all directions and guidelines and read, word-for-word, the instructions to students in the “SAY” boxes to ensure standardization of test administration. Additionally, the *CAASPP Online Test Administration Manual* provided information to test administrators regarding the systems involved in testing, including sections describing the Test Delivery System (TDS), so they could become familiar with the testing application used by their students (CDE, 2024d).

##### *CAASPP Online Test Administration Manual*

The *CAASPP Online Test Administration Manual* (CDE, 2024d) contained information and instructions on overall procedures and guidelines for all LEA and test site staff involved in the administration of computer-based assessments. Sections included the following topics:

* Test administration resources
* Test security
* Administration preparation and planning
* General test administration
* Test administration directions and scripts for test administrators
* Overview of the student testing application
* Instructions for steps to take before, during, and after testing

Appendices included definitions of common terms and descriptions of different aspects of the assessment and systems associated with the assessment.

##### *Test Operations Management System User Guide*

TOMS is a web-based application that allows LEA CAASPP coordinators to set up test administrations, add and manage users, submit computer-based student test settings, and order paper–pencil tests (PPTs).

TOMS modules described in the *TOMS User Guide* included the following (CDE, 2024e):

* **Test Administration Setup—**This module allowed LEAs to determine and calculate dates for the LEA 2023–24 administration of the CAASPP, including the Smarter Balanced Summative Assessments.
* **Adding and Managing Users—**This module allowed LEA CAASPP coordinators to add site CAASPP coordinators and test administrators to TOMS so that the designated user could administer, monitor, and manage the CAASPP computer-based assessments. The manual contained descriptions of the roles and responsibilities of those involved with CAASPP testing.
* **Reports—**This module allowed LEA CAASPP coordinators and site CAASPP coordinators access to the various reports in TOMS.
* **STAIRS/Appeals—**This module allowed LEA CAASPP coordinators and site CAASPP coordinators access to create new STAIRS cases or search for STAIRS/Appeals cases.
* **Student Profile—**This module allowed LEA CAASPP coordinators, site CAASPP coordinators, and test administrators and test examiners to view and manage student’s test assignments and test settings.

##### Other System Manuals

Other manuals were created to assist LEA CAASPP coordinators and others with the technological components of the CAASPP System and are listed next.

* ***CAASPP and English Language Proficiency Assessments for California (ELPAC) Technical Specifications and Configuration Guide for Online Testing*—**This manual provided information, tools, and recommended configuration details to help technology staff prepare computers and install the secure browser to be used for the computer-based CAASPP (CDE, 2024c).
* ***CAASPP and ELPAC Security Incidents and Appeals Procedure Guide*—**This manual provided information on how to report a testing incident and submit an Appeal to reset, reopen, invalidate, or restore individual computer-based student assessments (CDE, 2024b).
* ***CAASPP and ELPAC Accessibility Guide*—**This manual provided descriptions of the accessibility features for computer-based assessments as well as information about supported hardware and software requirements for administering assessments to students using accessibility resources, including those with a braille accommodation using Job Access With Speech® (software) or a braille embosser (hardware) (CDE, 2024a).

### Local Educational Agency Training

Each year, ETS, in collaboration with the CDE and its Assessment Validity and Outreach contractor, the Sacramento County Office of Education (SCOE), establishes and implements a comprehensive training plan for LEA assessment staff and educators on all aspects of the assessment program. The ETS and SCOE annual training plans specify the audience, topics, frequency, and mode (synchronous or asynchronous) of the training, including such elements as format, participants, and organization.

ETS and SCOE make every effort to make the information available in a variety of ways that allowed educators flexibility in accessing training at a time that best fits in with their day-to-day activities. This includes offering training events on multiple days and times, livestreaming events, recording and archiving training, and converting training to self-paced modules that can be taken any time, at the learner’s convenience.

All training opportunities are posted in one centralized location on the CAASPP & ELPAC Website. LEA staff can register for training opportunities in one place, on the Upcoming and On-Demand Trainings web page. Archived training is also made available on this web page, making it easier for educators to find a training they missed and providing easy access to recorded trainings. Participants can also register to receive a copy of the training materials without registering to attend the live training. Training materials are developed in such a way that educators can consume the information independently by accessing the training materials.

#### Synchronous and Asynchronous Training

All synchronous training for 2023–24 was offered on Zoom, recorded, and made available for on-demand viewing. Zoom provides an opportunity for educators to ask questions and get answers in real time. Coffee Sessions were live streamed on YouTube.

ETS and SCOE use various strategies to increase engagement during synchronous trainings. Live polls are presented to get real-time feedback about attendees’ knowledge of a particular topic, allowing presenters to tailor presentations to the audience’s level of understanding. The chat functionality is enabled to give participants an opportunity to interact with each other or to provide open-ended feedback, or it is disabled to minimize distraction and drive attendees’ focus to the information being presented. Breakout groups are used in smaller group trainings, as appropriate. Breaks and processing time are incorporated into presentations to give attendees opportunities to attend to other responsibilities that could otherwise distract from the training.

Working closely with the CDE, ETS and SCOE provide informal support to educators by offering monthly Coffee Sessions. Coffee Sessions include CDE and ETS staff who can answer questions about all aspects of testing. ETS also offers several Office Hours for coordinators where support staff are generally available from 9 a.m. to 3 p.m., allowing coordinators to join as needed and get customized support. SCOE continues to offer Assessment and Accountability Information Meetings intended to provide LEA coordinators with regular updates about California’s assessment and accountability systems. All trainings and meetings are recorded and archived for on-demand viewing on the Upcoming and On-Demand Trainings web page on the CAASPP & ELPAC Website.

#### Videos and Guides

ETS produced videos on various aspects of administering the CAASPP, including how to perform functions within TOMS, such as setting up a test administration window, adding users, assigning assessments to students, and uploading test settings. SCOE produced the accompanying quick reference guides, providing multiple avenues of support for educators administering the assessments.

In addition to the standard administration videos, ETS produced additional videos to support administration. Some videos were geared toward parents/guardians to help them understand the Student Score Reports (SSRs). Other videos were intended to help coordinators or other users complete a process, such as how to administer an assessment from start to finish, administering a practice or training test, starting and stopping a test session, how to monitor student completion, and how to complete second scoring that is required for some of the assessments. This list is a sampling of the available videos intended to capture the major areas of support for various interest holders. The comprehensive suite of training videos can be found on the CAASPP & ELPAC Website.

#### Training for Proper Identification and Assignment of Designated Supports and Accommodations

ETS produced and updated additional short demonstration videos to maintain the library of videos for every embedded accessibility resource. The videos demonstrate how to use the resource for educators, students, and parents/guardians and were available in both English and Spanish on the Accessibility Resource Demonstration Videos web page on the CAASPP & ELPAC Website. Demonstration videos for the most frequently used non-embedded accessibility resources are also available. These videos were linked within the Individual Student Assessment Accessibility Profile (ISAAP) Tool, increasing access to the demonstration videos. Educators using the ISAAP Tool to determine the student’s needs could view the corresponding demonstration video without having to navigate away from the tool.

A video on how to use the ISAAP Tool was also available to support educators in the process of creating an individual student profile and matching accessibility resources to student needs to ensure a fair and valid testing experience for all students.

For the 2023–24 CAASPP administration, ETS produced a two-part asynchronous training module. Module A, Matching Accessibility Resources to Students’ Needs, focused on providing educators with an understanding of the importance of accessibility resources, the categories of accessibility resources, and the process for matching students with appropriate accessibility resources for daily instruction and on assessments. Module B, Using Accessibility Resources in Daily Instruction, focused on the importance of removing barriers to student learning and using accessibility resources in daily instruction. Educators could complete the training independently or had the option to attend one of two live sessions held by ETS to extend and deepen the learning experience.

At the California Assessment Conference, SCOE offered four sessions on accessibility:

1. “Building Bridges to Accessibility: Creating a Systematic Approach for Successful Implementation of Accessibility Resources” focused on approaches to implementing accessibility resources in instruction and assessment.
2. “From Awareness to Action: Implementing Accessibility Resources in the Classroom” aimed to bridge the gap between awareness and action, sharing knowledge and strategies for creating inclusive learning environments; its focus was supporting students with disabilities and English learner (EL) students.
3. “Advancing Equity Through Assessment Accessibility” explored using Universal Design for Learning principles to meet student needs.
4. The “Accessibility Tools for Matching Resources to students” focused on the resources available to students participating in CAASPP and provided information on matching student needs and using those resources.

#### Feedback for Continuous Improvement Survey

The CAASPP and ELPAC programs solicit feedback annually from various interest holder groups, including LEA CAASPP coordinators, LEA ELPAC coordinators, site CAASPP coordinators, site ELPAC coordinators, test administrators, test examiners, and users with the IA Administrator Only (used to administer only interim assessments) role through the CAASPP and ELPAC Feedback for Continuous Improvement Survey. Feedback was collected via a post-test survey sent to more than 293,000 California educators and through focus groups. Educators provided valuable feedback for potential improvements to the future administration of CAASPP and the ELPAC—one or both—by reporting some lessons they learned in 2023–24. This year, additional efforts were made to collect information about the newly released CAST Interim Assessments and ELPAC Interim Assessments.

Improvements made in response to survey results are detailed in [chapter 12](#_Continuous_and_Systematic). The CDE and ETS used key recommendations from educators to implement positive changes in the next test administration year.

##### Overview

The California educators who responded provided specific, actionable insights about their test administration experience. This survey gathered information and data from educators who were part of the administration of CAASPP, the ELPAC, or both programs. Its goal was to highlight successes and identify areas for improvement, both immediate and long term.

Overall, California educators continue to express positive experiences in their preparations for administering CAASPP and the ELPAC.

##### Communication

During the 2023–24 test administration year, the CDE and ETS continued to streamline communications and provide LEAs with relevant information throughout the year. CAASPP and ELPAC monthly communications were sent throughout the administration with timely reminders and training announcements. In addition, proactive communications were sent to help remind LEA CAASPP coordinators of important actions needed for a successful administration, such as reminders to set up a test administration window, order materials, or enter scores into the Data Entry Interface, if needed.

### Accessibility Resources

The Every Student Succeeds Act reaffirms the importance of ensuring that assessments are accessible to special student populations, and the Individuals with Disabilities Education Act lays out monitoring requirements for students with disabilities. This section describes the accessibility resources used to support students in the Smarter Balanced Summative Assessments, as well as the procedures to identify and assign students with accommodations and designated supports. Finally, the number of students who were assigned accessibility resources is reported on the basis of available data.

The 2023–24 Smarter Balanced Summative Assessments offered commonly used accessibility resources available through the CAASPP computer-based testing platform, where applicable for the tested construct.

#### Accessibility Resource Categories

The purpose of universal tools, designated supports, and accommodations in testing is to provide *all* students with the opportunity to demonstrate what they know and what they are able to do. Universal tools, designated supports, and accommodations minimize or remove barriers that could otherwise prevent students from demonstrating their knowledge, skills, and achievement in a specific content area.

The CDE California Assessment Accessibility Resources Matrix (Accessibility Matrix) (CDE, 2023) is intended for school-level personnel and individualized education program (IEP) and Section 504 plan teams to select and administer the appropriate universal tools, designated supports, and accommodations as deemed necessary for individual students.

##### Universal Tools

Universal tools were available to all students by default, although they could be disabled if a student found them distracting. Each universal tool fell into one of two categories: embedded and non-embedded. Embedded universal tools were provided through the TDS (through the CAASPP secure browser), although they could be turned off by a test administrator.

The universal tools in the following subsections were available in the 2023–24 Smarter Balanced Summative Assessment administration.

###### Embedded

The following embedded universal tools were available to students testing in the secure browser:

* Breaks
* Calculator (grades six through eight and grade eleven)
* Digital notepad
* English dictionary (writing)
* English glossary
* Expandable items
* Expandable passages
* Global notes (writing)
* Highlighter
* Keyboard navigation
* Line reader
* Mark for review
* Mathematics tools (for example, ruler, protractor)
* Spell check
* Strikethrough
* Thesaurus
* Writing tools (for example, bold, italic, bullets, undo/redo)
* Zoom (in/out)

###### Non-Embedded

The following non-embedded universal tools were available to students testing in the secure browser:

* Breaks
* English dictionary (writing)
* Scratch paper

The following non-embedded universal tools were available to students taking the PPT:

* Breaks
* Calculator (for calculator-allowed mathematics sessions only)
* English dictionary
* English glossary
* Highlighter
* Line reader
* Mark for review
* Scratch paper
* Strikethrough
* Thesaurus

##### Designated Supports

Designated supports were available to all students when determined for use by an educator or team of educators (with parent/guardian and student input, as appropriate) or specified in the student’s IEP or Section 504 plan. These are assigned through the test settings in TOMS. The designated supports each fell into one of two categories: embedded and non-embedded. Embedded designated supports were provided through the Student Testing Interface (through the CAASPP secure browser).

The designated supports in the following subsections were available in the 2023–24 Smarter Balanced Summative Assessment administration.

###### Embedded

The following embedded designated supports were available to students testing in the secure browser:

* Color contrast
* Masking
* Mouse pointer (size and color)
* Permissive mode
* Print (font) size
* Streamline
* Text-to-speech (items and stimuli)
* Text-to-speech in Spanish (mathematics items)
* Translated text directions (Spanish)
* Translations (glossary)
* Translations (Spanish stacked–dual language)
* Turn off any universal tool(s)

###### Non-Embedded

The following non-embedded designated supports were available to students testing in the secure browser:

* Amplification
* Bilingual dictionary
* Color contrast
* Color overlay
* Magnification
* Medical supports
* Noise buffers
* Read aloud (items)
* Read aloud (Spanish)
* Scribe (items)
* Separate setting (for example, most beneficial time, special lighting or acoustics, adaptive furniture)
* Simplified test directions
* Translated test directions

The following non-embedded designated supports were available to students taking the PPT:

* Bilingual dictionary
* Illustration glossary
* Magnification
* Masking
* Medical supports
* Noise buffers
* Printed test directions in English
* Read aloud (items)
* Scribe
* Separate setting (for example, most beneficial time, special lighting or acoustics, adaptive furniture)
* Simplified test directions
* Translated test directions (including American Sign Language [ASL])
* Translations (glossary)

##### Accommodations

Accommodations are changes in procedures or materials that increased equitable access during CAASPP administration and are permitted to all eligible students if specified in the student’s IEP or Section 504 plan. Assessment accommodations for students who needed them generated valid assessment results; they allowed these students to show what they know and can do. Accommodations did not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

The accommodations in the following subsections were available in the 2023–24 Smarter Balanced Summative Assessment administration.

###### Embedded

The following embedded accommodations were available to students testing in the secure browser:

* American Sign Language (videos)
* Audio transcript
* Braille (embosser and refreshable)
* Closed-captioning (allowed but not currently used)
* Speech-to-text
* Text-to-speech (passages)
* Word prediction

###### Non-Embedded

The following non-embedded accommodations were available to students testing in the secure browser:

* 100s number table
* Abacus
* Alternate response options
* Calculator (grades six through eight and grade eleven)
* Multiplication table (grades four through eight and grade eleven)
* Print-on-demand
* Read aloud (passages)
* Scribe (writing)
* Speech-to-text
* Word prediction

The following non-embedded accommodations were available to students taking the PPT:

* 100s number table
* Abacus
* Alternate response options
* American Sign Language
* Braille
* Breaks
* Calculator (for calculator-allowed mathematics sessions only)
* Large print
* Multiplication table
* Read aloud (passages)
* Scribe

##### Unlisted Resources

An unlisted resource is an instructional support a student regularly uses in daily instruction, assessment, or both, and has not been previously identified as a universal tool, designated support, or accommodation. The CDE Accessibility Matrix included an inventory of unlisted resources that were already identified and were preapproved (CDE, 2023). During the 2023–24 CAASPP administration, an LEA CAASPP coordinator or a site CAASPP coordinator would use TOMS to submit a request for use of an unlisted resource. A preidentified, preapproved unlisted resource request was automatically approved. A request for an unlisted resource that was not preidentified was sent to the CDE for review and adjudication.

Unlisted resources are non-embedded resources that are made available if specified in the eligible student’s IEP or Section 504 plan and only upon approval by the CDE. Unlisted resources that changed the construct of an assessment and were approved were flagged as causing a change in construct. Test results for a student using an unlisted resource that was approved but that changed the construct of what was being tested were considered invalid for reporting purposes.

Preidentified unlisted resources applicable to the Smarter Balanced Summative Assessments are as follows:

* Bilingual dictionary
* Calculator (mathematics only)
* English dictionary
* Math tools (mathematics only)
* Signed exact English
* Thesaurus
* Translated word lists
* Translations (not provided by Smarter Balanced)

The LEA CAASPP coordinator or site CAASPP coordinator was required to submit a request for the use of an unlisted resource to the CDE a minimum of 10 business days before the student’s first day of testing.

#### Identification and Selection

All eligible students enrolled in a California public school participate in the CAASPP System, including students with disabilities and EL students. The Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines* (Smarter Balanced, 2023) and the CDE Accessibility Matrix (CDE, 2023) are intended for school-level personnel and IEP and Section 504 plan teams to select and administer the appropriate universal tools, designated supports, and accommodations as deemed necessary for individual students.[[6]](#footnote-7) The Smarter Balanced Summative Assessments follow the Smarter Balanced recommendations for use (Smarter Balanced, 2018).

The *Guidelines* apply to all participating students and promote an individualized approach to the implementation of assessment practices. Another web page, the Smarter Balanced Accessibility Strategies web page on the Tools for Teachers website (Smarter Balanced, 2024), connects the assessment resources described in the *Guidelines* with associated classroom practices.

The full list of the universal tools, designated supports, and accommodations used in CAASPP computer-based assessments, including the Smarter Balanced Summative Assessments, is documented in the CDE Accessibility Matrix. Most embedded and non-embedded universal tools, designated supports, and accommodations listed in parts 1, 2, and 3 of the Accessibility Matrix are available for the Smarter Balanced Summative Assessments through the computer-based testing interface or, in the case of non-embedded resources, from the school or LEA. Part 5 of the Accessibility Matrix includes approved unlisted resources. School-level personnel, IEP teams, and Section 504 teams used the CDE Accessibility Matrix when deciding how best to support the student’s test-taking experience. Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014),provides suggestions for implementation of these resources.

Test administrators and test examiners are given the opportunity to administer the CAASPP practice and training tests so that students have the opportunity to familiarize themselves with a designated support or accommodation prior to testing. (Refer to section [*5.5 Practice and Training Tests*](#_Practice_and_Training) for additional information.)

#### Assignment

Designated supports and accommodations are assigned to individual students on the basis of identified student need. Such assignments are implemented in TOMS by the LEA CAASPP coordinator or site CAASPP coordinator, either through individual assignment through the student’s profile in TOMS or in a batch upload for multiple students. When the batch upload process was used, settings were uploaded into TOMS using a spreadsheet with data that had either been entered into a template downloaded from TOMS; or created by selecting and entering information into the web-based ISAAP Tool. The ISAAP Tool could be used by LEAs in conjunction with the *Guidelines* and the *2023–24* CAASPP and ELPAC Accessibility Guide (CDE, 2024a), as well as with state regulations and policies (such as the Accessibility Matrix) related to assessment accessibility*.*

The embedded designated supports and accommodations were delivered to the student through the TDS at the time of testing; the non-embedded designated supports and accommodations were provided at the time of testing to the student by the LEA. Refer to section [*1.9 Systems Overview and Functionality*](#_Systems_Overview_and) in [*Chapter 1: Introduction*](#_Introduction) for more details regarding the TDS.

Once a student’s IEP or Section 504 plan team decided which accessibility resource(s) the student should use, LEA CAASPP coordinators and site CAASPP coordinators used TOMS to assign designated supports and accommodations to students prior to the start of a test session.

There were three ways a student’s accessibility resource(s) could be assigned:

1. Using the ISAAP Tool to identify the accessibility resource(s) and then uploading the spreadsheet it creates into TOMS (This process is discussed in more detail in subsection [*5.4.2 Identification and Selection*](#_Identification_and_Selection_2).)
2. Using the Online Student Test Settings template to enter students’ assignments and then uploading the spreadsheet into TOMS
3. Entering assignments for each student individually in TOMS

If a student’s IEP or Section 504 plan team identified and designated a resource not identified in the CDE Accessibility Matrix, the LEA CAASPP coordinator or site CAASPP coordinator needed to submit a request for an unlisted resource to be approved by the CDE. The CDE then determined whether the requested unlisted resource changed the construct being measured before the student started testing. Once the requested unlisted resource was approved, the student could begin the assessment using the approved resource.

[Appendix 5.A](#_Appendix_5.A:_Accessibility_1) provides information on the number of students who were assigned accommodations and designated supports.

#### Delivery of Embedded and Non-Embedded Resources to Students

Universal tools, designated supports, and accommodations can be delivered as either embedded or non-embedded resources. Embedded resources are digitally delivered features or settings available as part of the technology platform for Smarter Balanced Summative Assessment testing. Examples of embedded resources include the expandable items, color contrast, and masking.

Non-embedded resources are available, when provided by the LEA, for both computer-based assessments and PPTs. These resources are not part of the technology platform for the computer-administered Smarter Balanced Summative Assessments. Examples of non-embedded resources include magnification, noise buffers, and the use of a scribe.

Refer to subsection [*5.4.1 Accessibility Resource Categories*](#_Accessibility_Resource_Categories_2) for a detailed description of the accessibility resources available to students taking the Smarter Balanced Summative Assessments.

#### Usage of Designated Supports and Accommodations

LEA CAASPP coordinators and site CAASPP coordinators were responsible for assigning their students’ test settings in TOMS before testing occurred and providing the necessary resources during testing. If a test setting was not applied before testing, then a STAIRS incident was to be submitted to reset the assessment so the student could be retested with the correct accommodation or designated support. If a test setting was accidentally assigned to a student, then a STAIRS incident was also to be submitted to reset the assessment so the student could be retested without the accommodation or designated support.

After schools and LEAs assigned eligible students to accommodations or designated supports, the Cambium Assessment, Inc. (CAI) TDS provided and captured whether a certain accommodation or designated support (or multiple accommodations or designated supports) was used by a student—that is, the student interacted with a control for the resource, such as a button—as the student progressed through the assessment.

Table 5.2 and table 5.3 report the number of students who, based on the availability of data, were assigned to a certain accommodation or designated support and actually used this accommodation or designated support at least once in ELA and mathematics, respectively. Embedded accessibility resources are those that are part of the computer-based TDS, whereas non-embedded accessibility resources are provided outside of that system. However, because the TDS does not capture the usage of all embedded resources and cannot capture the usage of any non-embedded resources, these tables report only on a limited subset of the embedded resources.

Types of accommodations and designated supports—labeled “ACC” and “DS” in the *Resource Type* column—included in table 5.2 and table 5.3 are as follows:

* **Text-to-Speech:** Text is read aloud to the student via embedded text-to-speech technology. It includes text-to-speech passages and text-to-speech items for accommodations.
* **Speech-to-Text:** Voice recognition allows students to use their voices as input devices to the computer to dictate responses.
* **ASL:** ASL videos are available for any item that has a listening component. The ASL human signer and the signed test content are viewed on the same screen.
* **Print-on-Demand:** Paper copies of passages and stimuli, items, or all of these are printed for students.
* **Masking:** This resource involves blocking off content that is not of immediate need or that may be distracting to the student.
* **Audio Transcript:** This resource allows students to view a transcript of the audio content for the current test page. This is useful for students with visual impairments who are accustomed to accessing information presented via audio in the form of braille.

Results show that the number of students who were assigned for the accessibility resources was greater than the number of students who actually used the accessibility resources across assessments.

Table 5.2 Summary of Accommodations and Designated Supports Used by Students—ELA

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Content Area** | **Grade Level** | **Opportunity** | **Support** | **Resource Type** | **Students Assigned** | **Students Used** |
| ELA | All Grades | All | Any Tracked Resource | N/A | 511,848 | 278,923 |
| ELA | All Grades | CAT | Any Tracked Resource | N/A | 507,397 | 232,512 |
| ELA | All Grades | PT | Any Tracked Resource | N/A | 511,066 | 232,378 |
| ELA | All Grades | CAT | Embedded American Sign Language | ACC | 1,025 | 364 |
| ELA | All Grades | CAT | Embedded Audio Transcript | ACC | 1,078 | 41 |
| ELA | All Grades | CAT | Embedded Speech-to-Text | ACC | 79,940 | 10,081 |
| ELA | All Grades | CAT | Embedded Text-to-Speech Passages | ACC | 203,442 | 105,131 |
| ELA | All Grades | CAT | Non-Embedded Print-on-Demand | ACC | 662 | 100 |
| ELA | All Grades | CAT | Embedded Masking | DS | 77,324 | 3,690 |
| ELA | All Grades | CAT | Embedded Text-to-Speech Items | DS | 450,563 | 208,907 |
| ELA | All Grades | PT | Embedded American Sign Language | ACC | 1,047 | 0 |
| ELA | All Grades | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | All Grades | PT | Embedded Speech-to-Text | ACC | 81,232 | 39,485 |
| ELA | All Grades | PT | Embedded Text-to-Speech Passages | ACC | 205,572 | 124,584 |
| ELA | All Grades | PT | Non-Embedded Print-on-Demand | ACC | 676 | 180 |
| ELA | All Grades | PT | Embedded Masking | DS | 77,706 | 4,121 |
| ELA | All Grades | PT | Embedded Text-to-Speech Items | DS | 454,287 | 173,204 |
| ELA | 3 | All | Any Tracked Resource | N/A | 85,707 | 53,336 |
| ELA | 3 | CAT | Any Tracked Resource | N/A | 84,633 | 43,064 |
| ELA | 3 | PT | Any Tracked Resource | N/A | 85,572 | 47,704 |
| ELA | 3 | CAT | Embedded American Sign Language | ACC | 139 | 69 |
| ELA | 3 | CAT | Embedded Audio Transcript | ACC | 132 | 1 |
| ELA | 3 | CAT | Embedded Speech-to-Text | ACC | 12,270 | 272 |
| ELA | 3 | CAT | Embedded Text-to-Speech Passages | ACC | 25,805 | 13,485 |
| ELA | 3 | CAT | Non-Embedded Print-on-Demand | ACC | 99 | 15 |
| ELA | 3 | CAT | Embedded Masking | DS | 10,008 | 391 |
| ELA | 3 | CAT | Embedded Text-to-Speech Items | DS | 76,808 | 40,352 |
| ELA | 3 | PT | Embedded American Sign Language | ACC | 142 | 0 |
| ELA | 3 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 3 | PT | Embedded Speech-to-Text | ACC | 12,600 | 8,470 |
| ELA | 3 | PT | Embedded Text-to-Speech Passages | ACC | 26,308 | 18,185 |
| ELA | 3 | PT | Non-Embedded Print-on-Demand | ACC | 94 | 33 |
| ELA | 3 | PT | Embedded Masking | DS | 10,068 | 476 |
| ELA | 3 | PT | Embedded Text-to-Speech Items | DS | 77,796 | 39,514 |
| ELA | 4 | All | Any Tracked Resource | N/A | 88,240 | 54,430 |
| ELA | 4 | CAT | Any Tracked Resource | N/A | 87,344 | 43,993 |
| ELA | 4 | PT | Any Tracked Resource | N/A | 88,137 | 47,834 |
| ELA | 4 | CAT | Embedded American Sign Language | ACC | 126 | 45 |
| ELA | 4 | CAT | Embedded Audio Transcript | ACC | 142 | 5 |
| ELA | 4 | CAT | Embedded Speech-to-Text | ACC | 15,121 | 358 |
| ELA | 4 | CAT | Embedded Text-to-Speech Passages | ACC | 31,495 | 17,588 |
| ELA | 4 | CAT | Non-Embedded Print-on-Demand | ACC | 117 | 28 |
| ELA | 4 | CAT | Embedded Masking | DS | 11,762 | 605 |
| ELA | 4 | CAT | Embedded Text-to-Speech Items | DS | 79,038 | 40,800 |
| ELA | 4 | PT | Embedded American Sign Language | ACC | 132 | 0 |
| ELA | 4 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 4 | PT | Embedded Speech-to-Text | ACC | 15,442 | 10,137 |
| ELA | 4 | PT | Embedded Text-to-Speech Passages | ACC | 31,885 | 22,647 |
| ELA | 4 | PT | Non-Embedded Print-on-Demand | ACC | 118 | 36 |
| ELA | 4 | PT | Embedded Masking | DS | 11,835 | 688 |
| ELA | 4 | PT | Embedded Text-to-Speech Items | DS | 79,892 | 37,749 |
| ELA | 5 | All | Any Tracked Resource | N/A | 88,348 | 53,145 |
| ELA | 5 | CAT | Any Tracked Resource | N/A | 87,513 | 44,509 |
| ELA | 5 | PT | Any Tracked Resource | N/A | 88,278 | 45,315 |
| ELA | 5 | CAT | Embedded American Sign Language | ACC | 161 | 49 |
| ELA | 5 | CAT | Embedded Audio Transcript | ACC | 167 | 6 |
| ELA | 5 | CAT | Embedded Speech-to-Text | ACC | 16,639 | 360 |
| ELA | 5 | CAT | Embedded Text-to-Speech Passages | ACC | 33,416 | 19,519 |
| ELA | 5 | CAT | Non-Embedded Print-on-Demand | ACC | 120 | 17 |
| ELA | 5 | CAT | Embedded Masking | DS | 12,308 | 622 |
| ELA | 5 | CAT | Embedded Text-to-Speech Items | DS | 79,183 | 41,414 |
| ELA | 5 | PT | Embedded American Sign Language | ACC | 166 | 0 |
| ELA | 5 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 5 | PT | Embedded Speech-to-Text | ACC | 16,962 | 9,527 |
| ELA | 5 | PT | Embedded Text-to-Speech Passages | ACC | 33,840 | 23,450 |
| ELA | 5 | PT | Non-Embedded Print-on-Demand | ACC | 125 | 33 |
| ELA | 5 | PT | Embedded Masking | DS | 12,396 | 692 |
| ELA | 5 | PT | Embedded Text-to-Speech Items | DS | 79,970 | 35,060 |
| ELA | 6 | All | Any Tracked Resource | N/A | 75,327 | 42,980 |
| ELA | 6 | CAT | Any Tracked Resource | N/A | 74,819 | 37,345 |
| ELA | 6 | PT | Any Tracked Resource | N/A | 75,235 | 35,215 |
| ELA | 6 | CAT | Embedded American Sign Language | ACC | 145 | 44 |
| ELA | 6 | CAT | Embedded Audio Transcript | ACC | 161 | 7 |
| ELA | 6 | CAT | Embedded Speech-to-Text | ACC | 13,230 | 4,878 |
| ELA | 6 | CAT | Embedded Text-to-Speech Passages | ACC | 32,841 | 19,142 |
| ELA | 6 | CAT | Non-Embedded Print-on-Demand | ACC | 89 | 17 |
| ELA | 6 | CAT | Embedded Masking | DS | 11,655 | 775 |
| ELA | 6 | CAT | Embedded Text-to-Speech Items | DS | 66,779 | 32,541 |
| ELA | 6 | PT | Embedded American Sign Language | ACC | 145 | 0 |
| ELA | 6 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 6 | PT | Embedded Speech-to-Text | ACC | 13,376 | 5,716 |
| ELA | 6 | PT | Embedded Text-to-Speech Passages | ACC | 33,090 | 21,543 |
| ELA | 6 | PT | Non-Embedded Print-on-Demand | ACC | 93 | 24 |
| ELA | 6 | PT | Embedded Masking | DS | 11,676 | 762 |
| ELA | 6 | PT | Embedded Text-to-Speech Items | DS | 67,180 | 25,478 |
| ELA | 7 | All | Any Tracked Resource | N/A | 66,676 | 33,575 |
| ELA | 7 | CAT | Any Tracked Resource | N/A | 66,301 | 28,363 |
| ELA | 7 | PT | Any Tracked Resource | N/A | 66,533 | 26,494 |
| ELA | 7 | CAT | Embedded American Sign Language | ACC | 152 | 56 |
| ELA | 7 | CAT | Embedded Audio Transcript | ACC | 139 | 7 |
| ELA | 7 | CAT | Embedded Speech-to-Text | ACC | 10,231 | 2,211 |
| ELA | 7 | CAT | Embedded Text-to-Speech Passages | ACC | 31,617 | 15,462 |
| ELA | 7 | CAT | Non-Embedded Print-on-Demand | ACC | 60 | 10 |
| ELA | 7 | CAT | Embedded Masking | DS | 10,203 | 550 |
| ELA | 7 | CAT | Embedded Text-to-Speech Items | DS | 58,598 | 24,314 |
| ELA | 7 | PT | Embedded American Sign Language | ACC | 159 | 0 |
| ELA | 7 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 7 | PT | Embedded Speech-to-Text | ACC | 10,294 | 3,130 |
| ELA | 7 | PT | Embedded Text-to-Speech Passages | ACC | 31,779 | 17,638 |
| ELA | 7 | PT | Non-Embedded Print-on-Demand | ACC | 62 | 22 |
| ELA | 7 | PT | Embedded Masking | DS | 10,205 | 621 |
| ELA | 7 | PT | Embedded Text-to-Speech Items | DS | 58,857 | 17,429 |
| ELA | 8 | All | Any Tracked Resource | N/A | 63,901 | 28,696 |
| ELA | 8 | CAT | Any Tracked Resource | N/A | 63,562 | 24,529 |
| ELA | 8 | PT | Any Tracked Resource | N/A | 63,795 | 21,579 |
| ELA | 8 | CAT | Embedded American Sign Language | ACC | 151 | 44 |
| ELA | 8 | CAT | Embedded Audio Transcript | ACC | 164 | 12 |
| ELA | 8 | CAT | Embedded Speech-to-Text | ACC | 9,083 | 1,620 |
| ELA | 8 | CAT | Embedded Text-to-Speech Passages | ACC | 30,465 | 13,784 |
| ELA | 8 | CAT | Non-Embedded Print-on-Demand | ACC | 72 | 8 |
| ELA | 8 | CAT | Embedded Masking | DS | 10,320 | 432 |
| ELA | 8 | CAT | Embedded Text-to-Speech Items | DS | 56,318 | 21,118 |
| ELA | 8 | PT | Embedded American Sign Language | ACC | 151 | 0 |
| ELA | 8 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 8 | PT | Embedded Speech-to-Text | ACC | 9,164 | 2,079 |
| ELA | 8 | PT | Embedded Text-to-Speech Passages | ACC | 30,668 | 15,193 |
| ELA | 8 | PT | Non-Embedded Print-on-Demand | ACC | 76 | 19 |
| ELA | 8 | PT | Embedded Masking | DS | 10,368 | 450 |
| ELA | 8 | PT | Embedded Text-to-Speech Items | DS | 56,549 | 13,750 |
| ELA | 11 | All | Any Tracked Resource | N/A | 43,649 | 12,761 |
| ELA | 11 | CAT | Any Tracked Resource | N/A | 43,225 | 10,709 |
| ELA | 11 | PT | Any Tracked Resource | N/A | 43,516 | 8,237 |
| ELA | 11 | CAT | Embedded American Sign Language | ACC | 151 | 57 |
| ELA | 11 | CAT | Embedded Audio Transcript | ACC | 173 | 3 |
| ELA | 11 | CAT | Embedded Speech-to-Text | ACC | 3,366 | 382 |
| ELA | 11 | CAT | Embedded Text-to-Speech Passages | ACC | 17,803 | 6,151 |
| ELA | 11 | CAT | Non-Embedded Print-on-Demand | ACC | 105 | 5 |
| ELA | 11 | CAT | Embedded Masking | DS | 11,068 | 315 |
| ELA | 11 | CAT | Embedded Text-to-Speech Items | DS | 33,839 | 8,368 |
| ELA | 11 | PT | Embedded American Sign Language | ACC | 152 | 0 |
| ELA | 11 | PT | Embedded Audio Transcript | ACC | 0 | 0 |
| ELA | 11 | PT | Embedded Speech-to-Text | ACC | 3,394 | 426 |
| ELA | 11 | PT | Embedded Text-to-Speech Passages | ACC | 18,002 | 5,928 |
| ELA | 11 | PT | Non-Embedded Print-on-Demand | ACC | 108 | 13 |
| ELA | 11 | PT | Embedded Masking | DS | 11,158 | 432 |
| ELA | 11 | PT | Embedded Text-to-Speech Items | DS | 34,043 | 4,224 |

Table 5.3 Summary of Accommodations and Designated Supports Used by Students—Mathematics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Content Area** | **Grade Level** | **Opportunity** | **Support** | **Resource** **Type** | **Students Assigned** | **Students Used** |
| Mathematics | All Grades | All | Any Tracked Resource | N/A | 485,900 | 199,290 |
| Mathematics | All Grades | CAT | Any Tracked Resource | N/A | 484,384 | 158,345 |
| Mathematics | All Grades | PT | Any Tracked Resource | N/A | 485,423 | 165,803 |
| Mathematics | All Grades | CAT | Embedded American Sign Language | ACC | 1,023 | 550 |
| Mathematics | All Grades | CAT | Embedded Speech-to-Text | ACC | 78,929 | 1,164 |
| Mathematics | All Grades | CAT | Non-Embedded Print-on-Demand | ACC | 719 | 69 |
| Mathematics | All Grades | CAT | Embedded Masking | DS | 77,091 | 2,203 |
| Mathematics | All Grades | CAT | Embedded Text-to-Speech | DS | 435,059 | 155,630 |
| Mathematics | All Grades | PT | Embedded American Sign Language | ACC | 1,031 | 435 |
| Mathematics | All Grades | PT | Embedded Speech-to-Text | ACC | 79,276 | 22,085 |
| Mathematics | All Grades | PT | Non-Embedded Print-on-Demand | ACC | 723 | 98 |
| Mathematics | All Grades | PT | Embedded Masking | DS | 77,220 | 1,366 |
| Mathematics | All Grades | PT | Embedded Text-to-Speech | DS | 436,105 | 156,068 |
| Mathematics | 3 | All | Any Tracked Resource | N/A | 86,880 | 50,068 |
| Mathematics | 3 | CAT | Any Tracked Resource | N/A | 86,543 | 41,838 |
| Mathematics | 3 | PT | Any Tracked Resource | N/A | 86,818 | 43,647 |
| Mathematics | 3 | CAT | Embedded American Sign Language | ACC | 142 | 88 |
| Mathematics | 3 | CAT | Embedded Speech-to-Text | ACC | 12,194 | 272 |
| Mathematics | 3 | CAT | Non-Embedded Print-on-Demand | ACC | 91 | 10 |
| Mathematics | 3 | CAT | Embedded Masking | DS | 10,047 | 188 |
| Mathematics | 3 | CAT | Embedded Text-to-Speech | DS | 80,986 | 41,557 |
| Mathematics | 3 | PT | Embedded American Sign Language | ACC | 143 | 87 |
| Mathematics | 3 | PT | Embedded Speech-to-Text | ACC | 12,289 | 5,425 |
| Mathematics | 3 | PT | Non-Embedded Print-on-Demand | ACC | 90 | 15 |
| Mathematics | 3 | PT | Embedded Masking | DS | 10,061 | 150 |
| Mathematics | 3 | PT | Embedded Text-to-Speech | DS | 81,253 | 41,643 |
| Mathematics | 4 | All | Any Tracked Resource | N/A | 87,878 | 46,275 |
| Mathematics | 4 | CAT | Any Tracked Resource | N/A | 87,573 | 37,282 |
| Mathematics | 4 | PT | Any Tracked Resource | N/A | 87,832 | 39,849 |
| Mathematics | 4 | CAT | Embedded American Sign Language | ACC | 132 | 81 |
| Mathematics | 4 | CAT | Embedded Speech-to-Text | ACC | 14,950 | 305 |
| Mathematics | 4 | CAT | Non-Embedded Print-on-Demand | ACC | 121 | 15 |
| Mathematics | 4 | CAT | Embedded Masking | DS | 11,810 | 261 |
| Mathematics | 4 | CAT | Embedded Text-to-Speech | DS | 80,539 | 36,917 |
| Mathematics | 4 | PT | Embedded American Sign Language | ACC | 132 | 68 |
| Mathematics | 4 | PT | Embedded Speech-to-Text | ACC | 15,029 | 6,374 |
| Mathematics | 4 | PT | Non-Embedded Print-on-Demand | ACC | 123 | 25 |
| Mathematics | 4 | PT | Embedded Masking | DS | 11,824 | 193 |
| Mathematics | 4 | PT | Embedded Text-to-Speech | DS | 80,812 | 37,371 |
| Mathematics | 5 | All | Any Tracked Resource | N/A | 87,408 | 43,124 |
| Mathematics | 5 | CAT | Any Tracked Resource | N/A | 87,182 | 33,393 |
| Mathematics | 5 | PT | Any Tracked Resource | N/A | 87,319 | 36,879 |
| Mathematics | 5 | CAT | Embedded American Sign Language | ACC | 159 | 87 |
| Mathematics | 5 | CAT | Embedded Speech-to-Text | ACC | 16,480 | 280 |
| Mathematics | 5 | CAT | Non-Embedded Print-on-Demand | ACC | 127 | 16 |
| Mathematics | 5 | CAT | Embedded Masking | DS | 12,406 | 389 |
| Mathematics | 5 | CAT | Embedded Text-to-Speech | DS | 79,240 | 32,888 |
| Mathematics | 5 | PT | Embedded American Sign Language | ACC | 159 | 69 |
| Mathematics | 5 | PT | Embedded Speech-to-Text | ACC | 16,532 | 5,738 |
| Mathematics | 5 | PT | Non-Embedded Print-on-Demand | ACC | 126 | 16 |
| Mathematics | 5 | PT | Embedded Masking | DS | 12,418 | 305 |
| Mathematics | 5 | PT | Embedded Text-to-Speech | DS | 79,406 | 34,526 |
| Mathematics | 6 | All | Any Tracked Resource | N/A | 71,672 | 28,238 |
| Mathematics | 6 | CAT | Any Tracked Resource | N/A | 71,445 | 22,064 |
| Mathematics | 6 | PT | Any Tracked Resource | N/A | 71,612 | 22,725 |
| Mathematics | 6 | CAT | Embedded American Sign Language | ACC | 139 | 76 |
| Mathematics | 6 | CAT | Embedded Speech-to-Text | ACC | 13,032 | 171 |
| Mathematics | 6 | CAT | Non-Embedded Print-on-Demand | ACC | 104 | 12 |
| Mathematics | 6 | CAT | Embedded Masking | DS | 11,587 | 448 |
| Mathematics | 6 | CAT | Embedded Text-to-Speech | DS | 63,991 | 21,584 |
| Mathematics | 6 | PT | Embedded American Sign Language | ACC | 140 | 69 |
| Mathematics | 6 | PT | Embedded Speech-to-Text | ACC | 13,080 | 2,968 |
| Mathematics | 6 | PT | Non-Embedded Print-on-Demand | ACC | 105 | 18 |
| Mathematics | 6 | PT | Embedded Masking | DS | 11,603 | 273 |
| Mathematics | 6 | PT | Embedded Text-to-Speech | DS | 64,173 | 21,222 |
| Mathematics | 7 | All | Any Tracked Resource | N/A | 60,723 | 16,322 |
| Mathematics | 7 | CAT | Any Tracked Resource | N/A | 60,545 | 12,888 |
| Mathematics | 7 | PT | Any Tracked Resource | N/A | 60,651 | 11,603 |
| Mathematics | 7 | CAT | Embedded American Sign Language | ACC | 156 | 94 |
| Mathematics | 7 | CAT | Embedded Speech-to-Text | ACC | 9,987 | 78 |
| Mathematics | 7 | CAT | Non-Embedded Print-on-Demand | ACC | 77 | 5 |
| Mathematics | 7 | CAT | Embedded Masking | DS | 10,173 | 395 |
| Mathematics | 7 | CAT | Embedded Text-to-Speech | DS | 53,377 | 12,431 |
| Mathematics | 7 | PT | Embedded American Sign Language | ACC | 158 | 54 |
| Mathematics | 7 | PT | Embedded Speech-to-Text | ACC | 10,013 | 758 |
| Mathematics | 7 | PT | Non-Embedded Print-on-Demand | ACC | 77 | 10 |
| Mathematics | 7 | PT | Embedded Masking | DS | 10,197 | 161 |
| Mathematics | 7 | PT | Embedded Text-to-Speech | DS | 53,478 | 11,015 |
| Mathematics | 8 | All | Any Tracked Resource | N/A | 57,641 | 12,563 |
| Mathematics | 8 | CAT | Any Tracked Resource | N/A | 57,490 | 9,196 |
| Mathematics | 8 | PT | Any Tracked Resource | N/A | 57,556 | 9,275 |
| Mathematics | 8 | CAT | Embedded American Sign Language | ACC | 149 | 69 |
| Mathematics | 8 | CAT | Embedded Speech-to-Text | ACC | 8,968 | 51 |
| Mathematics | 8 | CAT | Non-Embedded Print-on-Demand | ACC | 87 | 5 |
| Mathematics | 8 | CAT | Embedded Masking | DS | 10,382 | 355 |
| Mathematics | 8 | CAT | Embedded Text-to-Speech | DS | 50,035 | 8,796 |
| Mathematics | 8 | PT | Embedded American Sign Language | ACC | 149 | 47 |
| Mathematics | 8 | PT | Embedded Speech-to-Text | ACC | 8,994 | 709 |
| Mathematics | 8 | PT | Non-Embedded Print-on-Demand | ACC | 89 | 6 |
| Mathematics | 8 | PT | Embedded Masking | DS | 10,402 | 161 |
| Mathematics | 8 | PT | Embedded Text-to-Speech | DS | 50,089 | 8,707 |
| Mathematics | 11 | All | Any Tracked Resource | N/A | 33,698 | 2,700 |
| Mathematics | 11 | CAT | Any Tracked Resource | N/A | 33,606 | 1,684 |
| Mathematics | 11 | PT | Any Tracked Resource | N/A | 33,635 | 1,825 |
| Mathematics | 11 | CAT | Embedded American Sign Language | ACC | 146 | 55 |
| Mathematics | 11 | CAT | Embedded Speech-to-Text | ACC | 3,318 | 7 |
| Mathematics | 11 | CAT | Non-Embedded Print-on-Demand | ACC | 112 | 6 |
| Mathematics | 11 | CAT | Embedded Masking | DS | 10,686 | 167 |
| Mathematics | 11 | CAT | Embedded Text-to-Speech | DS | 26,891 | 1,457 |
| Mathematics | 11 | PT | Embedded American Sign Language | ACC | 150 | 41 |
| Mathematics | 11 | PT | Embedded Speech-to-Text | ACC | 3,339 | 113 |
| Mathematics | 11 | PT | Non-Embedded Print-on-Demand | ACC | 113 | 8 |
| Mathematics | 11 | PT | Embedded Masking | DS | 10,715 | 123 |
| Mathematics | 11 | PT | Embedded Text-to-Speech | DS | 26,894 | 1,584 |

### Practice and Training Tests

Practice and training tests are available publicly to LEA staff, students, parents/guardians, and any other individual for the Smarter Balanced Summative Assessments. These tests simulate the experience of the computer-based Smarter Balanced Summative Assessments to allow anyone to experience the assessment. For the 2023–24 school year, accommodated versions of Smarter Balanced Summative Assessment practice and training tests were developed to include all accessibility resources—including braille, closed-captioning, text-to-speech, and audio transcripts—available on the assessment.

Students can access practice and training tests using a web browser. They allow students and administrators to familiarize themselves with the user interface and components of the TDS and help maintain the standardization of test administration. Practice and training tests are available through the Practice and Training Test website linked on the Practice and Training Tests web page on the CAASPP & ELPAC Website.

The practice tests, offered at each grade level, were released to prepare students for the Smarter Balanced Summative Assessments. These tests more closely simulate the Smarter Balanced Summative Assessments’ length and complexity and align with the Smarter Balanced Summative Assessment blueprint.

The grade-level specific training tests can be taken by students in all tested grade levels. Many unique item types available on the operational assessment are covered in the training tests. The scoring guides for the practice and training tests are available on the Practice and Training Test Resources web page on the CAASPP & ELPAC Website.

### Test Security and Confidentiality

For the operational Smarter Balanced Summative Assessments, every person who worked with the assessments, communicated test results, or received testing information was responsible for maintaining the security and confidentiality of the assessments, including CDE staff, ETS staff, ETS subcontractors, LEA staff connected with assessments, school assessment coordinators, students, parents/guardians, and teachers. The ETS Code of Ethics required that all test information, including tangible materials (for example, test items), confidential files (for example, those containing personally identifiable student information), and processes related to test administration (for example, the configurations of secure servers), were kept secure. ETS had systems in place that maintained tight security for test items and test results, as well as for student data. To ensure security for all assessments that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI), which is described in the next subsection.

All assessments within the CAASPP System, as well as the confidentiality of student information, should be protected to ensure the validity, reliability, and fairness of the results. As stated in *Standard 7.9* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), “The documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session” (p. 128).

This section of the *Smarter Balanced Summative Assessments Technical Report* describes the measures intended to prevent potential test security incidents prior to testing and the actions that were taken to handle security incidents occurring during or after the testing window using the STAIRS process.

#### The ETS Office of Testing Integrity

The OTI is a division of ETS that provides quality-assurance services for all testing programs managed by ETS. This division resides in the ETS legal department. The Office of Professional Standards Compliance at ETS publishes and maintains the *ETS Standards for Quality and Fairness* (2014), which supports OTI goals and activities. The *ETS Standards for Quality and Fairness* provides guidelines to help ETS staff design, develop, and deliver technically sound, fair, and beneficial products and services and help the public and auditors evaluate those products and services.

The OTI mission is to

* prevent test security violations;
* minimize any testing security violations that can impact the fairness of testing;
* minimize and investigate any security breach that threatens the validity of the interpretation of test scores; and
* report on security activities.

The OTI helps prevent misconduct on the part of students and administrators, detects potential misconduct through empirically established indicators, and resolves situations involving misconduct in a fair and equitable way that reflects the laws and professional standards governing the integrity of testing. The OTI also implements policies designed to detect and block technologies used to gain an unfair advantage.

In its pursuit of enforcing secure testing practices, the OTI strives to safeguard the various processes involved in an assessment development and administration cycle. For the Smarter Balanced Summative Assessments, those processes included the following:

* Assessment development
* Item and data review
* Item banking
* Transfer of forms and items to the CDE and CAI
* Security of electronic files using a firewall
* Test administration
* Test delivery
* Processing and scoring
* Data management
* Statistical analysis
* Student confidentiality

#### Procedures to Maintain Standardization of Test Security

Test security requires the accounting of all secure materials—including computer-based summative test items and student data—before, during, and after each test administration. The LEA CAASPP coordinator is responsible for keeping all electronic test materials secure, keeping student information confidential, and making sure the site CAASPP coordinators and test administrators are properly trained regarding security policies and procedures.

The site CAASPP coordinator is responsible for mitigating test security incidents at the test site and for reporting incidents to the LEA CAASPP coordinator.

The test administrator is responsible for reporting testing incidents to the site CAASPP coordinator and securely destroying printed and digital media for items and passages generated by the print-on-demand feature of the TDS (CDE, 2024b).

The following measures ensured the security of the CAASPP:

* LEA CAASPP coordinators and site CAASPP coordinators must have electronically signed and submitted a “CAASPP Test Security Agreement for LEA CAASPP coordinators and site CAASPP coordinators” form in TOMS before ETS can grant the coordinators access to TOMS (5 *CCR* Section 859[d]).
* Anyone having access to the testing materials must have electronically signed and submitted a “Test Security Affidavit for Test Examiners, Test Administrators, Proctors, Translators, Scribes, and Any Other Person Having Access to CAASPP Tests” form in TOMS before receiving access to any testing materials (5 *CCR* Section 859[c]).
* Anyone having access to the testing materials but not having access to TOMS must have signed the CAASPP *Test Security Affidavit for Non-TOMS Users*, which was available as a web-based form, before receiving access to any testing materials.

In addition, it was the responsibility of every participant in the CAASPP System to report immediately any violation or suspected violation of test security or confidentiality. The test administrator reported to the site CAASPP coordinator or LEA CAASPP coordinator, who then submitted the incident using the STAIRS/Appeals process. Breach incidents were to be reported by the LEA CAASPP coordinator to the California Technical Assistance Center (CalTAC) and entered into STAIRS within 24 hours of the incident (5 *CCR* Section 859[e]).

#### Test Security Monitoring

The LEA and school testing staff were responsible for maintaining the security and confidentiality of testing materials and devices during the testing window and reporting any irregularities or breaches that occurred. ETS performed site visits and testing procedure audits at randomly selected LEAs and test sites throughout California during the test administration of CAASPP and the ELPAC operational assessments. Audits were performed before, during, and after test administrations to observe adherence to published procedures regarding the handling of testing materials and test administration guidelines.

For Tier 1 audits, 200 randomly selected sites, with their respective LEAs, responded to an online survey. ETS program staff conducted Tier 2 audits of LEAs in virtual meetings on the basis of significant security concerns such as the number of STAIRS cases submitted, communications from the LEA, security concerns reported to the CDE or ETS, or their responses to the audit survey. While there were no Tier 3 audits conducted during the 2023–24 test administration, these would have been conducted in person by a member of the ETS California program team if security issues were deemed to have disregarded rules of testing that resulted in a severe security breach. If, during a Tier 3 audit, the ETS staff member deemed the security issue to be of highest concern, then ETS OTI would be asked to escalate the investigation.

ETS provided a final summary report of audit findings to the CDE at the end of the test administration. ETS program management reported a summary of these findings to the CDE after a site visit.

#### Security of Electronic Files Using a Firewall

A firewall is software that prevents unauthorized entry to files, email, and other organization-specific information. All ETS data exchanges and internal email remain within the ETS firewall at all ETS locations, ranging from Princeton, New Jersey; to San Antonio, Texas; to Sacramento, California.

All electronic applications that are included in TOMS remain protected by the ETS firewall software at all times. Because of the sensitive nature of the student information processed by TOMS, the firewall plays a significant role in maintaining assurance of confidentiality among the users of this information.

Refer to section [*1.9 Systems Overview and Functionality*](#_Systems_Overview_and) in[*Chapter 1: Introduction*](#_Introduction) for more information on TOMS.

#### Transfer of Scores via Secure Data Exchange

Because of the confidential nature of test results, ETS uses secure file transfer protocol (SFTP) and encryption for all data file transfers; test data is never sent via email. SFTP is a method for reliable and exclusive routing of files. Files reside on a password-protected server that only authorized users can access. ETS shares an SFTP server with the CDE. On that site, ETS posts Microsoft Word and Excel files, Adobe Acrobat PDFs, or other document files for the CDE to review; the CDE returns reviewed materials in the same manner. Files are deleted upon retrieval.

The SFTP server is used as a conduit for the transfer of files; secure test data is stored only temporarily on the shared SFTP server. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems.

For the 2023–24 Smarter Balanced Summative Assessments, ETS entered information about the deliverable into a web form on a SharePoint website when a file was posted. A CDE staff member monitored this log throughout the day for updates to the status of deliverables and downloaded and deleted the file from the SFTP server when its status showed that it had been posted.

#### Data Management in the Secure Database

ETS maintains a secure database to house all student demographic data and assessment results. Information associated with each student has a database relationship to the LEA, school, and grade codes as the data is collected during testing. Only individuals with the appropriate credentials can access the data. ETS builds all interfaces with the most stringent security considerations, including interfaces with data encryption for databases that store test items and student data. ETS applies best and up-to-date security practices, including system-to-system authentication and authorization, in all solution designs.

All stored test content and student data is encrypted. Industry-standard secure protocols are used to transfer test content and student data from the ETS internal data center to any external systems. ETS complies with the Family Educational Rights and Privacy Act (20 *United States Code* [*USC*] § 1232g; 34 *Code of Federal Regulations* Part 99) and the Children’s Online Privacy Protection Act (15 *USC* §§ 6501-6506, P.L. No. 105–277, 112 Stat. 2681–1728).

In TOMS, staff at LEAs and test sites have different levels of access appropriate to the role assigned to them (CDE, 2024e).

#### Statistical Analysis on Secure Servers

During CAASPP testing, ETS information technology staff members retrieve data files from CAI and load those files into a database. The ETS Data Quality Services staff extract the data from the database and perform quality-control procedures (for example, the values of all variables are as expected) before passing files to the ETS statistical analysis group. The statistical analysis staff store the files on secure servers. All staff members involved with the data adhere to the ETS Code of Ethics and the ETS Information Protection Policies to prevent any unauthorized access to data.

#### Student Confidentiality

To meet the requirements of the Every Student Succeeds Act, as well as state requirements, LEAs must collect demographic data about students’ ethnicity, disabilities, parent/guardian education, and so forth during the school year. ETS takes every precaution to prevent any of this information from becoming public or being used for anything other than for testing and score-reporting purposes. These procedures are applied to all documents in which student demographic data appears, such as technical reports.

#### Student Test Results

##### Types of Results

The following deliverables are produced for reporting of the Smarter Balanced Summative Assessments:

* Individual student results for computer-based assessments in the California Educator Reporting System (CERS)
* Individual SSRs (electronic)
* Internet reports—available on the CDE Test Results for California’s Assessments website—aggregated by content area and state, county, LEA, or test site

##### Security of Results Files

ETS takes measures to protect files and reports that show students’ scores and reporting levels. ETS is committed to safeguarding all secure information in its possession from unauthorized access, disclosure, modification, or destruction. ETS has strict information security policies in place to protect the confidentiality of both student and client data. Staff access to production databases is limited to personnel with a business need to access the data. User IDs for production systems must be person-specific or for systems use only.

ETS has implemented network controls for routers, gateways, switches, firewalls, network tier management, and network connectivity. Routers, gateways, and switches represent points of access between networks. However, these do not contain mass storage or represent points of vulnerability, particularly for unauthorized access or denial of service.

ETS has many facilities, policies, and procedures to protect computer files. Software and procedures such as firewalls, intrusion detection, and virus control are in place to provide for physical security, data security, and disaster recovery. ETS is certified in both the ISO 27001 standard for information security and the ISO 22301 standard for business continuity, and conducts disaster recovery exercises annually.

Access to the ETS Computer Processing Center is controlled by employee and visitor identification badges. The Center is secured by doors that can be unlocked only by the badges of personnel who have functional responsibilities within its secure perimeter. Authorized personnel accompany visitors to the ETS Computer Processing Center at all times. Extensive smoke detection and alarm systems, as well as a preaction fire-control system, are installed in the Center.

##### Security of Individual Results

ETS protects individual students’ results during the following conditions:

* Data collection from the TDS
* Scoring
* Transfer of scores by means of secure data exchange
* Reporting
* Posting of aggregate data
* Storage

In addition to protecting the confidentiality of testing materials, The ETS Code of Ethics further prohibits ETS employees from financial misuse, conflicts of interest, and unauthorized appropriation of ETS property and resources. Specific rules are also given to ETS employees and their immediate families who may take an assessment developed by ETS. The ETS OTI verifies that these standards are followed throughout ETS. This verification is conducted, in part, by periodic on-site security audits of departments, with follow-up reports containing recommendations for improvement.

#### Security and Test Administration Incident Reporting System Process

Test security incidents, such as improprieties, irregularities, and breaches, are prohibited behaviors that give a student an unfair advantage or compromise the secure administration of the assessments, which, in turn, compromise the reliability and validity of test results (CDE, 2024b). Whether intentional or unintentional, failure by staff or students to comply with security rules constitutes a test security incident. Test security incidents impact scoring and affect students’ performance on the assessment.

LEA CAASPP coordinators and site CAASPP coordinators ensured that all test security and test administration incidents were documented by following the prompts in TOMS that guided coordinators in their submittal. An Appeal is a request to reset, restore, reopen, invalidate, or grant a grace period extension to a student’s assessment. If an Appeal to a student’s assessment was warranted, TOMS provided additional prompts to file the Appeal.

After a case was submitted, an email containing a case number and next steps was sent to the submitter (and to the LEA CAASPP coordinator, if the case was submitted by the site CAASPP coordinator). The STAIRScase in TOMS provided the LEA CAASPP coordinator, the CDE, and the LEA Outreach Administrator with the opportunity to interact and communicate regarding the STAIRS process (CDE, 2024b).

Prior to the assessment administration, ETS and the CDE agreed that the following types of STAIRS cases would also be forwarded to the CDE:

* Student cheating or accessing unauthorized devices
* Security breach (where a student exposed secure materials)
* Student unable to review previous answers (that is, 20-minute pause rule)

Appeals requests were reviewed by the CDE or an ETS LEA Outreach Administrator. When a request to submit an Appeal was approved, the coordinator received a system-generated email with the Appeal type that was approved (CDE, 2024b).

Types of Appeals available during the 2023–24 Smarter Balanced Summative Assessment administration are described in table 5.4.

Table 5.4 Types of Appeals

|  |  |
| --- | --- |
| **Type of Appeal** | **Description** |
| Reset | Resetting a student’s assessment entailed removing that assessment from the TDS and enabling the student to start a new assessment from the beginning. |
| Invalidate | Invalidated assessments were scored, and scores were provided on the SSR with a note that an irregularity occurred. The student(s) was counted as participating in the calculation of the school’s participation rate for accountability purposes. |
| Reopen | Reopening an assessment allowed a student to access an assessment that had already been submitted or had expired. |
| Restore | Restoring an assessment returned an assessment from the Reset status to its prior status. This action could be performed only on assessments that were reset previously. |
| Grace Period Extension | Permitting a grace period extension allowed the student to review previously answered items upon logging back on to the assessment after expiration of the pause rule. Note that for a performance task, having the test administrator open a new testing session may be all that was needed to continue testing.  A grace period extension was granted only in cases where there was a disruption to a test session, such as a technical difficulty, fire drill, schoolwide power outage, earthquake, or other act beyond the control of the test administrator. |

##### Impropriety

A testing impropriety is an unusual circumstance that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the assessment, test security, or test validity. An example of an impropriety could be if students were making distracting gestures or sounds or talking during the test session that creates a disruption in the test session for other students, or a student left the test room without authorization.

An impropriety can be corrected and contained at a local level. An impropriety should be reported to the LEA CAASPP coordinator and site CAASPP coordinator immediately. The coordinator must report the incident within 24 hours, using the STAIRS/Appeals process in TOMS.

##### Irregularity

A testing irregularity is an unusual circumstance that impacts an individual or a group of students who are testing and may potentially affect student performance on the assessment or impact test security or test validity. An example of an irregularity could be that students were assigned an incorrect designated support or accommodation, or students cheated or provided answers to each other.

These circumstances can be corrected and contained at the local level and submitted using the STAIRS/Appeals process in TOMS. An irregularity must be reported to the LEA CAASPP coordinator and site CAASPP coordinator immediately. The coordinator must report the irregularity within 24 hours, using the online STAIRS/Appeals process in TOMS.

##### Breach

A testing breach is an event that poses a threat to the validity of the assessment. Examples may include such situations as a release of secure materials or a security or system risk. These circumstances have external implications for the CDE and may result in a decision to remove the test item(s) from the available secure item bank.

Breaches require immediate attention; a breach that was due to social media exposure on the part of a student or adult or due to media coverage of an administration was to be escalated to CalTAC via a telephone call from the LEA CAASPP coordinator. Following the call, the site CAASPP coordinator or LEA CAASPP coordinator must report the incident using the online STAIRS/Appeals process in TOMS within 24 hours. All other breaches were to be entered into STAIRS directly.

#### Appeals

For test security incidents reported in STAIRS that resulted in a need to invalidate, restore, or provide a grace period extension for individual computer-based student assessments, the request had to be approved by the CDE. Requests to reset and reopen assessments were processed by an LEA Outreach Administrator.

In most instances, an Appeal was submitted to address a test security breach or irregularity. The LEA CAASPP coordinator or site CAASPP coordinator submitted Appeals in TOMS. All submitted Appeals were available for retrieval and reviewed by LEA and site coordinators within a given organization. An Appeal could be requested only by the LEA CAASPP coordinator or site CAASPP coordinator if prompted while filing a STAIRS case in TOMS (CDE, 2024d). Types of Appeals available during the 2023–24 administration are described in table 5.4.

Table 5.5 and table 5.6 present the number of Appeals in STAIRS in the 2023–24 administration for ELA and mathematics, respectively, as well as the number of individual Statewide Student Identifiers (SSIDs) submitted and approved. The number in the *Appeals SSIDs Approved* column is the number of accepted cases that resulted in an Appeal, which may differ from the number in the *Number of Incidents* column because of incorrect entry or other factors.

Table 5.5 Number and Types of Incidents Submitted in STAIRS—ELA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Appeal Type** | **Number of Incidents** | **Total Number of SSIDs Submitted** | **Appeals SSIDs Approved** |
| Accessibility Issue | Reset | 1,070 | 1,753 | 1,643 |
| Administered Incorrect Assessment | Reset or Reopen or No Appeal | 441 | 633 | 374 |
| Administration Error | No Appeal | 0 | 0 | 0 |
| Data Entry Issue | Reset or Reopen or Invalidate or No Appeal | 0 | 0 | 0 |
| Expired or Accidentally Submitted Test | Reopen | 5,536 | 10,512 | 10,428 |
| Exposing Secure Materials | Invalidate or No Appeal | 0 | 0 | 0 |
| Incorrect SSID Used | Reset or No Appeal | 148 | 181 | 129 |
| Restore from Reset | Restore | 11 | 13 | 8 |
| Student Cheating or Accessing Unauthorized Devices | Invalidate | 278 | 301 | 276 |
| Student Disruption | No Appeal | 64 | 0 | 0 |
| Technical Issues | Grace Period Extension or No Appeal | 162 | 1,710 | 1588 |
| Validity Issue | Invalidate or Reset | 83 | 235 | 206 |

Table 5.6 Number and Types of Incidents Submitted in STAIRS—Mathematics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Appeal Type** | **Number of Incidents** | **Total Number of SSIDs Submitted** | **Appeals SSIDs Approved** |
| Accessibility Issue | Reset | 1,064 | 2,022 | 1,876 |
| Administered Incorrect Assessment | Reset or Reopen or No Appeal | 179 | 313 | 220 |
| Administration Error | No Appeal | 0 | 0 | 0 |
| Data Entry Issue | Reset or Reopen or Invalidate or No Appeal | 0 | 0 | 0 |
| Expired or Accidentally Submitted Test | Reopen | 1,363 | 2,140 | 2,120 |
| Exposing Secure Materials | Invalidate or No Appeal | 0 | 0 | 0 |
| Incorrect SSID Used | Reset or No Appeal | 116 | 133 | 92 |
| Restore from Reset | Restore | 12 | 12 | 9 |
| Student Cheating or Accessing Unauthorized Devices | Invalidate | 332 | 364 | 346 |
| Student Disruption | No Appeal | 46 | 0 | 0 |
| Technical Issues | Grace Period Extension or No Appeal | 96 | 704 | 660 |
| Validity Issue | Invalidate or Reset | 44 | 132 | 123 |

Table 5.7 and table 5.8 present the number of Appeals approved and rejected, respectively, in ELA and mathematics by Appeal type in STAIRS.

Table 5.7 Number of Appeals Approved in STAIRS

|  |  |  |
| --- | --- | --- |
| **Appeal Type** | **Number of Appeals Approved in ELA** | **Number of Appeals Approved in Mathematics** |
| Reset | 2,090 | 2,086 |
| Invalidate | 549 | 500 |
| Reopen | 10,437 | 2,188 |
| Restore | 8 | 9 |
| Grace Period Extension | 1,588 | 660 |

Table 5.8 Number of Appeals Rejected in STAIRS

|  |  |  |
| --- | --- | --- |
| **Appeal Type** | **Number of Appeals Rejected in ELA** | **Number of Appeals Rejected in Mathematics** |
| Reset | 279 | 168 |
| Invalidate | 48 | 55 |
| Reopen | 85 | 20 |
| Restore | 5 | 3 |
| Grace Period Extension | 122 | 39 |

### Processing and Scoring

Each day, new constructed-response (CR) data and TDS-scored data flow from the TDS to ETS. The TDS is capable of scoring a variety of item types, referred to as “machine-scored” items, which are described in subsection [*7.1.2 Scoring Item Responses*](#_Scoring_Item_Responses). Outcomes of CR items are scored by artificial intelligence or by human scoring.

Targeted efforts are made to recruit California educators for participation as raters in the human-scoring portion of the Smarter Balanced assessments. Raters are certified on the basis of their ability to use a rubric and accurately score sample responses. Once approved, raters are trained to access the Measurement Incorporated and ETS scoring interfaces and Smarter Balanced–specific scoring policies and procedures and are provided interactive training to practice scoring sample responses with feedback from the scoring leader.

Raters work in shifts and are supervised by a scoring leader who has received special training in scoring and monitoring. Raters are provided Smarter Balanced materials to aid scoring; these materials include anchor sets, scoring rubrics, validity samples, qualifying sets, and condition codes. (Refer to section [*7.2 Scoring Constructed-Response Items and Extended Writing Tasks*](#_Scoring_Constructed-Response_Items) for the definitions of these materials.) A scoring leader gives direct feedback to raters for additional content support. Scoring of California student responses is given priority routing to raters who are California-based educators.

### References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

*California* *Code of Regulations*,Title 5, Education, Division 1, Chapter 2, Subchapter 3.75, Article 2.

California Department of Education. (2023). *California assessment accessibility resources matrix*. Sacramento, CA: California Department of Education.

California Department of Education. (2024a). *CAASPP and ELPAC accessibility guide*. Sacramento, CA: California Department of Education.

California Department of Education. (2024b). *CAASPP and ELPAC security incidents and Appeals procedure guide*.Sacramento, CA: California Department of Education.

California Department of Education. (2024c). *CAASPP and ELPAC technical specifications and configuration guide for online testing*. Sacramento, CA: California Department of Education.

California Department of Education. (2024d). *CAASPP online test administration manual*. Sacramento, CA: California Department of Education.

California Department of Education. (2024e). *Test Operations Management System user guide*. Sacramento, CA: California Department of Education.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.

Every Student Succeeds Act of 2015, 20 U.S.C. 20 § 6301 (2015).

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide*.Los Angeles: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced resources and practices comparison crosswalk*. Los Angeles: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2023). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Los Angeles: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2024). *Accessibility strategies* (web page). Los Angeles: Smarter Balanced Assessment Consortium.

### Appendix 5.A: Accessibility Resource Assignment

*This content is located in a separate file.*

## Standard Setting

### Description

Standard setting, which also is referred to as achievement level setting, refers to a class of methodologies by which one or more thresholds are used to determine achievement levels. The California Department of Education set four achievement levels—*Standard Not Met, Standard Nearly Met*, *Standard Met,* and *Standard Exceeded*—with three threshold cuts for each grade level and content area.

In coordination with its member states, the Smarter Balanced Assessment Consortium implemented an extensive achievement level–setting process involving software development, item mapping, review panels, committees, workshops, and extensive validity research to set the final thresholds and achievement level descriptors. For detailed information regarding this process, refer to the *Smarter Balanced Assessment Consortium 2013–14 Technical Report* (Smarter Balanced, 2016).

### Reference

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2013–14 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

## Scoring and Reporting

To determine individual students’ scores for the Smarter Balanced Summative Assessments, student item responses were scored, and individual student scores were calculated on the basis of the item responses. In addition, student test scores were aggregated to produce information for schools and local educational agencies (LEAs).

This chapter describes how various types of student responses were scored, as well as the various types of scores and score reports that were produced at the end of administration of the Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics.

### Scoring Process

#### Structure of the Assessments

To understand the basis of the scoring approach, an understanding of the structure of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced computer-based summative assessments is necessary. These assessments are designed to gather evidence that can be used to make inferences about student mastery of the Common Core State Standards (CCSS). The assessments are based on claims and targets.

Claims are inferences made about a student based on the student’s test score. They are broad statements about learning outcomes. These statements require evidence that articulates the types of data and observations that support interpretations of progress toward the achievement of the claim. Claims identify the set of knowledge and skills being measured. The following is an example of a mathematics claim:

**Claim 1: Concepts and Procedures—**Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.

Targets describe the evidence that can be used to support a claim about a student. Targets are specific to claims. The following is a target associated with the previous claim:

**Target C—**Understand the connections between proportional relationships, lines, and linear equations.

The items are designed on the basis of a variety of task models that define item characteristics such as item type, allowable stimuli, prompt feature, and item interactions.

For the 2023–24 CAASPP Smarter Balanced administration, claim results at the individual level were not reported because of the impact that the adjusted blueprint has on the reliability of individual claim scores. (In general, fewer items means lower reliability.) Individual claim scores were also not reported to LEAs; however, they were reported to the California Department of Education (CDE).

Composite claims were reported for individual students taking the adjusted blueprint. The composite claim scores are calculated using maximum likelihood estimation, as described in the Smarter Balanced scoring specifications (Smarter Balanced, 2023); the scores are based on the items contained in a particular claim or combination of claims. In ELA, composite claim scores are computed for Claim 1 and Claim 3 (without speaking items) combined and Claim 2 and Claim 4 combined. In mathematics, composite claim scores are computed for Claim 1 as well as for Claim 2, Claim 3, and Claim 4 combined.

#### Scoring Item Responses

In accordance with the Smarter Balanced Online Summative Assessment specifications, students were administered a computer adaptive test (CAT) component and a randomly selected performance task (PT) (Smarter Balanced, 2017a through 2017i [ELA], 2018a through 2018k [mathematics]).

The combination of the CAT and PT components fulfilled the content requirements for the test blueprint (refer to [appendix 4.A](#_Appendix_4.A:_Smarter), [appendix 4.B](#_Appendix_4.B:_Performance), and [appendix 4.C](#_Appendix_4.C:_Item)).

CAASPP Smarter Balanced computer-based summative assessments included traditional selected response items, short constructed-response (CR) items, writing extended-response (WER) items, and technology-enhanced items. Some items were machine-scored, which means that they could be scored by the Test Delivery System (TDS). Other items were human-scored by a trained rater or via a hybrid approach of human and automated scoring. The scoring approach used depended on the item type and scoring requirements provided by the Smarter Balanced item specifications.

##### Constructed-Response

Item types that required a student to provide a response by entering words or numbers are called “constructed-response” items. Both the CAT and PT included CR items. The CAT section contained both machine-scored items worth 0–1 or 0–2 points, as well as short-text items worth 0–2 points. The PT section contained machine-scored items, short-text items worth 0–2 points, and WER items worth 0–6 points.[[7]](#footnote-8) A small number of mathematics PTs included CR items with a 0–3 point range. CR items for CAASPP Smarter Balanced assessments included the following item types:

* ***Short-answer text response items*** required students to respond with words, phrases, short sentences, mathematical expressions, or numbers. These items have a value of 0–2 points, with a small number of mathematics short-answer items having values ranging from 0–3 points. These items were scored holistically on the basis of a rubric. Holistic scoring gave students a single, overall assessment score for the response as a whole.
* ***WER items (full-write response)*** required students to write one or more paragraphs. The WER is scored for three writing traits (organization and purpose, evidence and elaboration, and conventions). These items were scored analytically on the basis of rubrics; readers assigned a score based on each dimension. WER items appear on ELA assessments only.

Short-answer items and WER items for ELA, and short-answer items for mathematics, for Smarter Balanced Summative Assessments were scored either by human raters (that is, hand scoring) or by an automated scoring engine (that is, automated scoring) called Project Essay Grade (PEG). An ETS subcontractor, Measurement Incorporated (MI), was responsible for scoring CR items, where some were hand scored, and some were scored via a hybrid approach using both automated scoring and hand scoring. Note that with hybrid scoring, the decision to use hand or automated scoring was at the response level, not the item level. These two scoring processes are described in subsections [*7.2.1 Rater Selection*](#_Rater_Selection) through [*7.2.5 Rater Monitoring*](#_Rater_Monitoring), which describe the hand scoring process; and subsection [*7.2.6 Automated Scoring*](#_Automated_Scoring), which describes automated scoring, including the hybrid process.

##### Selected Response (Machine-Scored)

Selected responses are scored by the TDS. Table 7.1 lists the types of items that were machine-scored.

Table 7.1 Machine-Scored Computer-based Item Types

|  |  |  |
| --- | --- | --- |
| **Item Type** | **Description** | **Content Area** |
| Equation | The student enters an equation or numeric response using an on-screen panel containing mathematical characters. | Mathematics only |
| Evidence-based selected response | This is a two-part item. The student responds to a multiple-choice item and then responds to a multiple-select item. | ELA only |
| Grid item—Drag and drop | The student responds by dragging and dropping a single choice (“source”) into the appropriate location (“target”). The scoring key is a set of numeric identifiers that specifies which source needs to be placed in which target to answer the item correctly. | Mathematics only |
| Grid item—Graphing | The student plots points, lines, and multisegmented lines on a graph. Items can be answered by looking at a graph. For some items, students must manipulate the elements in the graph to respond. | Mathematics only |
| Hot text | The student is presented with a stem that contains multiple underlined words or phrases from which the student selects sections of text or drag-and-drop sections of text. | ELA only |
| Match interaction | Text or images in rows are provided, and students are instructed to match these to values in columns by selecting a box where a match is valid. | ELA and mathematics |
| Multiple choice | Three to five answer choices are provided, and the student can select only one choice to respond. | ELA and mathematics |
| Multiple select | Five to eight answer choices are provided, and the student is instructed to select one or more choices to respond. These item types can have multiple keys; the student may be awarded partial credit for partially correct answers or may need to select all correct answers to receive credit. | ELA and mathematics |
| Table interaction | The student is required to respond by making a keyboard entry into one or more cells in a table grid. The response can be restricted to one selection of row, column, or table, or no restrictions. | Mathematics only |

##### Hybrid of Automated Scoring and Hand Scoring

The CDE elected to use a hybrid of automated scoring and hand scoring for the 2023–24 test administration. This hybrid scoring solution includes the following:

* Determining eligible items for automated scoring
* Training the models using data sources from multiple Smarter Balanced states and reliable expert raters (based on empirical evidence)
* Developing several models per item type, where needed. (For example, the WER items may have several models developed to handle each trait and, in some instances, condition codes as well.)
* Validating the models according to a two-step validation process:
* Initial validation is done using the training data; initial validation includes human-to-human and human-to-machine validation.
* Secondary validation is done annually, using operational data early in the test administration window to ensure the model generalizes as intended.
* The two-step validation process ensures confidence in the use of the models for operational scoring; if there are issues detected during either of the validation stages, the model can be retrained, or the item can be hand scored.

Once the validation steps are completed and both have been passed, the student responses are then routed for either automated scoring or human scoring, depending on the confidence of the scoring model for scoring each response. PEG can predict the human score and predict the accuracy of the autoscore, which allows more challenging (that is, low-confidence) responses to be routed to an expert rater (approximately 15%) to assign the score of record.

Because the secondary validation is done on an annual basis, this approach mitigates a concern regarding how automated scoring models generalize over time.

#### Certification of the Scoring System

ETS staff from the Assessment Development, Enterprise Score Key Management (eSKM), Psychometric Analysis & Research (PAR), Constructed Response Scoring, Systems & Capabilities, and Information Technology divisions participated in the certification of the scoring system. Each team followed procedures required by the ETS Office of Quality for operational readiness and Standard 7.8 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

ETS staff reviewed operational answer keys and scoring rubrics provided by the Smarter Balanced Assessment Consortium. In addition, item parameter estimates for items were loaded into the ETS operational scoring system. Central aspects of the validity of the CAASPP computer-based summative test scores are the degree to which scoring rubrics are related to the appropriate assessment targets and claims based on Smarter Balanced assessments. A key facet of validity is the degree to which scoring rules are applied accurately throughout the scoring sessions.

### Scoring Constructed-Response Items and Extended Writing Tasks

The procedures for hand scoring items are specified by Smarter Balanced. The scoring process described next is used to score responses to all CR short-answer and essay items. Subsections [7.2.1](#_Rater_Selection) through [7.2.5](#_Rater_Monitoring) are associated with hand scoring; while [7.2.6](#_Automated_Scoring) describes automated scoring, including the hybrid process.

#### Rater Selection

ETS and MI have developed a pool of more than five thousand raters who are experienced in scoring the Smarter Balanced assessments. The recruiting team first recruited qualified raters who had experience scoring these assessments. Rater accuracy data, collected during prior administration scoring, was used to prioritize recruitment of the most accurate, experienced raters. Once recruited, experienced raters were assigned to the content area and grade band(s) with which they were most experienced.

To supplement this core pool, the recruiting team contacted other raters in the database who have experience successfully scoring other large-scale assessments. These raters were assigned to the grade level, content area, and item type for which they were most qualified based on their performance on similar projects. Returning staff were selected on the basis of experience and performance, as well as attendance, punctuality, and cooperation with work procedures and organizational policies.

ETS and MI maintain evaluations and performance data for all staff who work on each scoring project to determine employment eligibility for future projects. Finally, ETS and MI target recruitment of new raters as needed, which allows for continual identification of talent across the country who will best fulfill the hand scoring requirements and represent a diverse population across age, ethnicity, gender, and other demographic groups.

Raters should have met the following requirements prior to being hired:

* Candidates must have a bachelor’s degree and be eligible to work in the United States (and be e-verified prior to hire).
* Teaching experience is strongly preferred.
* Graduate students and substitute teachers are encouraged to apply.
* Bilingual English and Spanish speakers are encouraged to apply.
* Candidates must complete training and achieve qualifications through the certification process.

California educators with teaching credentials, retired educators, and school administrators were encouraged to apply.

In selecting team leaders to monitor the raters, the scoring leadership reviewed records of all returning staff. They sought people who were experienced team leaders with a record of good performance on previous projects. They also considered raters who had been recommended for promotion to the team leader position or otherwise displayed exemplary performance.

All team leaders and raters were required to qualify before scoring and were informed of what they were expected to achieve to qualify. Refer to [*7.2.2 Training* *Overview*](#_Training_Overview) for a more complete description of this training.

All hand scoring project staff (scoring directors, team leaders, raters, and clerical staff) were required to sign a confidentiality and nondisclosure agreement before receiving any training or viewing any secure project materials. The employment agreement indicates that no participant in training or scoring may reveal information about the assessment, the scoring criteria, or the scoring methods to any person.

#### Training Overview

All raters hired for Smarter Balanced assessment hand scoring were trained using the rubric(s), anchor sets, and training and qualifying sets provided by Smarter Balanced. Many of these sets were created during the original field test scoring in 2014 and approved by Smarter Balanced. Additional sets were created as new items were field-tested. The same anchor sets are used each year. Additionally, an annual review of the rater agreement and scoring materials is conducted to inform the development of item-specific, supplemental training materials.

Supplemental materials are developed each summer and implemented in the subsequent operational administration. These additional materials are developed with a focus on challenging areas identified during the previous operational administration, as indicated by suboptimal rater accuracy (based on validity responses) or rater agreement. Supplemental materials are also created for newly operational items for which MI identifies a need for additional examples. For instance, MI may find an approach to a mathematics item that was not encountered during field testing but appears frequently during operational scoring, or an uncommon but valid way to address a research prompt that is not reflected in the existing rubric. In these cases, MI provides examples of these specific approaches along with guidance on how to score them correctly. MI also supplements materials to provide raters with additional guidance for content-wide challenging spots—such as full-write conventions—or to help them more accurately identify responses that should be flagged as nonscorable.

Once hired, raters are placed into a scoring group that corresponds to the content area and grade level that they are deemed best suited to score (based on work history and performance on past scoring projects). Raters are trained to score a specific item group of either short-answer (research, brief write, reading, and mathematics) or WER items. Within each item group, raters are divided into teams and supervised by team leaders and a scoring director. Each scoring director, team leader, and rater is assigned a unique ID used to track their scoring work throughout the scoring effort. The number of items an individual rater scores is minimized so that the rater becomes highly experienced in scoring responses to a given set of items.

All raters, experienced or not, are required to train on all the anchor and training sets. Following training and practice, all raters were required to pass a qualification to prove that they understand and can apply the criteria accurately. The scoring director and team leaders had access to all practice and qualification results, which were reviewed to identify frequently misscored responses and inform initial monitoring and feedback needs.

Until a rater has trained and qualified successfully, the rater is not permitted to score any student responses. Training is structured carefully so raters understand that all scoring decisions must be grounded in the training materials. In addition, raters learn how to navigate the anchor set, develop the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retain the necessary consistency to score all responses accurately.

#### Training Processes

During the 2023–24 scoring, all scoring personnel logged on to the MI secure Scoring Resource Center (SRC). The SRC included all online training modules, served as the portal to the MI Virtual Scoring Center (VSC) interface, and maintained the data repository of all scoring reports used for rater monitoring. The MI training system (VSC Train) provided a remote, secure application for training both team leaders and raters. VSC Train delivered a training lesson for each item that allowed each trainee to complete the following steps:

1. Review the anchor set(s)
2. Score the practice set(s)
3. Review an annotated version of the practice set(s) after submitting scores
4. Score the qualification sets

Training design varied slightly depending on Smarter Balanced item type:

* **WER Items (Full Writes):** Raters trained and qualified on a baseline training lesson for a grade level and writing purpose (for example, grade three narrative, grade six argumentative, etc.). After qualifying on the baseline, raters then took qualifying sets for each item in that grade level and purpose. Raters scored only those items for which they passed the qualifying set.
* **ELA Brief Writes, Reading, and Research:** Raters trained and qualified on a baseline lesson within a specific grade band and target. Qualification on the baseline lesson qualified the rater to score all items in that grade band and target.
* **Mathematics:** Raters trained and qualified on baseline lessons within a specific grade band. Qualification on a baseline lesson qualified the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

An additional validation stage supplemented full-write, brief-write, reading, and research rater qualification. Following the training and qualification steps described previously, all prospective full-write, brief-write, reading, and research raters were required to score, for most items, a 20-response set of prescored student responses sourced from the prior test administration. Like the qualification step, raters were required to meet accuracy standards during this validation to score operational responses for a given item. Any raters who failed to meet validation accuracy standards were automatically disqualified from scoring the item despite having passed qualification. This additional validation matches the full-write qualification methods that have been in place since the start of Smarter Balanced scoring in 2015 and adds an additional level of quality assurance (QA).

Rater training time varies by grade level and content area. Training for brief-write, reading, research, and mathematics items can typically be accomplished in one day, while training for WER items may take up to five days to complete. Raters generally worked three to seven hours per day. The hours worked per day were flexible, based on the raters’ shift preference and item(s) being scored. At a minimum, most raters scored 15 hours per week (day shift) or 10 hours per week (evening shift), with many scoring more than 30 hours per week (day shift) or 20 hours per week (evening shift).

In addition to item-specific scoring specifications, a variety of substantive procedural and policy information was provided to each trainee during training. These included instructions for how to identify and flag particular types of responses as well as how to communicate with leadership during hand scoring.

Raters were trained to recognize nonscorable responses; these responses were systematically routed to scoring supervisors for final condition-code assignment per Smarter Balanced requirements. For some item types, such as essays, condition-code responses were scored by scoring leaders trained to specialize in the scoring of these types of responses.

An “alerts” procedure was explained to raters during training sessions, where raters were trained to recognize “alerts” in their various forms, including those for suicide, criminal activity, alcohol or drug use, extreme depression, violence, rape, sexual or physical abuse, self-harm, intent to harm others, and neglect.

The training process, including this additional information, ensured that raters were fully prepared to hand score responses and understood all responsibilities and scoring requirements before they began operational scoring.

Following training, all materials were made available during scoring via the VSC Score Resource Library to ensure that all scoring staff could refer to them throughout the scoring effort. This library included the item and rubric, the annotated anchor and practice sets, and any associated supplemental materials.

When scoring, raters had access only to those items for which they had successfully trained and qualified. The hand scoring system sorts individual student responses into small sets of 5 to 10, grouped by item. When a rater is qualified to score multiple items, this approach eases cognitive load by presenting the rater with a scoring set in which all responses relate to the same item.

Multiple strategies were employed to minimize rater bias. First, raters did not have access to any student identifiers. Unless the students signed their names, wrote about their hometowns, or in some way provided other identifying information as part of their response, the raters had no knowledge of student characteristics. Second, all raters were trained using Smarter Balanced–provided materials, which were approved as unbiased examples of responses at the various score points. Training involved constant comparisons with the rubric and anchor papers so that raters’ judgments were based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback was maintained to identify any issues. Specifically, raters were closely monitored during scoring; and any instance of a rater making scoring decisions based on anything except the criteria was discussed with the rater. After this feedback was provided, raters were further monitored. Any rater who continued to exhibit bias after receiving a reasonable amount of feedback was dismissed.

MI also implemented a series of automated score verifications to ensure the accuracy of scores. For example, MI conducted a blank check that reset scores when a condition code of “blank” was assigned to a response with one or more characters in the response string (for example, a response composed of spaces or tabs). In this case, the score was recorded only after three independent raters assigned a condition code of “blank” to a response that appeared blank but included characters in the response string. A similar check was run when a score or condition code other than “blank” was assigned to a response that included no characters in the response string.

Automatic resetting of double-scored responses when two raters assigned nonadjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provided an additional score verification. In addition to automatically resetting and rescoring these responses, the raters’ information was captured in a report and reviewed by scoring directors, one of many tools used to determine retraining needs.

##### English Language Arts/Literacy Performance Task Extended Writing Tasks

Baseline anchor sets for each writing purpose (for example, informational writing or explanatory writing) were used to train raters on each of the writing traits―organization and purpose, evidence and elaboration or development and elaboration, and conventions—within a particular grade band. The writing purposes are narrative, informational, and opinion at grades three through five; narrative, informational, and argumentative at grades six through eight; and explanatory and argumentative at grade eleven.

For all writing purposes, organization and purpose is the first trait and conventions is the third trait. Evidence and elaboration is the second trait for the opinion, argumentative, informational, and explanatory writing purposes. Development and elaboration is the second trait for the narrative writing purpose.

What follows are writing traits for opinion, argumentative, informational, or explanatory writing:

* Organization and Purpose
* Opinion (grades three through five)
* Argumentative (grades six through eight and grade eleven)
* Informational (grades three through eight)
* Explanatory (grade eleven)
* Narrative (grades three through eight)
* Evidence and Elaboration
* Opinion (grades three through five)
* Argumentative (grades six through eight and grade eleven)
* Informational (grades three through eight)
* Explanatory (grade eleven)
* Conventions
* Opinion (grades three through five)
* Argumentative (grades six through eight and grade eleven)
* Informational (grades three through eight)
* Explanatory (grade eleven)
* Narrative (grades three through eight)

What follows are writing traits for narrative writing:

* Organization and Purpose
* Opinion (grades three through five)
* Argumentative (grades six through eight and grade eleven)
* Informational (grades three through eight)
* Explanatory (grade eleven)
* Narrative (grades three through eight)
* Development and Elaboration
* Narrative (grades three through eight)
* Conventions
* Opinion (grades three through five)
* Argumentative (grades six through eight and grade eleven)
* Informational (grades three through eight)
* Explanatory (grade eleven)
* Narrative (grades three through eight)

A chart that presents the traits to their purposes is shown in figure 7.1.

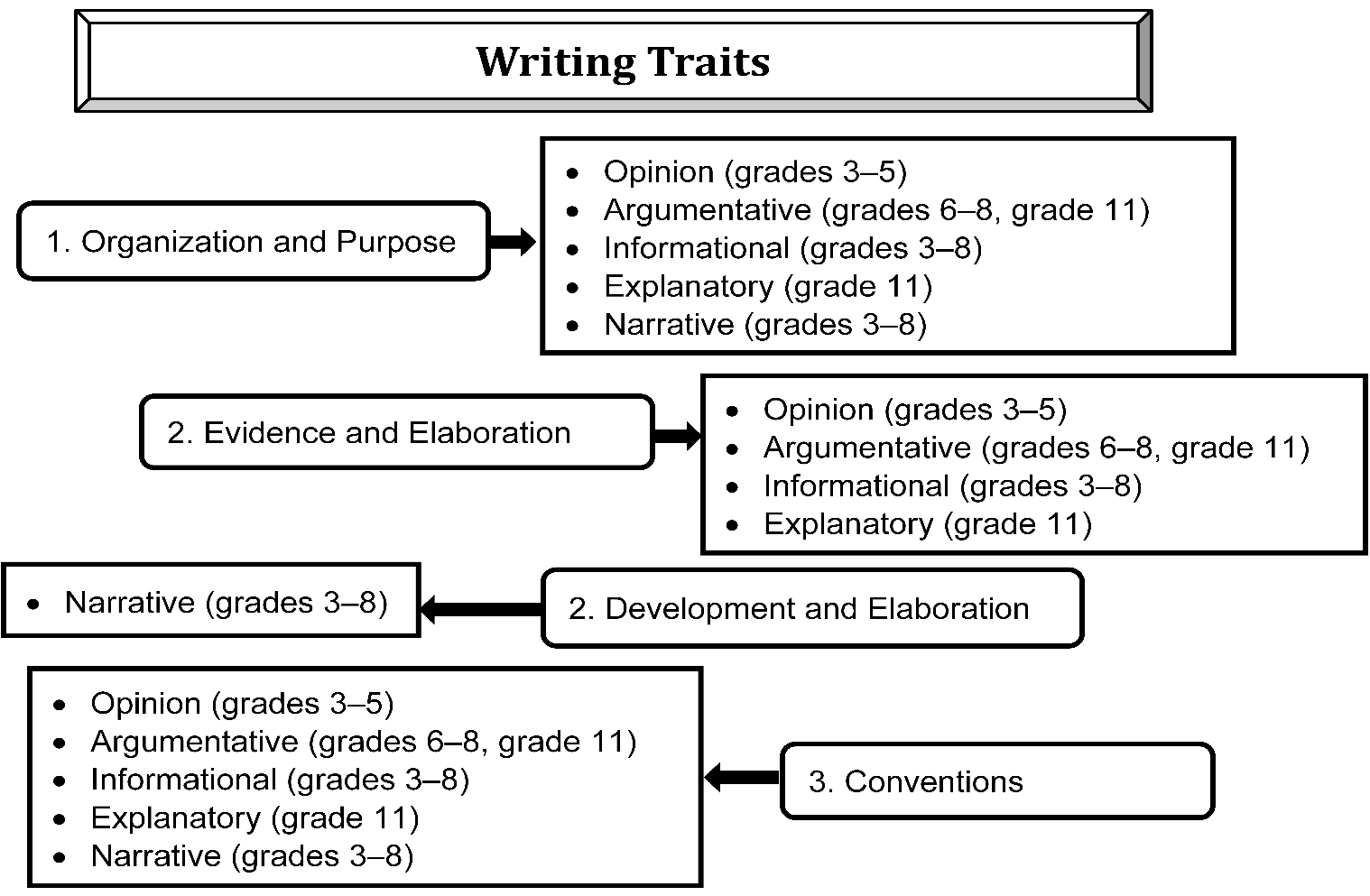


Figure 7.1 Writing traits

The training process for extended writing tasks is described next. The training steps are described in the following list.

**Training steps include the following:**

1. Trainees read the task, rubrics, and source materials for the WER items in a particular grade band and writing purpose (for example, grades three through five informational). Trainees read sample responses and annotations.
2. Trainees read a training set of five responses to the same item (Essay 1) and scored those responses for organization and purpose.
3. Trainees reviewed the correct scores and the scoring rationale for the organization and purpose scores for those responses.
4. Trainees read another training set of five responses to that item (Essay 1) and scored those responses for evidence and elaboration. They then reviewed the correct scores and the scoring rationale for the evidence and elaboration scores for those responses.
5. Trainees read another training set of five responses to that item (Essay 1) and scored those responses for conventions. They then reviewed the correct scores and the scoring rationale for the conventions scores for those responses.
6. Trainees read another training set of five responses to that item (Essay 1) and scored each of those responses for all three traits.
7. Trainees reviewed the scoring rationale for the training responses and answered training questions.
8. Trainees scored a qualification round (10 papers) for all three traits for Essay 1.
9. Trainees scored a second qualification round (10 papers) for all three traits for Essay 1.
10. Once qualified on the baseline item, trainees moved into item qualification for specific items. Trainees were presented with stimulus materials and the prompt. After review, trainees scored a qualification round (10 papers) for all three traits. Qualified raters—those who met the standard in the qualification round—began scoring. Trainees who did not meet the qualification standard on their first attempt had an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.

The training materials are described in the following list.

**Materials for training raters of WER items, at each grade level include the following:**

1. Baseline anchor sets approved during Smarter Balanced pre–range finding (Range finding activities included the review of student responses against item rubrics, the validation of rubric effectiveness, and the selection of anchor papers used by human scoring for the larger population of responses.)
2. Field test prompt and stimulus materials
3. Purpose- and task-specific rubrics
4. Conventions charts approved by the Smarter Balanced Assessment Consortium
5. Supplemental scoring guidelines approved by the Smarter Balanced Assessment Consortium
6. Training sets specific to the first WER task for each grade level and purpose
7. Qualification sets generally administered in two rounds of approximately 10 responses per WER task

##### English Language Arts/Literacy Short-Answer Items

The process for training raters to score short-answer items was also organized by grade band (grades three through five, grades six through eight, or grade eleven). These training steps are described in the following list.

**Training steps include the following:**

1. Trainees read the rubrics and scoring notes for the short-answer items in a particular grade band and purpose category (for example, grades three through five evidence). Trainees read sample responses to a prompt and the associated annotations.
2. Trainees reviewed the scoring rationale for each of the anchors (that is, anchor sets for the claim, target, and subclaim).
3. Trainees scored the training set (5–10 responses) for the short-answer claim, target, and subclaim.
4. Trainees reviewed the correct scores and scoring rationale for the training set.
5. Trainees read the prompt, source materials, or stimuli for the first short-answer item in the claim, target, and subclaim (for example, Grade 6, Claim 1, Reading Item 1).
6. Trainees scored a qualification round.
7. Trainees who did not meet the qualification standard on their first attempt had an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.
8. Qualified raters began scoring.

The training materials are described in the following list.

**Materials for short-answer item training include the following:**

1. Anchors and training sets by grade band, claim, target, and subcategory
2. Prompts and source materials or stimuli
3. Item-specific rubrics
4. Two qualification sets with 10 responses per item

##### Mathematics Items

The training process for mathematics items is described next. The training steps for scoring mathematics items are described in the following list.

**Training steps include the following:**

1. Trainees reviewed the items that are represented in the anchor and training sets, any associated source materials or stimuli, and the item-specific rubrics.
2. Trainees read the associated source materials or stimuli, as appropriate.
3. Trainees scored the training set for the item category.
4. Trainees reviewed the correct scores and scoring rationale for the training set.
5. Trainees scored a qualification round.
6. Trainees scored a second qualification round.
7. Trainees who did not meet the qualification standard on their first attempt had an opportunity to review correct scores and the scoring rationale with a scoring leader before making a second attempt.
8. Qualified raters began scoring.

The training materials are described in the following list.

**Materials for mathematics training include the following:**

1. Anchors and training sets by PT grade level, family, and item category or by CAT item
2. Prompts and source materials or stimuli
3. Item-specific rubrics
4. One or two qualification rounds per item category, depending on item complexity, with 10 responses per round

Unlike ELA PTs, mathematics PTs could contain interdependencies among the items within a task. Each mathematics PT was made up of four to six items. Items could be dependent on any of the previous items within the PT. For example, if item 6 was dependent on items 3 and 5, the rubric for item 6 specified the correct response based on prior correct responses to items 3 and 5. Raters were responsible for determining the appropriate response to item 6 and awarding credit accordingly, even when the student’s responses to items 3 and 5 were incorrect. It also was possible for the first two of the six items to be machine-scored while two or more of the other four were human-scored.

The proper handling of tasks with dependencies was addressed in the training process. Raters had practice working through PT responses and recognizing correct work that was based on previous incorrect values. PTs are composed of items based on several different task models. In general, training materials were organized so raters trained on a task model rather than on a complete PT. However, when PT items that were dependent on previous items in the set were presented in training, the entire set of items and responses was included. This allowed raters to identify the previous responses that served as the basis for the item that is being scored.

#### Score Summary for Condition Codes

Scoring condition codes allowed raters to categorize certain responses as unscorable. The code indicated the reason that the response could not be scored. Responses with condition codes were routed to scoring supervisors for final code assignment. Supervisors required detailed training on the Smarter Balanced condition codes and definitions (Smarter Balanced, 2014a).

Table 7.2 presents the valid condition codes used for scoring, along with descriptions of the responses that would warrant the assignment of the different codes.

Table 7.2 Scoring Condition Codes

|  |  |  |
| --- | --- | --- |
| **Condition Code** | **Reason** | **Use** |
| B | Blank | No response |
| I | Insufficient | 1. Use the “I” code when a student has not provided a meaningful response; this may include the following:  * Random keystrokes * Undecipherable text * “I hate this test” * “I don’t know, IDK” * “I don’t care” * “I like pizza!” (in response to a reading passage about helicopters) * Response consisting entirely of profanity  1. For ELA WER items, use the “I” code for responses described previously and also if  * the student’s original work is insufficient for the rater to determine whether the student is able to organize, cite evidence and elaborate, and use conventions as defined in the rubrics; or * the response is too brief to make a determination regarding whether it is on purpose or on topic. |
| L | Nonscorable Language | * ELA: Language other than English * Mathematics: Language other than English or Spanish |
| T | Off-Topic for ELA WER Items Only | * The response is unrelated to the task or sources or shows no evidence that the student has read the task or the sources (especially for informational or explanatory and opinion or argumentative). “Off topic” responses are generally substantial responses. |
| M | Off-Purpose for ELA WER Items Only | The student has clearly not written to the purpose designated in the task:   * An off-purpose response addresses the topic of the task but not the purpose of the task. * Students may use narrative techniques in an explanatory essay or use argumentative or persuasive techniques to explain, for example, and still be on purpose. * Off-purpose responses are generally developed responses (essays, poems, etc.) clearly not written to the designated purpose. |

#### Rater Monitoring

During operational scoring, 5 percent of the responses scored comprised preapproved validity responses. Validity responses serve as benchmark responses, as the most appropriate score for each validity response is predetermined by key interest holders. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The validity pool includes anchor validity responses originating from the field test administration.[[8]](#footnote-9) The pool of validity responses was selected to be generally representative of operational responses while ensuring sufficient examples of each score point. Validity results compared the score assigned by a rater to a validity response with the benchmark score of the same response. Validity responses provide a more direct measurement of rating quality than measures of interrater reliability (Raczynski et al., 2015).

MI calibrated validity responses to fit a unidimensional item response theory (IRT) model for each content area and item type. This approach involved transforming raters’ validity response scores into accuracy scores. Specifically, if the rater’s score matched the “true” score of the validity response, an accuracy score of 2 was assigned. If the rater’s score was adjacent to the score of the validity response, an accuracy score of 1 was assigned. Otherwise, for scores that were nonadjacent, an accuracy score of 0 was assigned. All accuracy score data for validity responses and readers were then fitted to a generalized partial credit model (GPCM) IRT model. Using the resulting IRT parameters, MI then calculated accuracy values for each rater based on a given set of validity responses. This calculation was conducted several times each day during scoring, providing real-time measures of rater accuracy.

In addition to validity responses, 10 percent of hand scored responses received blind second reads, the results of which were used to calculate interrater reliability. To support interpretability, second reads were conducted exclusively by expert (that is, highly accurate) raters, described later in this subsection.

The MI VSC system automatically and randomly routes the requisite number of responses to raters for second reads and validity in a blind manner. Raters did not know if they were scoring a first read, a second read, or a validity response. This system also prohibits raters from being eligible to score second reads for responses they have already scored.

Scoring accuracy during hand scoring was maintained by continuously assessing rater performance using validity responses. MI specifically evaluated how closely raters’ scores aligned with the benchmark scores of these validity responses. Key performance measures included the agreement between rater and benchmark scores, quantified using quadratic-weighted kappa (QWK),[[9]](#footnote-10) and the comparison of mean score differences between the distributions of benchmark and rater-assigned scores.

During scoring, raters received automated feedback based on recent performance. The automated feedback system identified raters requiring additional feedback—based on accuracy metrics—and automatically generated a custom set of responses for the rater to review. The system functions at the item level, thus providing feedback even to those raters with relatively high accuracy when the data identifies there are one or more items on which they can improve.

VSC Reports provided real-time reports throughout the scoring effort. These reports are available for access by hand scoring management. Interrater reliability reports provide the percentage of exact, adjacent, and nonadjacent agreement for scorable responses. Score point frequency distribution reports provide the percentage per score point and include the mean and standard deviation (SD) for each item. Validity performance reports provide the percentage of exact, adjacent, and nonadjacent agreement for validity responses and are used to monitor and correct drift at the group level. If the data indicated that raters as a group were scoring validity responses either consistently high or consistently low, leadership recalibrated the group by having raters review key training responses that reflect the types of responses being missed in validity. Leadership may also provide raters with a supplemental set of responses that help reinforce the lines for the various score points and reanchor the raters to the proper position, arresting groupwide drift.

Automated removal of raters and score resets are performed when item and rater performance fail to meet accuracy expectations. In these cases, all responses scored by a rater during a period of poor performance were reset and redistributed to other qualified raters for rescoring. By limiting raters to scoring relatively fewer items, this approach also maximized accuracy across items.

In addition to the automated feedback, scoring leadership provided individualized feedback to raters based on their performance. When leadership noticed a rater with low interrater reliability, leadership reviewed the rater’s misscored validity responses and associated data and look for a trend that suggested the scorer drifted from the anchored responses. If such a trend was present, leadership would tailor feedback specific to that rater, typically by presenting the rater with live responses the rater has misscored in a way that is reflective of the rater’s overall drift from the anchor set criteria and providing targeted, thoughtful rationales for the “correct” scores.

The system automatically generates performance metrics several times a day based on the most recent data, providing raters and scoring managers with daily, automated summaries of rater performance. This ensured that all hand scoring staff were kept informed of their current performance and any issues that needed attention. In addition to these daily summaries, detailed manager-level reports were produced to identify raters who required retraining or, if necessary, removal due to accuracy or productivity concerns. These reports enabled scoring management to direct scoring leaders to specific VSC reports, allowing them to pinpoint the areas where individual raters needed improvement.

The MI monitoring system afforded the objective, dynamic identification of the most accurate raters, to whom MI referred as “expert raters.” Specifically, expert raters are those with a demonstrated ability to score validity responses, including anchor validity responses originating from the field test administration,[[10]](#footnote-11) with a high degree of accuracy. Rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Expert rater status was a precondition for conducting second readings (refer to subsection [*9.3.1.2.2 Monitoring Raters*](#_Monitoring_Raters_1)).

Finally, as a supplement to automated assessments, team leaders spot-checked (that is, “read behind”) raters’ scoring to ensure that the raters were on target, conducting one-on-one retraining sessions to address any problems found. At the beginning of a project, team leaders read behind every rater every day; they become more selective about the frequency and number of read-behinds as raters became more proficient at scoring.

##### Rater Retraining and Dismissal

Retraining is an ongoing process once scoring is underway. Daily analysis of the rater status reports enables management personnel to identify individual or group retraining needs. If it becomes apparent that a whole team or group is having difficulty with a particular type of response, large-group training sessions are conducted.

When read-behinds or daily statistics identify a rater who cannot maintain acceptable agreement rates, the rater is retrained and monitored by scoring leadership personnel. A rater may be released from the project if retraining is unsuccessful. In these situations, all items scored by a rater during the timeframe in question can be identified, reset, and released back into the scoring pool. The aberrant rater’s scores are deleted, and the responses are redistributed to other qualified raters for rescoring.

##### Rater Agreement

The final statistics of the human scoring for interrater reliability for all items at all grade levels are presented in [appendix 8.I](#_Appendix_8.I:_Interrater). These statistics include the percentage of perfect agreement and adjacent agreement between the two readers and the QWK statistic. QWK is a statistic used to measure the degree of association between two ratings with values ranging from 0.0 (indicating no agreement) to 1.0 (indicating perfect agreement). Refer to subsection [*8.6.8.1.3 Quadratic-Weighted Kappa*](#_Quadratic-Weighted_Kappa_3) for detailed information on QWK.

Hand scoring interrater reliability is computed only on the basis of scorable responses (numeric scores) scored by two independent raters. Nonscorable responses (for example, off-topic, off-purpose, or foreign-language responses) that are scored by scoring leadership—not by two independent raters—are excluded from interrater reliability computations. In ELA, WER items (full writes) are scored in three dimensions: conventions (0–2 rubric), evidence and elaboration or development and elaboration (depending on the prompt purpose; 0–4 rubric), and organization and purpose (0–4 rubric). The short-answer items are scored with a 0–2 rubric. Mathematics short-answer items are scored using 0–1, 0–2, or 0–3 rubrics.

Smarter Balanced provided flagging criteria (Smarter Balanced, 2016a) based on the statistics that follow for identifying items to be reviewed for potential elimination after scoring was completed. Polytomous items were flagged if any of the following conditions occurred:

* Adjacent plus exact agreement < 0.80
* Exact agreement < 0.60
* QWK < 0.20

Dichotomous items were flagged if either of the following conditions occurred:

* Exact agreement < 0.80
* QWK < 0.20

ETS followed the Smarter Balanced recommended exact and adjacent agreement rates criteria. However, ETS used a more stringent agreement criterion for QWK; that is, dichotomous and polytomous items were flagged if their QWK was below 0.70.

Table 7.3 shows the number of human-scoring items flagged by content area, grade level, and scoring method. Note that human–automated scoring interrater reliability is evaluated in the hybrid model in subsection [*7.2.6 Automated Scoring*](#_Automated_Scoring). Results of human–automated scoring interrater reliability are presented in table 7.11 through table 7.16.

There were 307 items flagged among 1,043 scored items across all grade levels in ELA and mathematics. The summary results show that 71 percent of all CR scored items had interrater reliabilities that met or exceeded the minimum agreement rate thresholds. In general, mathematics short-answer items tended to be flagged less often than ELA short-answer or WER items. ELA short-answer items were flagged particularly often in grades six through eight. All flags were due to low QWK values (QWK < 0.70). Note that in table 7.3, “N/A” was placed in ELA dichotomous items, because there were no dichotomous CR items in ELA.

Table 7.3 Number of Human-Scored CR Items Flagged, by Content Area and Grade Level

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Scoring Method** | **Content Area** | **Grade Level** | **Flagged Polytomous Items** | **Flagged Dichotomous Items** | **Total Flagged Items** | **Total Number of Scored Items** | **Percentage Flagged** |
| Human-to-Human Short Answer (SA) | ELA | 3 | 2 | 0 | 2 | 9 | 22 |
| Human-to-Human SA | ELA | 4 | 3 | 0 | 3 | 9 | 33 |
| Human-to-Human SA | ELA | 5 | 0 | 0 | 0 | 5 | 0 |
| Human-to-Human SA | ELA | 6 | 31 | 0 | 31 | 40 | 78 |
| Human-to-Human SA | ELA | 7 | 27 | 0 | 27 | 44 | 61 |
| Human-to-Human SA | ELA | 8 | 22 | 0 | 22 | 41 | 54 |
| Human-to-Human SA | ELA | 11 | 25 | 0 | 25 | 80 | 31 |
| Human-to-Human SA | Mathematics | 3 | 0 | 2 | 2 | 36 | 6 |
| Human-to-Human SA | Mathematics | 4 | 1 | 4 | 5 | 30 | 17 |
| Human-to-Human SA | Mathematics | 5 | 3 | 1 | 4 | 73 | 5 |
| Human-to-Human SA | Mathematics | 6 | 9 | 4 | 13 | 41 | 32 |
| Human-to-Human SA | Mathematics | 7 | 1 | 3 | 4 | 18 | 22 |
| Human-to-Human SA | Mathematics | 8 | 1 | 7 | 8 | 28 | 29 |
| Human-to-Human SA | Mathematics | 11 | 2 | 3 | 5 | 34 | 15 |
| Human-to-Human WER | ELA | 3 | 32 | N/A | 32 | 75 | 43 |
| Human-to-Human WER | ELA | 4 | 20 | N/A | 20 | 81 | 25 |
| Human-to-Human WER | ELA | 5 | 19 | N/A | 19 | 81 | 23 |
| Human-to-Human WER | ELA | 6 | 17 | N/A | 17 | 57 | 30 |
| Human-to-Human WER | ELA | 7 | 24 | N/A | 24 | 90 | 27 |
| Human-to-Human WER | ELA | 8 | 23 | N/A | 23 | 90 | 26 |
| Human-to-Human WER | ELA | 11 | 21 | N/A | 21 | 81 | 26 |
| **Overall:** | **N/A** | **N/A** | **283** | **24** | **307** | **1,043** | **29** |

#### Automated Scoring

The MI PEG automated scoring technology is used to score eligible short-answer and WER items in ELA and short-answer items in mathematics. This subsection describes PEG, the training and validation sample and process, and the automated scoring process.

##### Project Essay Grade

Figure 7.2 presents the architecture of the MI PEG engine. During engine training, this architecture allows PEG to generate hundreds of custom linguistic (rule-based) features, which are determined by codified English linguistic rules such as syntax and semantics and extracted from representative student responses. In addition to rule-based features, PEG also includes features extracted by Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) procedures.

PEG item- and trait-specific scoring models use computed features from the training responses, along with the scores assigned to them by expert human raters. Using hundreds of parameterizations across several machine-learning algorithms, via cross validation and optimization, PEG determines which algorithms best predict the expert-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate linear and nonlinear classification and regression models. These approaches typically result in 100 candidate models for a single item or trait. PEG then uses an ensembling procedure to combine the best models into a robust final model. The ensembling procedure uses a linear regression, where the objective is to maximize a continuous relaxation of the QWK metric, thus maximizing PEG agreement with the expert human raters.

*Refer to the* [*Alternative Text for Figure 7.2*](#_Alternative_Text_for_14) *for a description of this flow diagram.*

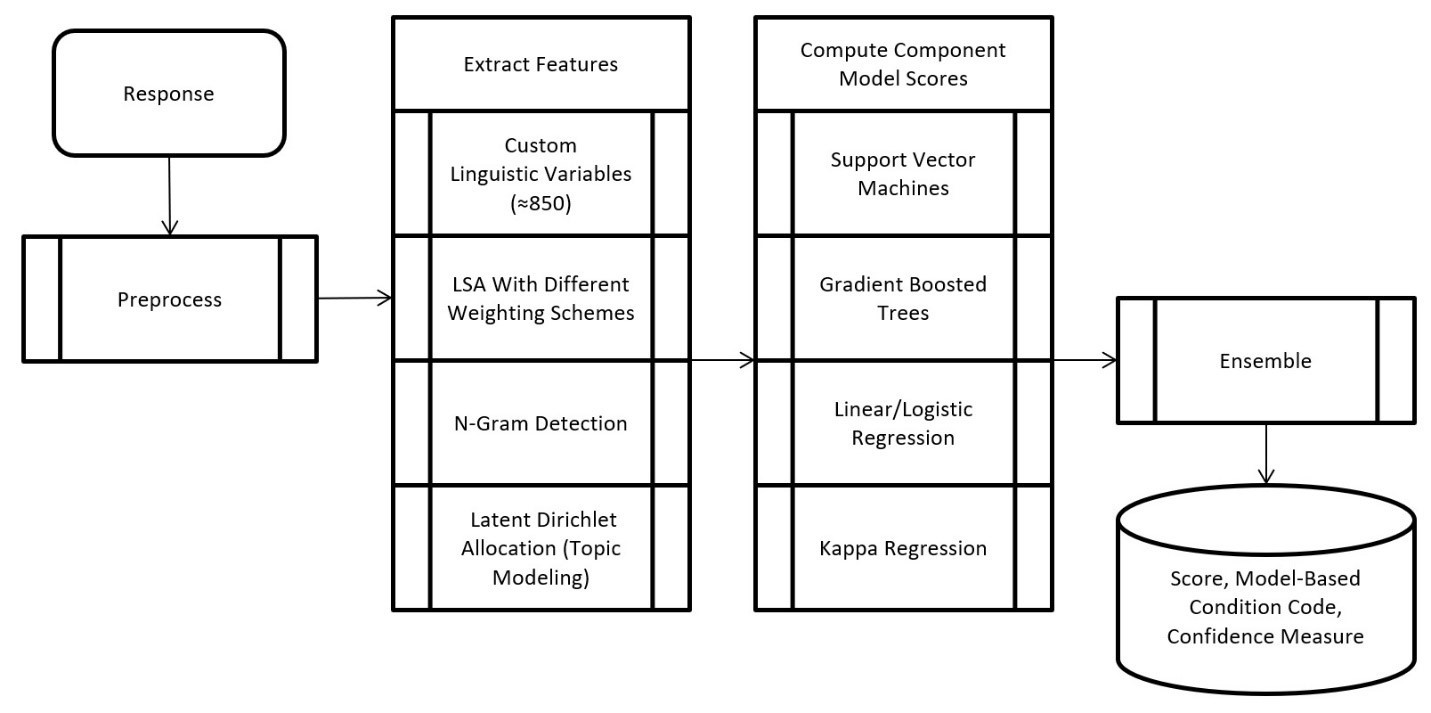


Figure 7.2 PEG engine overview

Since the 2014–15 Smarter Balanced administration, PEG has been used for operational scoring for California and other Smarter Balanced Assessment Consortium member states.

##### Sample

Automated scoring models were not created for items that had an insufficient quantity of training responses. This was the case for items that had low exposure to students, as dictated by the adaptive testing algorithm. Additionally, mathematics PT items that had multiple parts with scoring dependencies were not considered for automated scoring. Table 7.4 shows the number of items eligible for model training, by grade level and content area. Note that pretrained models existed for all 595 items, and no items were without pretrained models. Thus, no additional training was conducted in preparation for the spring 2024 administration. The remainder of this section describes the process used to train and validate the 595 existing models.

Table 7.4 Number of Items Eligible for Model Training, by Grade Level and Content Area

|  |  |  |  |
| --- | --- | --- | --- |
| **Grade Level** | **ELA Short Answer** | **ELA Essay** | **Mathematics** |
| 3 | 12 | 13 | 44 |
| 4 | 13 | 16 | 42 |
| 5 | 13 | 10 | 50 |
| 6 | 32 | 10 | 41 |
| 7 | 45 | 17 | 15 |
| 8 | 49 | 14 | 24 |
| 11 | 80 | 17 | 38 |
| **Totals:** | 244 | 97 | 254 |

##### Model Training and Validation

Student responses used for training and validation were sourced from the 2018–19, 2020–‍21, 2021–22, and 2022–23 Smarter Balanced operational administrations.

Responses were randomly sampled from available on-grade responses in the operational population. For all items, the sample included 1,500–2,000 responses, stratified by score point. The score of record used to train the engine was the score assigned to each response by an expert rater. Expert raters are those who demonstrate highly accurate and consistent scoring of validity responses.

For each item, the sample was divided as follows:

* Approximately 85 percent of the responses were assigned to a training set used to build the model.
* Approximately 15 percent of the responses were assigned to a validation set used to evaluate the accuracy of the model.

###### Model Training

Component model training requires input of response “features.” For items that assess writing quality (for example, WER items), PEG processes the responses and calculates approximately 850 linguistic variables that describe the responses in mathematical terms. These variables range in complexity from simple to highly complex. Examples of simple variables are measures such as word count or sentence length, word choice and spelling errors, and the number and severity of grammatical errors. The most complex variables measure patterns that represent style, fluidity, smoothness of transitions, clarity of communication, and other sophisticated concepts.

For content-based items (for example, short-answer mathematics items), the number of variables is unknown until the models are built. Because the content varies significantly from item to item, and therefore from model to model, PEG examines training responses and identifies the variables that most accurately capture the content in question using techniques like LSA, n-gram detection, and LDA (a type of topic modeling). To further refine the variable generation process, a computer language that enables the performance of a simultaneous search over semantic, lexographical, and syntactic features of responses was built.

To build an essay scoring model, PEG examines the variables and text features of responses, correlates them with the human scores previously assigned, and identifies those variables that have high predictive value.

To build a content scoring model, PEG analyzes training responses and calculates features that pertain to the content in question. PEG then sends the features to hundreds of different algorithms that compete to determine which algorithms best associate the features with the human-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate both linear and nonlinear models. Examples of approaches used include support vector machines, gradient boosted trees, and various regression approaches.

Note that building component models for each item—and for multidimensional items, each trait or dimension—prevents generalizing variables between items or traits, allowing PEG to faithfully reproduce humans’ application of the scoring rubrics. This means the resultant models are robust to gaming attempts, as each represents a unique valuation of the item- or trait-specific text features similarly valued by professional raters.

The approaches typically result in 100 models for a single item or essay trait. Ensembling is the process of selecting the “best of the best” models, to result in a small set of strong, yet dissimilar, component models. A linear-kappa regression is used to determine the model ensembling weights. The more accurate a given model is, the more weight it carries in the final score decision.

Scoring a response requires first preprocessing the response. The purpose of preprocessing is twofold:

1. Create raw and canonical representations of the response from which features can be extracted
2. Filter out responses for which the scoring model does not apply (for example, blank or insufficient responses)

The response is then scored with the associated component models. A final score is produced performing a weighted sum using the ensembling weights.

###### Model Validation

Model validation involves a two-phase approach: an initial validation using held-out training data and a secondary validation using operational data from the current administration.

Initial Validation

Initial validation is conducted by applying each model to score a respective validation set of responses. The validation set is independent of the training set, in that none of the responses it contains have been used to build the model. Two or more professional raters will not always agree on what score to give a student’s response; therefore, when the engine produces scores that agree with professional raters to the same or greater extent than the raters agree with each other, modeling is considered successful. The initial evaluation was made using the criteria shown in table 7.5, based on criteria proposed by Williamson, Xi, and Breyer (2012). This evaluation process was used for both the item-specific scoring models and the condition code models. While Williamson et al. (2012) recommend an agreement between human and machine scores of 0.70 QWK for normally distributed data, a QWK threshold of 0.65 was adopted because of the prevalence of skewed distributions in response data. The degradation (QWK) criterion of 0.07 is slightly more stringent than proposed by Williamson et al. (2012).

Note that, in table 7.5, QWK is quadratic-weighted kappa, SMD is standardized mean difference, H:H is human:human, and H:M is human:machine.

Table 7.5 Initial Model Evaluation Criteria

|  |  |
| --- | --- |
| **Criterion** | **Threshold** |
| Agreement of automated scores with human scores | QWKH:M ≥ 0.65 |
| Degradation from the human–human score agreement | QWKH:H − QWKH:M < 0.07 |
| Standardized mean score difference between human and automated scores | |SMDH:M| < 0.15 |

Bias Considerations

Student group differences in responses to CR items can introduce construct-irrelevant variance in scores, in turn threatening valid score interpretations. MI investigates potential sources of bias annually, for newly modeled items, as part of the initial validation process, using available data from the previous summative administration.

Table 7.6 shows the demographic variables and categories for the examined student groups.

Table 7.6 Demographic Variables and Categories for Student Groups

|  |  |
| --- | --- |
| **Demographic Variable** | **Categories** |
| Gender | * Male * Female |
| Race or Ethnicity | * American Indian or Alaska Native * Asian * Native Hawaiian or Other Pacific Islander * Filipino * Hispanic or Latino * Black or African American * White * Two or more races |
| English Learner (EL) Status | * EL * Non-EL |

For each new item modeled, analysis was performed on a student group if the number of observations (that is, H:M scores) was at least 10. A student group was flagged for bias if |SMD| ≥ 0.125 and if the SMD was significant at an overall significance level of 95 percent. A Bonferroni correction was used to adjust the significance level for each student group comparison. An item was flagged for bias if any student group comparison associated with the item was flagged.

Secondary Validation

All models associated with items that passed initial validation were subject to a secondary validation at the start of the spring 2024 administration using an early sample of operational responses from that administration. This sample was composed of the first available 500 responses per item across states, at a minimum. Responses from this sample were scored by both the automated scoring engine and an expert rater. During this interval, the human score was reported as the score of record. If the PEG scores were found to be consistent with the scores assigned by the expert rater, subsequent student responses for a given item were scored by PEG using a hybrid approach of human and automated scoring. If not, the item was hand scored.

Table 7.7 presents the secondary validation criteria. Note that since expert raters were the only humans that scored the secondary validation sample, a second human score was not collected, and thus QWK degradation is not part of the criteria. In this table, QWK is quadratic-weighted kappa, SMD is standardized mean difference, and H:M is human:machine.

Table 7.7 Secondary Validation Criteria

|  |  |
| --- | --- |
| **Criterion** | **Threshold** |
| Agreement of automated scores with human scores | QWKH:M ≥ 0.65 |
| Standardized mean score difference between human and automated scores | |SMDH:M| ≤ 0.15 |

Table 7.8 presents the secondary validation results. Of the 595 items with models subject to secondary validation, models associated with 454 of the items (76.3%) passed all secondary evaluation criteria.

Table 7.8 Summary of Secondary Validation Results by Grade Level and Content Area

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Items with All Models Passing Initial Validation Criteria: Short Answer** | **Items with All Models Passing Initial Validation Criteria: Essay** | **Items with All Models Passing Initial Validation Criteria: Mathematics** | **Items with All Models Passing Secondary Validation Criteria: Short Answer** | **Items with All Models Passing Secondary Validation Criteria: Essay** | **Items with All Models Passing Secondary Validation Criteria: Mathematics** |
| 3 | 12 | 13 | 44 | 12 | 3 | 44 |
| 4 | 13 | 16 | 42 | 13 | 6 | 40 |
| 5 | 13 | 10 | 50 | 13 | 5 | 47 |
| 6 | 32 | 10 | 41 | 19 | 5 | 40 |
| 7 | 45 | 17 | 15 | 27 | 9 | 15 |
| 8 | 49 | 14 | 24 | 31 | 9 | 22 |
| 11 | 80 | 17 | 38 | 46 | 10 | 38 |
| **Totals:** | **244** | **97** | **254** | **161** | **47** | **246** |

Live Training and Validation

In April and May 2024, when operational scoring was underway, a live training and validation effort was undertaken for those hand scored items lacking validated models from prior efforts but having sufficient 2023–24 operational responses to train and validate new models. In general, these items were associated with models that had previously failed an initial or secondary validation. In such cases, training with 2023–24 operational responses offered the potential to improve model performance. All models associated with these items were thus trained using either exclusively 2023–24 responses—when a minimum of 1,400 2023–‍24 responses per item existed—or 2023–24 responses supplemented with 2022–23 responses. In either case, the validation sets consisted exclusively of 2023–24 responses. Because live validation involved operational data, it was not necessary to conduct a secondary validation.

Table 7.9 summarizes the results of the live training and validation. Of the 356 items associated with models that underwent live training and validation, models associated with 211 of the items (59.3%) passed all evaluation criteria. While this pass rate is higher than the pass rates observed during the initial (49.8%), it is lower than the secondary (76.3%) validation efforts. This is most likely explained by the nature of the items modeled. Specifically, since all item models in this sample had failed a prior validation, by design the sample consisted of difficult-to-model items.

Table 7.9 Summary of Live Training and Validation Results by Grade Level and Content Area

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Items Trained: Short Answer** | **Items Trained: Essay** | **Items Trained: Mathematics** | **Items with All Models Passing Initial Validation Criteria: Short Answer** | **Items with All Models Passing Initial Validation Criteria: Essay** | **Items with All Models Passing Initial Validation Criteria: Mathematics** |
| 3 | 1 | 25 | 9 | 1 | 16 | 4 |
| 4 | 3 | 24 | 9 | 3 | 19 | 1 |
| 5 | 1 | 25 | 33 | 1 | 14 | 19 |
| 6 | 24 | 16 | 10 | 15 | 10 | 4 |
| 7 | 28 | 20 | 7 | 18 | 12 | 4 |
| 8 | 26 | 25 | 9 | 17 | 6 | 7 |
| 11 | 36 | 21 | 4 | 24 | 12 | 4 |
| **Totals:** | **119** | **156** | **81** | **79** | **89** | **43** |

Following initial validation, secondary validation, and live training and validation, a total of 665 items, composed of 240 ELA short answer, 136 essay, and 289 mathematics short answer, were scored using a hybrid process, described in the following subsection.

##### Hybrid Scoring Process

As models associated with a given item passed secondary validation (or live validation), subsequent student responses were scored using a hybrid human–automated scoring approach. If all models associated with a given item did not pass secondary validation, responses associated with the item were hand scored by the larger pool of raters. These raters were monitored using validity responses and backreads conducted by expert raters, and they and their supervisors (team leaders, scoring directors) received automated, daily reports of their performance (that is, accuracy and productivity).

Figure 7.3 shows the response routing rules under the hybrid scoring process. In the hybrid model, responses were first preprocessed for automated scoring and to filter alert responses and certain nonscorable cases (for example, insufficient text to score or high proportion of copied prompt text). This is achieved through the use of a series of three-digit flags used to indicate condition codes as defined in the hand scoring criteria (refer to table 7.10). For example, PEG flags responses that lack proper development, lack enough content to be scored, or are written in an unsupported language. Responses were then sent to the automated scoring engine, where text features were extracted, the scoring model(s) applied, and responses assigned a score and measure of score confidence (that is, an error estimate based on response features). Low-confidence responses straddle the lines between score point values on a rubric and are difficult to score accurately because they exhibit characteristics of multiple score points. Higher-confidence responses received the engine score as the score of record, while lower-confidence responses were routed directly to expert raters, who assigned the score of record. Note that the expert rater pool was dynamic, and raters were added or removed on a day-to-day basis based on their current performance. Overall, approximately 15 percent of responses to engine-scored items were flagged as low confidence and scored by expert raters.

*Refer to the* [*Alternative Text for Figure 7.3*](#_Alternative_Text_for_40) *for a description of this flow diagram.*

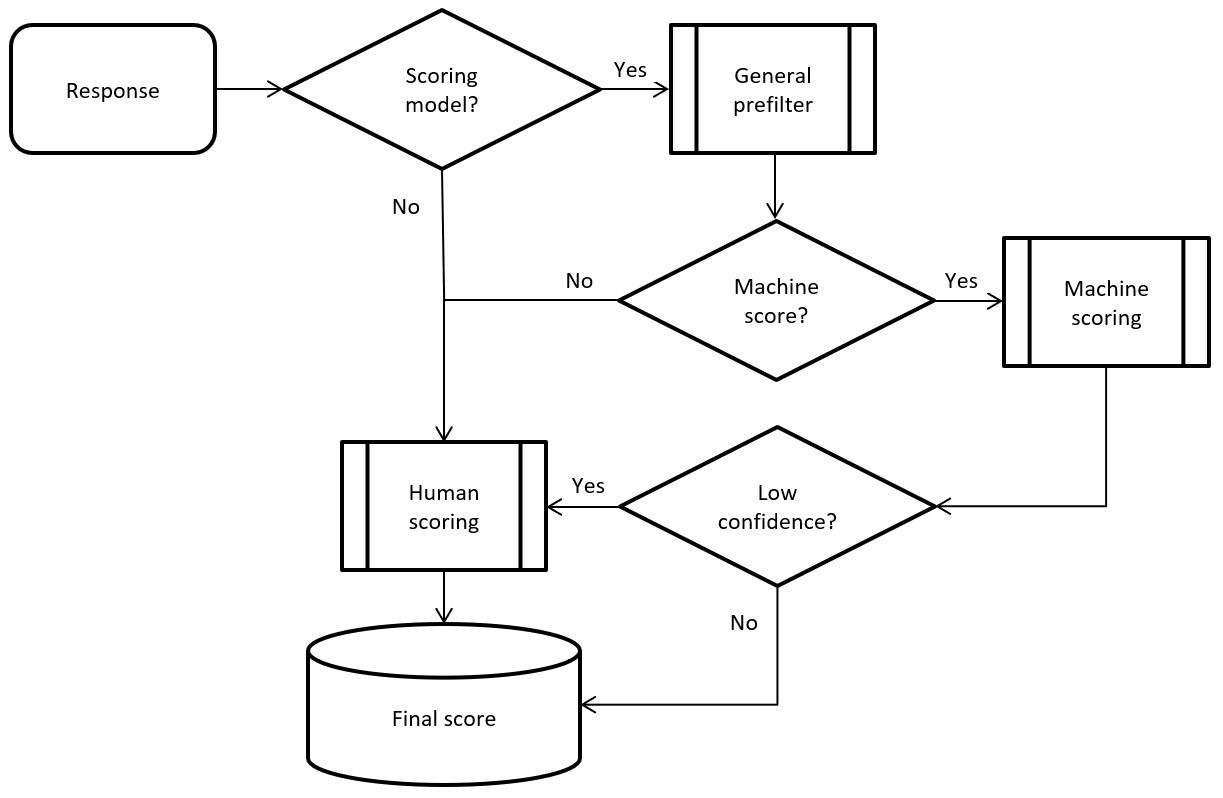


Figure 7.3 Hybrid scoring response routing rules

Upon receipt and validation of each response, MI routed responses for those items eligible for automated scoring to PEG and the remainder of the responses to the MI hand scoring system. Note that, in table 7.10 (which presents flags for alert responses and nonscorable cases), scorable flags indicate instances where PEG will return both the applicable flag *and* a score.

Table 7.10 Flags for Alert Responses and Nonscorable Cases

|  |  |  |
| --- | --- | --- |
| **Flag** | **Usage Description** | **Scorable?** |
| 0 | Standard scoring | Yes |
| 200 | Too few words (that is, blank, or extremely short response) | No |
| 240 | Too long (that is, too many characters submitted; 30,000 characters is the current limit) | No |
| 250 | Expected essay fields are null or empty; set when nulls are discovered within the processing pipeline; not client configurable | No |
| 400 | Unexpected item\_id (that is, the item\_id is not one of the items PEG artificial intelligence has modeled) | No |
| 620 | Applies when the ratio of copied characters exceeds specified threshold (for example, 0.5 means 50%) (can be used for all Smarter Balanced items for which prompt content was provided) | Yes |
| 650 | Insufficient Condition Code (I) (that is, response holds strong general resemblance to those marked “Insufficient” by human readers, but is nonetheless PEG scorable and so, scores are provided)  *PEG Configuration:* Item agnostic; but for 2020–21, applicable to ELA items only | Yes |
| 660 | Language Non-English Condition Code (L) (that is, response holds strong general resemblance to those marked “Non-English” by human readers, but is nonetheless PEG scorable and so, scores are provided)  *PEG Configuration:* Item agnostic; but for 2020–21, applicable to ELA items only | Yes |
| 670 | Off-Topic: Applicable to ELA WER items only and is item specific in the PEG environment | Yes |
| 680 | Off-Mode: Applicable to ELA WER items only and is item specific in the PEG environment | Yes |
| 900 | Timeout (that is, unable to complete essay score prediction within time limits); not client configurable | No |
| 950 | System error processing essay (that is, internal PEG error); not client configurable | No |

##### Quality Assurance

The MI hybrid scoring approach included numerous QA steps. First, models were trained using exclusively scores assigned by expert raters and the associated responses. Second, each automated scoring model was subjected to an evaluation process, as described in the previous model validation subsection. This involved evaluating the quality of the human-scored training data, as well as comparing the performance of the engine to the performance of expert raters. Third, for models trained using responses from prior administrations, the generalizability of each model to the 2023–24 operational responses was confirmed via a secondary validation. Finally, quality was further assured during scoring by routing a minimum of 15 percent of the responses that were most different from the training responses to expert raters and assigning the human score.

##### Project Essay Grade–Human–Automated Scoring Agreement

This subsection summarizes the human–automated scoring agreement for all items scored using a hybrid process in spring 2024, including the following:

* Items passing initial model validation
* Items passing secondary validation
* Items passing live validation

Table 7.11 through table 7.13 present the human–automated scoring agreement on the initial and secondary validation samples for ELA short-answer items, ELA essay items, and mathematics short-answer items, respectively.

Table 7.11 Human–Automated Scoring Agreement for ELA Short-Answer Items on Initial and Secondary Validation Samples by Grade Level

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Initial N of Items** | **Initial % Exact** | **Initial % Exact + Adjacent** | **Initial QWK** | **Secondary N of Items** | **Secondary % Exact** | **Secondary % Exact + Adjacent** | **Secondary QWK** |
| 3 | 12 | 79.6 | 99.6 | 0.81 | 12 | 82.3 | 99.5 | 0.77 |
| 4 | 13 | 80.1 | 99.8 | 0.84 | 13 | 80.9 | 99.8 | 0.80 |
| 5 | 13 | 75.4 | 99.6 | 0.81 | 13 | 77.4 | 99.8 | 0.78 |
| 6 | 19 | 78.7 | 99.5 | 0.81 | 19 | 79.1 | 99.6 | 0.77 |
| 7 | 27 | 76.3 | 99.4 | 0.79 | 27 | 76.4 | 99.4 | 0.75 |
| 8 | 31 | 76.2 | 99.5 | 0.78 | 31 | 75.8 | 99.4 | 0.75 |
| 11 | 46 | 77.2 | 99.5 | 0.79 | 46 | 76.1 | 99.5 | 0.77 |

Table 7.12 Human–Automated Scoring Agreement for ELA Essay Items on Initial and Secondary Validation Samples by Grade Level

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Trait** | **Initial N of Items** | **Initial % Exact** | **Initial % Exact + Adjacent** | **Initial QWK** | **Secondary Number of Items** | **Secondary % Exact** | **Secondary % Exact + Adjacent** | **Secondary QWK** |
| 3 | Conventions | 3 | 71.6 | 99.7 | 0.72 | 3 | 72.5 | 99.5 | 0.70 |
| 3 | Evidence and Elaboration | 3 | 77.9 | 99.2 | 0.82 | 3 | 78.2 | 99.7 | 0.77 |
| 3 | Organization and Purpose | 3 | 75.0 | 99.7 | 0.80 | 3 | 79.1 | 99.6 | 0.78 |
| 4 | Conventions | 6 | 69.2 | 99.0 | 0.74 | 6 | 69.7 | 99.3 | 0.74 |
| 4 | Evidence and Elaboration | 6 | 73.6 | 99.5 | 0.84 | 6 | 73.5 | 99.1 | 0.79 |
| 4 | Organization and Purpose | 6 | 72.2 | 99.2 | 0.82 | 6 | 74.2 | 99.2 | 0.79 |
| 5 | Conventions | 5 | 72.5 | 99.6 | 0.71 | 5 | 73.0 | 99.6 | 0.72 |
| 5 | Evidence and Elaboration | 5 | 73.0 | 99.0 | 0.82 | 5 | 72.6 | 99.6 | 0.80 |
| 5 | Organization and Purpose | 5 | 72.2 | 99.6 | 0.83 | 5 | 72.7 | 99.6 | 0.80 |
| 6 | Conventions | 5 | 75.5 | 99.0 | 0.72 | 5 | 73.5 | 99.5 | 0.74 |
| 6 | Evidence and Elaboration | 5 | 71.4 | 98.7 | 0.78 | 5 | 76.2 | 99.6 | 0.78 |
| 6 | Organization and Purpose | 5 | 69.8 | 98.9 | 0.78 | 5 | 76.2 | 99.6 | 0.78 |
| 7 | Conventions | 9 | 76.1 | 99.7 | 0.70 | 9 | 75.5 | 99.8 | 0.74 |
| 7 | Evidence and Elaboration | 9 | 75.6 | 99.7 | 0.83 | 9 | 81.7 | 99.8 | 0.84 |
| 7 | Organization and Purpose | 9 | 75.6 | 99.6 | 0.84 | 9 | 81.6 | 99.9 | 0.84 |
| 8 | Conventions | 9 | 77.0 | 99.1 | 0.71 | 9 | 76.1 | 99.7 | 0.74 |
| 8 | Evidence and Elaboration | 9 | 73.7 | 99.1 | 0.82 | 9 | 76.9 | 99.6 | 0.80 |
| 8 | Organization and Purpose | 9 | 75.1 | 99.7 | 0.84 | 9 | 77.2 | 99.6 | 0.80 |
| 11 | Conventions | 10 | 79.1 | 99.7 | 0.75 | 10 | 77.1 | 99.6 | 0.73 |
| 11 | Evidence and Elaboration | 10 | 76.5 | 99.7 | 0.86 | 10 | 75.6 | 99.9 | 0.84 |
| 11 | Organization and Purpose | 10 | 76.4 | 99.7 | 0.86 | 10 | 75.8 | 99.9 | 0.83 |

Note that in the final column of table 7.13, *Secondary QWK*, QWK is not presented for 0–1 items because of the binary score scale.

Table 7.13 Human–Automated Scoring Agreement for Mathematics Items on Initial and Secondary Validation Samples by Grade Level

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Score Point Range** | **Initial N of Items** | **Initial % Exact** | **Initial % Exact + Adjacent** | **Initial QWK** | **Secondary N of Items** | **Secondary % Exact** | **Secondary % Exact + Adjacent** | **Secondary QWK** |
| 3 | 0–1 | 10 | 94.2 | 100 | 0.86 | 10 | 94.1 | 100 | NA |
| 4 | 0–1 | 7 | 91.0 | 100 | 0.79 | 7 | 92.3 | 100 | NA |
| 5 | 0–1 | 7 | 92.6 | 100 | 0.81 | 7 | 93.5 | 100 | NA |
| 6 | 0–1 | 8 | 96.6 | 100 | 0.81 | 8 | 95.8 | 100 | NA |
| 7 | 0–1 | 7 | 96.9 | 100 | 0.85 | 7 | 96.8 | 100 | NA |
| 8 | 0–1 | 5 | 90.2 | 100 | 0.75 | 5 | 90.5 | 100 | NA |
| 11 | 0–1 | 16 | 95.6 | 100 | 0.87 | 16 | 94.2 | 100 | NA |
| 3 | 0–2 | 28 | 90.8 | 99.3 | 0.91 | 28 | 90.6 | 99.4 | 0.89 |
| 4 | 0–2 | 29 | 91.0 | 99.7 | 0.91 | 29 | 91.6 | 99.7 | 0.89 |
| 5 | 0–2 | 38 | 88.3 | 99.6 | 0.88 | 38 | 87.9 | 99.5 | 0.84 |
| 6 | 0–2 | 32 | 88.9 | 99.6 | 0.86 | 32 | 89.1 | 99.5 | 0.84 |
| 7 | 0–2 | 8 | 87.0 | 99.4 | 0.80 | 8 | 88.9 | 99.9 | 0.80 |
| 8 | 0–2 | 16 | 89.1 | 99.8 | 0.89 | 16 | 90.3 | 99.7 | 0.86 |
| 11 | 0–2 | 17 | 89.1 | 99.4 | 0.88 | 17 | 88.1 | 99.4 | 0.87 |
| 3 | 0–3 | 6 | 91.1 | 99.8 | 0.96 | 6 | 92.5 | 99.9 | 0.96 |
| 4 | 0–3 | 4 | 87.9 | 99.8 | 0.94 | 4 | 86.8 | 99.6 | 0.93 |
| 5 | 0–3 | 2 | 90.8 | 98.4 | 0.94 | 2 | 89.4 | 98.3 | 0.90 |
| 8 | 0–3 | 1 | 78.2 | 98.0 | 0.88 | 1 | 86.1 | 98.4 | 0.92 |
| 11 | 0–3 | 5 | 85.5 | 99.0 | 0.89 | 5 | 83.7 | 99.0 | 0.88 |

Table 7.14 through table 7.16 present the human‒automated scoring agreement on the live validation samples for ELA short-answer items, ELA essay items, and mathematics short-answer items, respectively. Live training did not involve a secondary validation as it involved operational data.

Table 7.14 Human–Automated Scoring Agreement for ELA Short-Answer Items on Live Validation Sample by Grade Level

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade Level** | **Live N of Items** | **Live % Exact** | **Live % Exact + Adjacent** | **Live QWK** |
| 3 | 1 | 73.8 | 99.3 | 0.66 |
| 4 | 3 | 79.7 | 99.7 | 0.81 |
| 5 | 1 | 70.4 | 98.0 | 0.73 |
| 6 | 15 | 77.6 | 99.5 | 0.73 |
| 7 | 18 | 78.5 | 99.7 | 0.74 |
| 8 | 17 | 76.1 | 99.6 | 0.74 |
| 11 | 24 | 76.5 | 99.6 | 0.77 |

Table 7.15 Human–Automated Scoring Agreement for ELA Essay Items on Live Validation Sample by Grade Level

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Trait** | **Live Number of Items** | **Live % Exact** | **Live % Exact + Adjacent** | **Live QWK** |
| 3 | Conventions | 16 | 70.5 | 99.6 | 0.71 |
| 3 | Evidence and Elaboration | 16 | 73.4 | 98.8 | 0.77 |
| 3 | Organization and Purpose | 16 | 72.8 | 99.0 | 0.77 |
| 4 | Conventions | 19 | 69.4 | 99.2 | 0.73 |
| 4 | Evidence and Elaboration | 19 | 72.2 | 98.9 | 0.78 |
| 4 | Organization and Purpose | 19 | 73.0 | 99.2 | 0.79 |
| 5 | Conventions | 14 | 70.8 | 99.5 | 0.70 |
| 5 | Evidence and Elaboration | 14 | 70.1 | 99.0 | 0.78 |
| 5 | Organization and Purpose | 14 | 70.2 | 99.1 | 0.79 |
| 6 | Conventions | 10 | 73.2 | 99.4 | 0.72 |
| 6 | Evidence and Elaboration | 10 | 73.6 | 99.3 | 0.79 |
| 6 | Organization and Purpose | 10 | 74.0 | 99.4 | 0.79 |
| 7 | Conventions | 12 | 71.5 | 99.6 | 0.72 |
| 7 | Evidence and Elaboration | 12 | 74.6 | 99.4 | 0.80 |
| 7 | Organization and Purpose | 12 | 74.8 | 99.4 | 0.81 |
| 8 | Conventions | 6 | 76.7 | 99.6 | 0.72 |
| 8 | Evidence and Elaboration | 6 | 76.9 | 99.8 | 0.84 |
| 8 | Organization and Purpose | 6 | 74.8 | 99.8 | 0.83 |
| 11 | Conventions | 12 | 75.8 | 99.5 | 0.73 |
| 11 | Evidence and Elaboration | 12 | 76.0 | 99.7 | 0.84 |
| 11 | Organization and Purpose | 12 | 76.2 | 99.8 | 0.84 |

Note that in the final column of table 7.16, *Live QWK*, QWK is not presented for 0–1 items because of the binary score scale.

Table 7.16 Human–Automated Scoring Agreement for Mathematics Items on Live Validation Samples by Grade Level

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Grade Level** | **Score Point Range** | **Live Number of Items** | **Live % Exact** | **Live % Exact + Adjacent** | **Live QWK** |
| 3 | 0–1 | 3 | 94.4 | 100 | NA |
| 4 | 0–1 | 1 | 88.7 | 100 | NA |
| 5 | 0–1 | 4 | 95.4 | 100 | NA |
| 6 | 0–1 | 1 | 91.4 | 100 | NA |
| 7 | 0–1 | 1 | 100 | 100 | NA |
| 8 | 0–1 | 3 | 87.8 | 100 | NA |
| 3 | 0–2 | 1 | 100 | 100 | 1 |
| 5 | 0–2 | 14 | 84.1 | 99.4 | 0.82 |
| 6 | 0–2 | 3 | 87.3 | 99.2 | 0.81 |
| 7 | 0–2 | 3 | 90.1 | 99.1 | 0.88 |
| 8 | 0–2 | 3 | 92.3 | 100 | 0.92 |
| 11 | 0–2 | 3 | 97.6 | 100 | 0.98 |
| 5 | 0–3 | 1 | 88.3 | 98.7 | 0.91 |
| 8 | 0–3 | 1 | 72.2 | 97.0 | 0.89 |
| 11 | 0–3 | 1 | 90.2 | 98.8 | 0.89 |

### Student Test Scores

ETS developed two parallel scoring systems to produce students’ scores: the eSKM scoring system, which scores and delivers individual students’ scores to the ETS reporting system; and the parallel scoring system developed by ETS Technology and Information Processing Services (TIPS), which computes individual students’ scores. The two scoring systems independently applied the same scoring algorithms and specifications. ETS psychometricians verified the eSKM scoring by comparing all individual student scores from TIPS and resolving any discrepancies. This process redundancy is an internal quality-control step that is in place to verify the accuracy of scoring. Students’ scores were reported only when the two parallel systems produced identical results with acceptable tolerance.

The ETS PAR and eSKM teams investigated and resolved any mismatching scores. (For example, the mismatch could be a result of a Smarter Balanced and CDE decision to not score an item as a problem was identified in a particular item or rubric.) ETS would apply a problem item notification (PIN) not to score the item through the systematic process in eSKM, which might result in a mismatch if TIPS were still in the process of applying the PIN in the parallel system when the student score was being compared. This real-time scoring check is designed to detect mismatches and track remediation.

All scores must comply with the ETS scoring specifications and the parallel scoring process to ensure the quality and accuracy of scoring and to support the transfer of scores into the database of the student records scoring system, the Test Operations Management System (TOMS).

#### Total Test Scores

##### Theta Scores

For all of the assessments, theta scores (IRT ability estimates) are obtained through maximum likelihood estimation (MLE) applied to item scores (Birnbaum, 1968). Items scored as one (correct) or zero (incorrect) are referred to as dichotomous items. Items scored from zero to some number of points greater than one are called polytomous items. The GPCM is applied to both types of items. The GPCM (Muraki, 1992) is presented in equation 7.1. *Refer to the* [*Alternative Text for Equation 7.1*](#_Alternative_Text_for_7) *for a description of this equation.*

 (7.1)

where,

*Pih (θj)* is the probability of student with proficiency *θj* obtaining score *h* on item *i,*

*ni* is the maximum number of score points for item *i*,

*ai* is the discrimination parameter for item *i*,

*bi* is the location parameter for item *i*,

*div* is the category parameter for item *i* on score *v*, and

*D* is a scaling constant of 1.7 that makes the logistic model approximate the normal ogive model.

When *ni* = 1, equation 7.1 becomes an expression of the two-parameter logistic model for dichotomous items.

The log-likelihood of a student with proficiency *θj*, given the observed response vector *v*, is presented in equation 7.2. *Refer to the* [*Alternative Text for Equation 7.2*](#_Alternative_Text_for_1) *for a description of this equation.*

 (7.2)

where,

*I* is the total number of items in the response vector,

*ni* is the maximum number of score points for item *i*,

*Pih* is the probability of the score *h* observed on item *i*, as expressed in equation 7.1, and

*uiv* is presented in equation 7.3. *Refer to the* [*Alternative Text for Equation 7.3*](#_Alternative_Text_for_35) *for a description of this equation.*

 (7.3)

The theta that is associated with the largest log-likelihood for a particular pattern of scores is the maximum likelihood theta estimate. The equation for the MLE cannot generally be solved explicitly, as it is nonlinear in nature (Hambleton & Swaminathan, 1985, p. 79). As a result, an iterative process such as the Newton-Raphson procedure is employed. At iteration *t*, a student’s estimated ability *θ* is calculated using equation 7.4. *Refer to the* [*Alternative Text for Equation 7.4*](#_Alternative_Text_for_2) *for a description of this equation.*

 (7.4)

where,

 is the first derivative of the log-likelihood at iteration *t*−*1*, and

 is the second derivative.

When the difference between the estimates in successive iterations becomes acceptably small (that is, difference is less than 0.0001), the process is said to converge. The convergence criterion determines the level of accuracy of estimation, provided that the process converges. Theta scores are the basis for scale scores but are not reported. Scale scores and the transformation from theta scores to scale scores are described in subsection [*7.3.1.4 Scale Scores for the Total Assessment*](#_7.4.1.2_Scale_Scores).

##### Inverse Test Characteristic Curve Method

There are some special cases in which the score reported for a student is not based on the MLE approach:

* The student got the lowest possible score on the total assessment, which would lead to an MLE of −∞.
* The student got the highest possible score on the total assessment, which would lead to an MLE of +∞.
* The student’s response pattern did not lead to a single most likely MLE of the student’s ability, or the likelihood function was so flat that its maximum was not much greater than the likelihood over a wide range of theta values.

In these cases, the student’s score is computed by the inverse test characteristic curve (TCC) method (Stocking, 1996). This method transforms the sum of the student’s item scores into an ability estimate. That estimate is the ability level at which the sum of the expected scores on the items the student took is equal to the sum of the scores that the student actually earned on those items.

The item characteristic curve for an item shows the probability of a correct answer to the item (in the case of dichotomous items) or the probability of responding in a score category (in the case of polytomous items) as a function of the student’s ability. The TCC for a set of items shows the expected total score on those items as a function of the student’s ability. Because information is lost by not using each student’s unique pattern of responses, this method is used only when the response pattern does not lead to one clear MLE of the student’s ability or the likelihood function is so flat that although it has a maximum, there is a wide range of theta values at which the likelihood is only slightly less than the maximum.

The lowest obtainable theta (LOT) and the highest obtainable theta (HOT) defined by the Smarter Balanced Assessment Consortium are presented in table 7.17 for each grade level and content area (Smarter Balanced, 2016b).[[11]](#footnote-12) The theta scores for grades three through eight and grade eleven are on a common vertical scale.

Table 7.17 Theta of Lowest and Highest Obtainable Scores

|  |  |  |
| --- | --- | --- |
| **Content Area and Grade Level** | **LOT** | **HOT** |
| ELA 3 | −4.5828 | 1.6527 |
| ELA 4 | −4.2914 | 2.1189 |
| ELA 5 | −3.5921 | 2.5851 |
| ELA 6 | −3.2424 | 3.0513 |
| ELA 7 | −2.8928 | 3.5175 |
| ELA 8 | −2.5431 | 3.9837 |
| ELA 11 | −2.4266 | 4.5664 |
| Mathematics 3 | −4.0971 | 1.8298 |
| Mathematics 4 | −3.9079 | 2.3342 |
| Mathematics 5 | −3.7188 | 2.8386 |
| Mathematics 6 | −3.5296 | 3.3430 |
| Mathematics 7 | −3.3405 | 3.8474 |
| Mathematics 8 | −3.1513 | 4.3518 |
| Mathematics 11 | −2.9622 | 4.8562 |

##### Scoring of Incomplete Cases

Sometimes students fail to complete their assessments. Depending on the nature of the missing data, different actions are taken. This subsection covers the following three situations:

1. Attemptedness/Test-taking rules that describe when an assessment is considered attempted or participated
2. When an assessment is scored
3. How and when incomplete assessments are scored

As defined in the Smarter Balanced scoring specifications, assessments are considered “complete” if students respond to at least the minimum number of operational items specified in the blueprint. Otherwise, the assessments are “incomplete.” (Refer to table 8.1 and table 8.2 for the minimum number of items in each composite claim in the adjusted, shortened-form blueprint, in the full-form blueprint, and the adjusted, shortened embedded field test blueprint, respectively.) In a fixed-form (that is, not CAT) assessment, unanswered items are treated as incorrect. However, in a CAT environment, the specific unanswered items are unknown, because the test administration terminates when a student stops responding to items. ETS implemented several procedures that score an incomplete assessment in a CAT environment; these procedures are presented in table 7.18.

Table 7.18 Treatment of Incomplete Assessments

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **If the student ...** | **Classify the student as participating?** | **Include the data in the student file?** | **Score the responses for the student?** | **Classify the student as attempting the assessment?** | **Report a score for the student?** |
| Logged on to both the CAT and PT but answered no items | Yes | Yes | No | No | No |
| Logged on to both the CAT and PT and answered at least one item for only CAT or PT | Yes | Yes | Lowest obtainable score for the assessment | No | No |
| Logged on to both the CAT and PT and answered at least 1 PT item and at least 1, but fewer than 10, CAT items | Yes | Yes | Lowest obtainable score for the assessment | Yes | No |
| Logged on to both the CAT and PT, answered at least 1 PT item and at least 10 CAT items, but did not answer a specified minimum number of items for a complete assessment | Yes | Yes | MLE (unanswered items in the middle of the assessment scored as incorrect), or for an incomplete assessment, estimate from equation 7.5 | Yes | Yes |

The number and percentage of all students and students in the demographic student groups specified in table 7.29 who took the assessments are presented in the tables of [appendix 8.A](#_Alternative_Text_for_38). Results in [appendix 8.A](#_Alternative_Text_for_38) show that, in general, more than 96 percent of students took the Smarter Balanced Summative Assessments in grades three through eight for ELA and mathematics. In high school, student test-taking rates were 94 percent in ELA and mathematics.

Among the demographic student groups, most student groups regardless of grade level or content area had very high test-taking rates. In general, high school had slightly lower test-taking rates (approximately 3 percent lower) than students in grades three through eight. In ELA, EL students and the homeless students had slightly lower test-taking rates, approximately 5 percent lower than the overall 96 percent of test taking rates.

Note that students in the English classification–No response and English classification–To be determined groups had lower participation rates than other groups. However, these student groups are not considered in the previous summary because of their small sample sizes.

The number of students overall and in the selected demographic student groups with different test completion conditions are presented in the tables of [appendix 7.E](#_Appendix_7.E:_Student). Among students who logged on to the assessments, a majority of students completed the ELA and mathematics assessments. Students who logged on to both the CAT and PT but either did not answer any items in both sessions; or responded to at least one item in either the CAT or PT but did not answer any items in the other; or responded to at least one item in each session but fewer than 10 CAT items (combining the first three rows in table 7.E.1 and table 7.E.2) ranged from 0.03 percent to 0.2 percent across all grade levels and content areas.

Sometimes a student stops answering items before the TDS has administered all the items the student is supposed to answer. When that happens, the student’s assessment is considered complete if the student has answered at least a specified minimum number of items (less than the number of items in the full assessment). Otherwise, the student’s score is based on an adjusted ability estimate calculated by the formula in equation 7.5. *Refer to the* [*Alternative Text for Equation 7.5*](#_Alternative_Text_for_3) *for a description of this equation.*

 (7.5)

where,

*θadj* is the student’s adjusted ability estimate;

*θachieved* is the theta estimate based on the incomplete assessment;

*θmin* is a predetermined theta estimate equal to −3.5, which is the average of the LOT values across all assessments (on the vertical theta scale); and

*PropAdj* is the proportion of the assessment completed by the student.

##### Scale Scores for the Total Assessment

After MLE scoring is performed on the theta scale and the scoring rules are implemented, the scaling constants are applied. Scale scores (SS) are on the Smarter Balanced vertical scale and are formed by linking across grade levels using common items in adjacent grade levels. The vertical scale score is the linear transformation of the post–vertically scaled IRT ability estimate (refer to subsection [*2.6.3 Vertical Scaling*](#_Vertical_Scaling) for the procedure). The student’s estimated theta score is converted to a scale score using equation 7.6 for ELA. *Refer to the* [*Alternative Text for Equation 7.6*](#_Alternative_Text_for_4) *for a description of this equation.*

For ELA: SS = 85.8 *θ* + 2508.2 (7.6)

The student’s estimated theta score is converted to a scale score using equation 7.7 formathematics. *Refer to the* [*Alternative Text for Equation 7.7*](#_Alternative_Text_for_5) *for a description of this equation.*

For mathematics: SS = 79.3 *θ* + 2514.9 (7.7)

There is a restriction that the scale score cannot be higher or lower than the specified highest and lowest possible scores for that content area and grade level. The revised lowest obtainable scale score (LOSS) and revised highest obtainable scale score (HOSS) for each assessment are displayed in table 7.19. The CDE revised LOSS and HOSS and started to implement the revisions in the 2020–21 administration to measure student growth over years more effectively. Scale scores are rounded to the nearest integer.

Table 7.19 Lowest and Highest Obtainable Scale Scores

|  |  |  |
| --- | --- | --- |
| **Content Area and Grade Level** | **LOSS** | **HOSS** |
| ELA 3 | 2115 | 2650 |
| ELA 4 | 2140 | 2690 |
| ELA 5 | 2200 | 2730 |
| ELA 6 | 2230 | 2770 |
| ELA 7 | 2260 | 2810 |
| ELA 8 | 2290 | 2850 |
| ELA 11 | 2300 | 2900 |
| Mathematics 3 | 2190 | 2660 |
| Mathematics 4 | 2205 | 2700 |
| Mathematics 5 | 2220 | 2740 |
| Mathematics 6 | 2235 | 2780 |
| Mathematics 7 | 2250 | 2820 |
| Mathematics 8 | 2265 | 2860 |
| Mathematics 11 | 2280 | 2900 |

Detailed information regarding the establishment of scale scores for the Smarter Balanced Summative Assessments can be found in chapter 10 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a) and the *Smarter Balanced Scoring Specification: 2014–2015 Administration* (American Institutes for Research [AIR], 2015). The original LOSS and HOSS were set by the Smarter Balanced Assessment Consortium in 2015.

##### Achievement Levels

Standard settings were performed by the Smarter Balanced Assessment Consortium, which defined four achievement levels based on overall scale scores. These achievement level categories were labeled *Standard Not Met,* *Standard Nearly Met,* *Standard Met,* and *Standard Exceeded*. The combined categories of *Standard Met* and *Standard Exceeded* are used to define students meeting the proficiency criterion for accountability purposes. Refer to *Chapter 10: Achievement Level Setting* of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a) for details related to the standard setting procedure; *Reporting Achievement Level Descriptors* (Smarter Balanced, 2015) for the descriptors used to describe Smarter Balanced achievement levels; and *Interpretation and Use of Scores and Achievement Levels* (Smarter Balanced, 2014b) for more information about using achievement levels.

Levels are defined as follows:

* **Level 1—Standard Not Met:** Student demonstrates minimal understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
* **Level 2—Standard Nearly Met:** Student demonstrates partial understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
* **Level 3—Standard Met:** Student demonstrates adequate understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.
* **Level 4—Standard Exceeded:** Student demonstrates thorough understanding of ELA and mathematics and the ability to apply the knowledge and skills for his or her grade level that are associated with college and career readiness.

The thresholds for the achievement levels vary by grade level and content area. Table 7.20 provides the theta thresholds for Standard Nearly Met, Met, and Exceeded at each grade level. For example, the threshold of −0.888 for *Standard Met* in grade three ELA means that a student must earn a theta score (𝜃) of −0.888 or higher to achieve that classification.

Table 7.20 Theta Thresholds for Achievement Levels

|  |  |  |  |
| --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Nearly Met** | **Standard Met** | **Standard Exceeded** |
| ELA 3 | −1.646 | −0.888 | −0.212 |
| ELA 4 | −1.075 | −0.410 | 0.289 |
| ELA 5 | −0.772 | −0.072 | 0.860 |
| ELA 6 | −0.597 | 0.266 | 1.280 |
| ELA 7 | −0.340 | 0.510 | 1.641 |
| ELA 8 | −0.247 | 0.685 | 1.862 |
| ELA 11 | −0.177 | 0.872 | 2.026 |
| Mathematics 3 | −1.689 | −0.995 | −0.175 |
| Mathematics 4 | −1.310 | −0.377 | 0.430 |
| Mathematics 5 | −0.755 | 0.165 | 0.808 |
| Mathematics 6 | −0.528 | 0.468 | 1.199 |
| Mathematics 7 | −0.390 | 0.657 | 1.515 |
| Mathematics 8 | −0.137 | 0.897 | 1.741 |
| Mathematics 11 | 0.354 | 1.426 | 2.561 |

Table 7.21 shows the scale score range of each achievement level for the ELA and mathematics assessments, respectively.

Table 7.21 Scale Score Ranges for Achievement Levels

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Not Met** | **Standard Nearly Met** | **Standard Met** | **Standard Exceeded** |
| ELA 3 | 2115–2366 | 2367–2431 | 2432–2489 | 2490–2650 |
| ELA 4 | 2140–2415 | 2416–2472 | 2473–2532 | 2533–2690 |
| ELA 5 | 2200–2441 | 2442–2501 | 2502–2581 | 2582–2730 |
| ELA 6 | 2230–2456 | 2457–2530 | 2531–2617 | 2618–2770 |
| ELA 7 | 2260–2478 | 2479–2551 | 2552–2648 | 2649–2810 |
| ELA 8 | 2290–2486 | 2487–2566 | 2567–2667 | 2668–2850 |
| ELA 11 | 2300–2492 | 2493–2582 | 2583–2681 | 2682–2900 |
| Mathematics 3 | 2190–2380 | 2381–2435 | 2436–2500 | 2501–2660 |
| Mathematics 4 | 2205–2410 | 2411–2484 | 2485–2548 | 2549–2700 |
| Mathematics 5 | 2220–2454 | 2455–2527 | 2528–2578 | 2579–2740 |
| Mathematics 6 | 2235–2472 | 2473–2551 | 2552–2609 | 2610–2780 |
| Mathematics 7 | 2250–2483 | 2484–2566 | 2567–2634 | 2635–2820 |
| Mathematics 8 | 2265–2503 | 2504–2585 | 2586–2652 | 2653–2860 |
| Mathematics 11 | 2280–2542 | 2543–2627 | 2628–2717 | 2718–2900 |

#### Claim and Composite Claim Scores and Claim Levels (Subscores)

Claims identify knowledge and skills being measured through a set of items. Groups of items in each combination of grade level and content area are formed on the basis of related content standards; outcomes for these groups of items are called claim scores. A claim score is a measure of a student’s performance on the items in that claim.

Note that when Smarter Balanced’s adjusted, shortened-form blueprint was implemented for the 2020–21 test administration, the number of items was not reliable enough to report performance levels for claims at the individual student level. Therefore, individual student claim-level information was not available in the Student Score Report (SSR). However, starting in 2023–24, composite claim results were reported for individual students and are summarized in [appendix 7.C](#_Appendix_7.C:_Summary).

Aggregated claim-level information was reported at group levels within research files available on the Test Results for California’s Assessments website. In addition, claim scores and claim-level information at the state level are analyzed and reported in the technical report.

##### Identified Claims

There are four claims for both ELA assessments and mathematics assessments. However, claims 2 and 4 of mathematics scores are combined because of content similarity and to provide flexibility for item development. Consequently, only three claim scores are reported with the overall mathematics score.

Like the overall assessment, results of each claim are reported as a theta score, a scale score, and a claim strength or weakness. The claims for ELA are identified in table 7.22 and are also available in the blueprints, which are provided in [appendix 4.A](#_Appendix_4.A:_Smarter).

Table 7.22 Claims Identified for ELA

|  |  |
| --- | --- |
| **Claim** | **Description** |
| 1. Reading | Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts. |
| 1. Writing | Students can produce effective and well-grounded writing for a range of purposes and audiences. |
| 1. Listening/‌Speaking | Students can employ effective listening skills for a range of purposes and audiences. |
| 1. Research | Students can engage in research and inquiry to investigate topics and to analyze, integrate, and present information. |

The claims for mathematics are identified in table 7.23 and are also available in the blueprints, which are provided in [appendix 4.A](#_Appendix_4.A:_Smarter).

Table 7.23 Claims Identified for Mathematics

|  |  |
| --- | --- |
| **Claim** | **Description** |
| 1. Concepts and Procedures | Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency. |
| 1. Problem Solving | Students can solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. |
| 1. Communicating/‌Reasoning | Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others. |
| 1. Model and Data | Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems. |

##### Scale Scores for Claims

Claim scores are calculated by applying the MLE approach to the items contained in a particular claim. The resulting ability estimates are converted to claim scale scores by applying [equation 7.6](#EQ7_6) for ELA assessments and [equation 7.7](#EQ7_7) for mathematics assessments. ELA scores are computed for each claim. Mathematics scores are computed for Claim 1, claims 2 and 4 combined, and Claim 3.

Claim scores are based on fewer items than total test scores. As a result, more students get scores of 0 or the maximum allowable score on claim scores than for the total test score. Since MLE cannot estimate proficiency when the score is 0 or the maximum allowable value, claim scores are more often not able to be estimated than total test scores. ETS uses the inverse TCC approach when MLE-derived theta estimates are not available for a claim.

##### Performance Levels for Claims

The relative strengths and weaknesses for each student are reported for each claim. The three performance levels for each claim are as follows:

1. **Above Standard**—Student clearly understands and can successfully apply his or her knowledge to the standards tested in this content area for his or her grade.
2. **Near Standard**—Student shows understanding and can apply his or her knowledge to the standards tested in this content area for his or her grade.
3. **Below Standard**—Student has limited understanding and difficulty applying his or her knowledge to the standards tested in this content area for his or her grade.

Because claim scores are based on fewer items than overall test scores, the standard error of the claim scale scores is included in the determination of the student’s performance level on a claim. *SSclaim* is a student’s estimated scale score on a claim.  is an estimated standard error of measurement (SEM) of the scale score on a claim. A range of possible student scale scores is calculated for each student from SS sub claim minus 1.5 times SE sub SS sub claim. to SS sub claim plus 1.5 times SE sub SS sub claim., each of which is converted to a scale score and rounded to an integer.

If the value at the high end of the score range is less than the minimum scale score associated with the overall *Standard Met* achievement classification, the claim performance level is reported as “Below Standard.” This achievement classification is also assigned when a student’s responses to all items associated with a claim are incorrect.

If the value at the low end of the range is greater than or equal to the minimum scale score associated with the overall *Standard Met* achievement classification, the claim performance level is reported as “Above Standard.” This claim performance level is also reported when all student responses are correct.

Scale score ranges that do not meet either of these classifications are reported as “Near Standard.”

##### Composite Claims

From the 2020–21 test administration, because of the impact by the novel coronavirus disease 2019 pandemic, the CDE adopted the adjusted test blueprint. Given the reduced reliability for the individual claims due to the reduced number of items in each claim, California did not report scores or achievement levels for individual claims at the student level. Starting from the 2023–24 test administration, Smarter Balanced recommended that some individual claims can be combined and reported. The composite claim scores are as follows:

* ELA Composite Claim 1: Reading and listening
* ELA Composite Claim 2: Writing and research
* Mathematics Composite Claim 1: Concepts and procedures
* Mathematics Composite Claim 2: Problem Solving, Communicating Reasoning, and Modeling and Data Analysis (also known as Mathematical Practices)

For individual students taking the adjusted blueprint, composite claim scores are calculated using MLE, as described in subsection [*7.3.1.1 Theta Scores*](#_Theta_Scores); however, the scores are based on the items contained in a particular composite claim. The achievement levels for the composite claims are calculated using the same method as described in [*7.3.2.3 Performance Levels for Claims*](#_7.4.2.2_Performance_Levels). Summary statistics for the composite claims are found in [appendix 7.C](#_Appendix_7.C:_Summary).

In ELA, composite claim scores are computed for Claim 1 and Claim 3 (without speaking items) combined and Claim 2 and Claim 4 combined. In mathematics, claim scores are computed for Claim 1 on its own and for Claim 2, Claim 3, and Claim 4 combined.

#### Assessment Target Reports

##### Overview of Assessment Target Reports

Assessment target results are reported at the aggregated level for groups of 30 students or more for targets with more than 10 items in the pool. Aggregate target reports are available in the California Educator Reporting System (CERS) only within the Custom Aggregate Report feature.

Assessment target standards are specific to each content domain and linked to the CCSS associated with claim areas. For Smarter Balanced assessments, assessment targets are intended to support the development of high-quality items and tasks that contribute evidence to the claims. The relationship between assessment targets and CCSS elements is made explicit in the Smarter Balanced content specifications (ETS, 2015a, 2015b).

Assessment target scores provide insight into strengths and weaknesses for a group of students relative to their performance on the assessment. Performance relative to the entire assessment indicates strengths and weaknesses relative to the test performance as a whole. For a selected group of students (for example, a classroom), if their performance on an assessment target is better than their performance on the assessment as a whole, the assessment target is an area of relative strength. Conversely, if the group of students did not perform as well on an assessment target in relation to the assessment as a whole, it would be an area of relative weakness. Performance relative to *Standard Met* is an indicator of whether students’ performance on a target was above, near, or below *Standard* *Met* (level 3). For a selected group of students, a “Below” indicator suggests that the students have not yet mastered the content assessed in a target; however, the students’ overall performance on the assessment may be near or above standard.

Assessment target scores are derived from item *residuals*, which are the differences between a student’s observed score and expected score for a particular item. For the selected group of students, the assessment target scores for each student are calculated by summing the differences between the observed and expected scores for each student for all items that the student attempted within a particular assessment target. The sum of these differences is then divided by the total number of points possible for items within a particular target. Next, the mean assessment target scores, as well as the standard error for all students in the selected student group, are calculated. Finally, strengths and weaknesses thresholds are established after the values for each assessment target are calculated. More details on the calculation of the assessment targets and the establishment of the strengths and weaknesses thresholds are described in an ETS memorandum, *Target Score Reporting* (ETS, 2015b).

Note, however, that while assessment targets are based on target standards, not all claim areas support assessment target reporting. For example, assessment targets are reported for all claims in ELA but only for Claim 1 in mathematics.

##### Limitations

When reported, caution should be used when reporting or interpreting assessment targets. First, assessment targets can only be meaningfully reported at the group level because they are neither reliable nor generalizable enough to support inferences for individual students. Second, because residuals are sensitive to model fit, student strengths and weaknesses evaluated this way are sometimes the result of a misfit in item calibration. Therefore, it is necessary to compute the average residuals of each item across all students within each assessment target to determine whether the average residuals across all students are uniformly close to zero. Finally, assessment targets that are based on 10 or fewer items in the item bank are not reported, except the WER items.

The extent to which the scores are *generalizable* depends on the total number of items administered from that domain across all students. A small number of items is not sufficient to broadly represent the target domain or to support the general conclusions required of actionable information.

##### Reporting

The distribution of the average assessment target scores depends both on the number of students in the defined group and on the number of items that these students answered in a target. As both numbers grow large, the average residuals increasingly cluster symmetrically around zero. To support California schools in making valid inferences based on the assessment target information, the number of items per target standard is considered when reporting the assessment target. In the 2023–24 test administration, only those targets with 11 or more items in the pool are included for reporting for non-WER targets. Scores are displayed in CERS for groups with 30 students or more.

#### Theta Score Standard Errors

A student’s true ability level or theta score and standard error of theta are not known. The SEM is the SD of the distribution of theta scores that the student would earn under different testing conditions. In IRT, the only differences taken into account in the SEM are those associated with different sets of items that could be presented to the student. An error band can be calculated from the student’s theta score minus one SEM to the student’s theta score plus one SEM. Over a large number of replications of this procedure, the error band will contain the student’s true score approximately 68 percent of the time. The error band is transformed to the scale score metric and reported for the CAASPP Smarter Balanced assessments. It is useful to take into account the size of measurement errors because no assessment measures student ability with perfect accuracy or consistency. (Error bands are also discussed in subsection [*7.3.6 Error Band*](#_Error_Band).)

In the framework of IRT, the SEM is the reciprocal of the square root of the test information function (TIF) based on the items taken by each student. It is also the estimate of standard error for the estimate of theta. The TIF is the sum of information from each item on the assessment. With MLE, the SEM for a student with proficiency *θj* is calculated using equation 7.8. *Refer to the* [*Alternative Text for Equation 7.8*](#_Alternative_Text_for_6) *for a description of this equation.*

 (7.8)

where,

*I(θj)* is the test information for student *j*, calculated as presented in equation 7.9(*Refer to the* [*Alternative Text for Equation 7.9*](#_Alternative_Text_for_8) *for a description of this equation.),*

 (7.9)

and *Ii*(*θj*)is the item information of item *i* for student *j*.

When item information is based on the GPCM for both dichotomous and polytomous items, it is calculated as presented in equation 7.10. *Refer to the* [*Alternative Text for Equation 7.10*](#_Alternative_Text_for_9) *for a description of this equation.*

 (7.10)

where,

*Si*(*θj*) is the expected item score for item *i* on a theta scale score *θj*, calculated as presented in equation 7.11 (*Refer to the* [*Alternative Text for Equation 7.11*](#_Alternative_Text_for_10) *for a description of this equation.),*

 (7.11)

and equation 7.12 *(Refer to the* [*Alternative Text for Equation 7.12*](#_Alternative_Text_for_11) *for a description of this equation.),*

 (7.12)

where,

*Pih*(*θj*) is the probability of an examinee with *θj* getting score *h* on item *i*, the computation of which is shown in equation 7.1; and

*ni* is the maximum number of score points for item *i*.

The SEM is calculated only on the basis of the answered item(s) for both complete and incomplete assessments. The upper bound of the SEM is set to 2.5 on the theta metric, and any value larger than 2.5 is truncated at 2.5, as is required by the Smarter Balanced Assessment Consortium (AIR, 2015).

#### Scale Score Standard Errors

Standard errors of the maximum likelihood theta estimates are also transformed onto the reporting scale. This transformation is calculated using equation 7.13. *Refer to the* [*Alternative Text for Equation 7.13*](#_Alternative_Text_for_12) *for a description of this equation.*

 (7.13)

where,

*SEθ* is the standard error of the ability estimate on the *θ* scale, and

*a* is the slope of the scaling constants that transform *θ* to the reporting scale.

The value of *a* is 85.8 for ELA and 79.3 for mathematics.

#### Error Band

A band of scale scores showing the measurement error associated with each scale score is reported. It is generated by developing a band of indeterminacy surrounding the scale score, as presented in equation 7.14. *Refer to the* [*Alternative Text for Equation 7.14*](#_Alternative_Text_for_13) *for a description of this equation.*

 (7.14)

where,

*SS* is the scale score,

*SEscaled* is the SEM associated with this scale score,

*SS − SEscaled* is the lower boundary of the error band, and

*SS + SEscaled* is the upper boundary of the error band.

#### Lexile® and Quantile® Scores

Students received Lexile scores (Lexile reader measure), derived from the Smarter Balanced Summative Assessment for ELA; and Quantile scores, derived from the Smarter Balanced Summative Assessment for Mathematics, on the SSRs for the first time in the 2020–‍21 administration.

The Lexile score describes a student’s reading ability and also connects the student’s reading ability with books and other useful reading materials that are at the appropriate difficulty level to challenge the student’s reading and promote reading improvement (MetaMetrics, 2020a).

The Quantile score indicates how well a student understands mathematical concepts and skills at the student’s grade level and connects targeted mathematical resources that appropriately match to the student for further learning (MetaMetrics, 2020b).

MetaMetrics Inc. and the Smarter Balanced Assessment Consortium conducted linking studies in ELA and mathematics, respectively, using the classical linear equating method. Two score scales—the Smarter Balanced for ELA scale and the Lexile scale; and the Smarter Balanced for Mathematics scale and the Quantile scale—were linked using linear equating. The results of the linking studies provide information that can be used to match students’ achievement with instructional resources to identify the materials, concepts, and skills a student should be matched with for successful instruction in ELA and mathematics, given the student’s performance on a Smarter Balanced Summative Assessment.

The final linking equation between Smarter Balanced overall scale scores and Lexile measures can be written as presented in equation 7.15. *Refer to the* [*Alternative Text for Equation 7.15*](#_Alternative_Text_for_21) *for a description of this equation.*

 (7.15)

where,

the slope and intercept are constants that convert Smarter Balanced ELA scale scores to Lexile measures, and

*g* represents the test levels.

The final linking equation between Smarter Balanced overall scale scores and Quantile measures can be written as presented in equation 7.16. *Refer to the* [*Alternative Text for Equation 7.16*](#_Alternative_Text_for_22) *for a description of the equation.*

 (7.16)

where,

the slope and intercept are conversion constants that convert Smarter Balanced mathematics scale scores to Quantile measures, and

*g* represents the test levels.

The slope and intercept for equations 7.15 and 7.16 are shown in table 7.C.5 and table 7.C.6 in [appendix 7.C](#_Appendix_7.C:_Summary).

Refer to the linking studies for detailed information about these studies (MetaMetrics, 2020a, 2020b).

### Overview of Score Aggregation Procedures

To provide meaningful results to the interest holders, test scores for a given grade level and content area are aggregated at the school, LEA or direct funded charter school, county, and state levels. The aggregate scores are generated both for selected groups and for the population. The next subsection contains a description of the types of aggregation performed on Smarter Balanced Summative Assessment computer-based assessment scores. Score aggregation includes only students with valid scores; refer to subsection [*7.5.2 Special Cases*](#_Special_Cases_3) for more information.

#### Score Distributions and Summary Statistics

Summary statistics that describe student performance on each assessment that contains only operational items are presented in table 7.24.

Table 7.24 Operational Mean and SD of Theta and Scale Scores for the Standard Assessment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Number of Students** | **Mean Scale Score** | **Scale Score SD** | **Mean Theta Score** | **Theta Score SD** |
| ELA 3 | 403,574 | 2410 | 103 | -1.15 | 1.20 |
| ELA 4 | 413,479 | 2452 | 109 | -0.66 | 1.26 |
| ELA 5 | 418,413 | 2491 | 111 | -0.20 | 1.30 |
| ELA 6 | 420,389 | 2515 | 107 | 0.08 | 1.25 |
| ELA 7 | 424,312 | 2536 | 116 | 0.33 | 1.35 |
| ELA 8 | 424,695 | 2549 | 116 | 0.47 | 1.35 |
| ELA 11 | 435,289 | 2591 | 128 | 0.96 | 1.50 |
| Mathematics 3 | 400,426 | 2423 | 94 | -1.16 | 1.19 |
| Mathematics 4 | 409,812 | 2461 | 97 | -0.67 | 1.22 |
| Mathematics 5 | 414,105 | 2483 | 106 | -0.40 | 1.34 |
| Mathematics 6 | 415,324 | 2500 | 119 | -0.19 | 1.50 |
| Mathematics 7 | 418,364 | 2513 | 124 | -0.02 | 1.56 |
| Mathematics 8 | 418,313 | 2524 | 135 | 0.12 | 1.71 |
| Mathematics 11 | 428,245 | 2548 | 139 | 0.41 | 1.75 |

Summary statistics based on the operational items for students who took the version with the embedded field test PT items for mathematics are presented in table 7.25.

Table 7.25 Operational Mean and SD of Theta and Scale Scores for Students Taking the Embedded Field Test PT Form Versions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Number of Students** | **Mean Scale Score** | **Scale Score SD** | **Mean Theta Score** | **Theta Score SD** |
| Mathematics 3 | 8,143 | 2425 | 96 | -1.13 | 1.21 |
| Mathematics 4 | 8,329 | 2464 | 96 | -0.65 | 1.20 |
| Mathematics 5 | 8,407 | 2485 | 106 | -0.38 | 1.33 |
| Mathematics 6 | 8,448 | 2501 | 118 | -0.17 | 1.49 |
| Mathematics 7 | 8,547 | 2515 | 127 | 0.00 | 1.60 |
| Mathematics 8 | 8,538 | 2527 | 135 | 0.15 | 1.70 |
| Mathematics 11 | 6,458 | 2548 | 140 | 0.42 | 1.76 |

Included in the tables are the number of students for each assessment and the mean and SD of student scores expressed in terms of both scale score and theta score. The mean thetas and corresponding scale scores increase as expected over increasing grade levels across the vertical scale.[[12]](#footnote-13)

For students who took only operational items, the number and the percentage of students in each achievement level and the number and the percentage who meet or exceed the standard are shown in table 7.26 and table 7.27. Except for grade three, more test takers met or exceeded standards in ELA than in mathematics across grade levels. The percentage of students meeting or exceeding standards is in the range of 43 to 56 percent in ELA and 28 to 46 percent in mathematics in grades three through eight and grade eleven.

In both tables, the numbers in the *Standard Met/Exceeded N* and *Standard Met/‌Exceeded %* may not exactly match the sum of Level 3 and Level 4 percentages because of rounding.

Percentages of students who took only operational items in each achievement level for ELA are in table 7.26, and their graphical representation is displayed in figure 7.4, which immediately follows.

Table 7.26 Achievement Levels for Students Who Took Only Operational Items for the Smarter Balanced for ELA

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Not Met N** | **Standard Not Met %** | **Standard Nearly Met N** | **Standard Nearly Met %** | **Standard Met N** | **Standard Met %** | **Standard Exceeded N** | **Standard Exceeded %** | **Standard Met/‌Exceeded N** | **Standard Met/‌Exceeded %** |
| ELA 3 | 139,596 | 35% | 91,255 | 23% | 78,986 | 20% | 93,737 | 23% | 172,723 | 43% |
| ELA 4 | 153,070 | 37% | 78,901 | 19% | 81,425 | 20% | 100,083 | 24% | 181,508 | 44% |
| ELA 5 | 138,997 | 33% | 81,199 | 19% | 105,325 | 25% | 92,892 | 22% | 198,217 | 47% |
| ELA 6 | 125,604 | 30% | 102,143 | 24% | 116,541 | 28% | 76,101 | 18% | 192,642 | 46% |
| ELA 7 | 130,652 | 31% | 93,407 | 22% | 126,853 | 30% | 73,400 | 17% | 200,253 | 47% |
| ELA 8 | 126,286 | 30% | 103,250 | 24% | 127,400 | 30% | 67,759 | 16% | 195,159 | 46% |
| ELA 11 | 101,774 | 23% | 90,937 | 21% | 128,396 | 29% | 114,182 | 26% | 242,578 | 56% |

Figure 7.4 presents a graphical representation of the percentage of students who took only operational items in each achievement level for ELA by grade level. These are the achievement levels for ELA shown in table 7.26.

Figure 7.4 Percentage of achievement levels in ELA, operational assessments

Percentages of students who took only operational items in each achievement level for mathematics are in table 7.27, and their graphical representation is displayed in figure 7.5, which immediately follows.

Table 7.27 Achievement Levels for Students Who Took Only Operational Items for the Smarter Balanced for Mathematics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Not Met N** | **Standard Not Met %** | **Standard Nearly Met N** | **Standard Nearly Met %** | **Standard Met N** | **Standard Met %** | **Standard Exceeded N** | **Standard Exceeded %** | **Standard Met/‌Exceeded N** | **Standard Met/‌Exceeded %** |
| Mathematics 3 | 130,034 | 32% | 87,684 | 22% | 99,701 | 25% | 83,007 | 21% | 182,708 | 46% |
| Mathematics 4 | 125,385 | 31% | 115,583 | 28% | 91,076 | 22% | 77,768 | 19% | 168,844 | 41% |
| Mathematics 5 | 167,230 | 40% | 101,837 | 25% | 63,179 | 15% | 81,859 | 20% | 145,038 | 35% |
| Mathematics 6 | 168,744 | 41% | 104,065 | 25% | 65,771 | 16% | 76,744 | 18% | 142,515 | 34% |
| Mathematics 7 | 173,754 | 42% | 103,167 | 25% | 70,122 | 17% | 71,321 | 17% | 141,443 | 34% |
| Mathematics 8 | 194,782 | 47% | 91,548 | 22% | 56,035 | 13% | 75,948 | 18% | 131,983 | 32% |
| Mathematics 11 | 218,181 | 51% | 90,625 | 21% | 66,807 | 16% | 52,632 | 12% | 119,439 | 28% |

Figure 7.5 presents a graphical representation of the percentage of students at each mathematics achievement level by grade level. These are the achievement levels for mathematics shown in table 7.27.

Figure 7.5 Percentage of achievement levels in mathematics, operational assessments

For students who took an embedded field test PT, the number and percentage of students in each achievement level and the number and percentage who met or exceeded the standard are shown for mathematics in table 7.28. The percentage of students meeting or exceeding standards is in the range of 29 to 46 percent in mathematics in grades three through eight and grade eleven. The numbers in the *Standard Met/Exceeded N* and *Standard Met/‌Exceeded %* may not exactly match the sum of Level 3 and Level 4 percentages because of rounding.

Percentages of students who took the embedded field test PTs in each achievement level for mathematics are in table 7.28, and their graphical representation is displayed in figure 7.6, which immediately follows.

Table 7.28 Achievement Levels for Students Who Took Embedded Field Test–Only PTs for the Smarter Balanced for Mathematics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Not Met N** | **Standard Not Met %** | **Standard Nearly Met N** | **Standard Nearly Met %** | **Standard Met N** | **Standard Met %** | **Standard Exceeded N** | **Standard Exceeded %** | **Standard Met/‌ExceededN** | **Standard Met/‌Exceeded%** |
| Mathematics 3 | 2,673 | 33% | 1,723 | 21% | 1,957 | 24% | 1,790 | 22% | 3,747 | 46% |
| Mathematics 4 | 2,483 | 30% | 2,408 | 29% | 1,851 | 22% | 1,587 | 19% | 3,438 | 41% |
| Mathematics 5 | 3,363 | 40% | 2,126 | 25% | 1,294 | 15% | 1,624 | 19% | 2,918 | 35% |
| Mathematics 6 | 3,426 | 41% | 2,122 | 25% | 1,344 | 16% | 1,556 | 18% | 2,900 | 34% |
| Mathematics 7 | 3,580 | 42% | 1,985 | 23% | 1,404 | 16% | 1,578 | 18% | 2,982 | 35% |
| Mathematics 8 | 3,925 | 46% | 1,854 | 22% | 1,217 | 14% | 1,542 | 18% | 2,759 | 32% |
| Mathematics 11 | 3,293 | 51% | 1,311 | 20% | 1,048 | 16% | 806 | 12% | 1,854 | 29% |

Figure 7.6 presents a graphical representation of the percentage of students who took the embedded field test PTs at each mathematics achievement level by grade level. These are the achievement levels for mathematics shown in table 7.28.

Figure 7.6 Percentage of achievement levels in mathematics, embedded field test PTs

Detailed score distribution information is available in the appendices. Table 7.A.1 and table 7.A.2 in [appendix 7.A](#_Appendix_7.A:_Overall) show the estimated distributions of theta scores for each assessment. Table 7.B.1 and table 7.B.2 in [appendix 7.B](#_Appendix_7.B:_Scale) present the selected percentiles of the scale score distributions. Table 7.B.3 through table 7.B.16 present the frequency distributions of scale scores for each assessment.

Table 7.C.1 and table 7.C.2 in [appendix 7.C](#_Appendix_7.C:_Summary) show the number of items presented within each composite claim, number of students with valid scores in each composite claim, and the mean and SD of student scores expressed in terms of both scale score and theta score. For a score to be considered valid, the student must not meet any of the special cases listed in subsection [*7.5.2 Special Cases*](#_Special_Cases_3) and must have answered enough items to receive a composite claim score. The number of students in each composite claim performance level are reported in table 7.C.3 and table 7.C.4.

#### Demographic Student Group Summaries

Statistics summarizing student performance by content area and grade level for selected groups of the students who took items are provided in [appendix 7.D](#_Appendix_7.D:_Demographic): for each assessment in table 7.D.1 through table 7.D.14 and for each field test claim for mathematics in table 7.D.15 through table 7.D.21. Table 7.D.22 through table 7.D.35 present statistical summaries of student performance by content area and grade level by composite claim for select student groups. Table 7.D.36 through table 7.D.42 present the same information as table 7.D.22 through table 7.D.35 for the embedded field test for mathematics.

In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status, disability status, migrant status, foster youth status, military status, homeless status, and ethnicity by economic status.

The tables show, for each student group, the number of students with a valid scale score, scale score mean and SD, and the percentage of students in each achievement level and claim performance level.

Table 7.29 lists the demographic student groups included in this technical report. The number and the percent of students for these demographic student groups are provided in [appendix 7.D](#_Appendix_7.D:_Demographic). Note that a student’s economic status was determined by the education level of the student’s parents/guardians and whether or not the student was eligible to participate in a free or reduced-price meal program. To protect privacy when the number of students in a student group is 10 or fewer, the summary statistics at the achievement and claim level are not reported, but are replaced by “N/A.” Additional information on specific demographic student groups can be found on the Test Results for California’s Assessments website.

Table 7.29 Demographic Student Groups to Be Reported

|  |  |
| --- | --- |
| **Category** | **Student Groups** |
| **Disability Status** | * Disability * No disability |
| **Economic Status** | * Economically disadvantaged * Not economically disadvantaged |
| **English Language Fluency** | * English learner * English only * Reclassified fluent English proficient * Initial fluent English proficient * Adult English learner * To be determined * English proficiency unknown |
| **Ethnicity** | * American Indian or Alaska Native * Asian * Native Hawaiian or Other Pacific Islander * Filipino * Hispanic or Latino * Black or African American * White * Two or more races |
| **Foster Youth Status** | * Foster youth * Not foster youth |
| **Gender** | * Male * Female |
| **Homeless Status** | * Homeless * Not homeless |
| **Migrant Status** | * Eligible for the Title I Part C Migrant Program (Migrant education) * Not eligible for the Title I Part C Migrant Program (Not migrant education) |
| **Military Status** | * Armed forces family member * Not armed forces family member |

### Reports Produced and Scores for Each Report

The assessments that make up the CAASPP computer-based assessments provide results or score summaries that are reported for different purposes. The four major purposes are to

1. help facilitate conversations between parents/guardians and teachers about student performance,
2. serve as a tool to help parents/guardians and teachers work together to improve student learning,
3. help schools and LEAs identify strengths and areas that need improvement in their educational programs, and
4. provide the public and policymakers with information about student achievement.

This section provides detailed descriptions of the uses and applications of CAASPP reporting for students.

#### Online Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage the CAASPP computer-based assessments and to inform the TDS. This system uses a role-specific design to restrict access to certain tools and applications based on the user’s designated role. Specific functions of TOMS include the following:

* Manage user access privileges
* Manage test administration calendars and testing windows
* Manage student test assignments
* Manage and confirm the accuracy of students’ test settings (that is, designated supports and accommodations) prior to testing
* Generate and download various reports

In addition to TOMS, another California online reporting system was used during the 2023–‍24 administration: CERS.

TOMS communicated with CERS, which provided authorized users with interactive and cumulative online reports for ELA and mathematics at the student, school, group, and LEA levels. CERS provided educators and administrators access to student results for each administered assessment available in the reporting system.

Based on the CAASPP reporting requirements, CERS provided the summative reports containing information outlining student knowledge and skills as they became available. CERS also permitted access to individual score reports, which provided score data for each administered assessment available in the reporting system. The online aggregate reports were available to be downloaded in PDF, Excel, and comma-separated value formats.

CERS was the primary source for LEA staff to analyze CAASPP results at the LEA, school, grade, classroom, or customized group level. CERS provided these reports, which can be downloaded and used to inform instruction. LEA staff with TOMS logon credentials could enter CERS through the CAASPP & English Language Proficiency Assessments for California (ELPAC) Website to access student assessment results.

#### Special Cases

Student scores were not reported for the following cases:

* The student had a medical emergency during testing and did not complete the assessment.
* The student’s parent/guardian requested exemption from testing and did not complete the assessment.
* The student did not log on to both CAT and PT portions.
* The student logged on to both parts (PT and CAT) without any recorded answers.
* The student logged on to one part (PT or CAT).

#### Types of Score Reports

There are four categories of CAASPP reports. Reports within each category are presented in this subsection.

1. **SSR—**The SSR was the official score report for parents/‌guardians. An SSR described the student’s results.
2. **LEA student data files—**LEA student data files were available for download on demand by the LEA in TOMS to coincide with availability of the SSRs.
3. **LEA aggregations—**Aggregate data was available to view in CERS and the Test Results for California’s Assessments website.
4. **CERS reports—**CERS allows users to produce customizable reports for students individually and on the aggregate. Individual reports are based on selectable criteria that includes the school year, content area, and assessment type; these reports include the student’s scale score and error band information, reporting category, and date of the assessment. CERS aggregate reports allow customization based on such criteria as student groups, content areas, and year tested.

##### Student Score Report

The CAASPP SSR is the official score report for parents/guardians and includes the following metrics:

* Scale score for each content-area assessment reported (The ranges of scale scores for both ELA and mathematics are provided in table 7.21.)
* Achievement level for each content-area assessment reported (Smarter Balanced achievement levels for both ELA and mathematics are *Standard Exceeded*, *Standard Met*, *Standard Nearly Met*, and *Standard Not Met*.)

Scores for students who were assigned accommodations or designated supports are reported in the same way as for students without accommodations or designated supports. (Refer to section [*2.4 Fairness and Accessibility*](#_Fairness_and_Accessibility_3) for more information about accessibility resources.)

LEAs had four options for accessing and distributing SSRs to parents/guardians:

1. Accessing electronic SSRs in PDF or HTML format using a locally provided parent/guardian or student portal
2. Downloading SSRs in PDF or HTML format from TOMS and making them available electronically using a secure local method
3. Downloading SSR PDFs from TOMS, printing them, and making them available locally
4. Purchasing video SSRs and distributing them to supplement the official SSR

The LEA CAASPP coordinator could securely deliver the appropriate reports to test sites. In the case of a locally printed Smarter Balanced Summative Assessment SSR, the LEA sent the printed report(s) to the child’s parent/guardian. Smarter Balanced Summative Assessment SSRs that included individual student results were not distributed beyond the student’s school.

Scores for students who were assigned accommodations or designated supports are reported in the same way as for students who were not assigned accommodations or designated supports. Detailed information about accessibility resources is described in subsection [*5.4.1 Accessibility Resource Categories*](#_Accessibility_Resource_Categories_2).

For the 2023–24 test administration, SSRs were made available to the LEAs in multiple supported languages, including English, Spanish, Arabic, Chinese (Traditional), Filipino, Korean, and Vietnamese. An SSR in a supported language was created if the student’s primary language as reported in the California Longitudinal Pupil Achievement Data System was one of these supported languages. The LEAs that received SSRs in supported languages received one SSR in English and another in the supported language. These reports were available as PDFs for the LEA to download from TOMS.

Further information about the SSR and its interpretation is provided on the CAASPP Starting Smarter website for California assessments.

###### Access via Student or Parent Portal

LEAs had the option to provide SSRs electronically using a locally provided parent or student portal.

Amazon Web Services—with the Amazon Simple Storage Service and the Amazon Key Management Service—ensured encrypted access for parents/guardians to view a child’s electronic SSR, which was available as a PDF.

###### Access via the Test Operations Management System

The LEA CAASPP coordinator downloaded the electronic PDFs directly from TOMS and could securely deliver the appropriate reports to test sites. Optionally, the LEA could download the SSR PDF, print the SSR PDF, and then provide the printed report(s) to the child’s parent/‌guardian.

##### Local Educational Agency Student Data Files and Aggregations

The CAASPP student data files for the LEA were available for the LEA CAASPP coordinator and site CAASPP coordinator to download from TOMS.

Student scores and aggregations were also available to LEAs prior to the release of final reports via electronic reporting, using CERS. This website permitted LEAs to view preliminary results data for all assessments taken as it became available.

Current and historical aggregate results are accessible to the public on the CDE Test Results for California’s Assessments website.

##### CERS Reports

CERS was the primary source for LEA staff to analyze CAASPP results at the LEA, school, grade, classroom, or customized group level. CERS provided these reports, which can be downloaded and used to inform instruction. LEA staff with TOMS logon credentials could enter CERS through the CAASPP & ELPAC Website to access student assessment results.

There are four additional report types in addition to the Individual Student Reports:

1. **Yearly Reports** are the basic performance report. These reports summarize summative assessment performance for student populations from one or more grade levels for one or more years of available data.
2. **Longitudinal Reports** track summative assessment performance for a single student population as that population progresses through different grades. In addition to presenting tabular data, these reports include a line graph showing how the performance of the population changed from grade to grade. These reports are only available if the reporting system includes summative assessment results.
3. **Claim and Domain Reports** are yearly reports that break down performance data by claim or domain for summative assessments.
4. **Target Reports** are yearly reports that break down performance by claim and target for the Smarter Balanced Summative Assessments. Target reports are available for all ELA claims and the mathematics Concepts and Procedures claim only.

#### Score Report Applications

CAASPP computer-based summative assessment results, presented in SSRs, provided parents/guardians with information about their child’s progress. The results are one tool for increasing communication and collaboration between parents/guardians and teachers. Along with the results from the Smarter Balanced Interim Assessments, the SSR could be used by parents/‌guardians while talking with teachers about ways to improve their child’s achievement of the CCSS.

Schools could use the CAASPP computer-based summative assessment results to help make decisions about how best to support student achievement. CAASPP computer-based summative assessment results, however, should never be used as the only source of information to make important decisions about a child’s education.

CAASPP computer-based summative assessment results help schools and LEAs identify strengths and weaknesses in their instructional programs. Each year, staff from schools and LEAs examine CAASPP test results at each grade level and content area tested. Their findings are used to help determine

* the extent to which students are learning the academic standards,
* instructional areas that can be improved,
* teaching strategies that can be developed to address the needs of students, and
* decisions about how to use funds to ensure that students achieve the standards.

CAASPP computer-based summative assessments results were used to rank the academic performance of schools, compare schools with similar characteristics (for example, size and ethnic composition), identify low-performing and high-performing schools, and set yearly targets for academic progress.

#### Criteria for Interpreting Test Scores

An LEA may use CAASPP computer-based summative assessment results to help make decisions about student placement, promotion, retention, or other considerations related to student achievement. However, it is important to remember that a single assessment can provide only limited information. Other relevant information should be considered as well. It is advisable for parents/guardians to evaluate their child’s strengths and weaknesses in the relevant topics by reviewing classroom work and progress reports in addition to the child’s CAASPP computer-based summative assessment results. It is also important to note that a student’s score in a content area could vary somewhat if the student were retested.

#### Criteria for Interpreting Score Reports

The information presented in various reports must be interpreted with caution when making performance comparisons. When comparing scale score and achievement-level results, the user is limited to comparisons within a content area. The score scales for English language arts/literacy (ELA) and mathematics are not comparable to each other. The user may compare scale scores for the same content area and grade level, within a school, between schools, or between a school and its LEA, its county, or the state.

The user can make comparisons within the same grade level and content area across years. Because the CAASPP results are scaled vertically, scale scores for an assessment at one grade level may be compared to scale scores at another grade level; this allows for the comparison of the same student’s performance over time, as well as comparison of student groups at different grade levels. Thus, users may say that proficiency for a given grade level was higher or lower one year as compared with another. However, caution should be taken when comparing scale scores from different grade levels, especially nonadjacent grade levels within a content area, because the curricula are different across grade levels. Comparing scores obtained in different content areas should be avoided because the results are not on the same scale.

However, comparing scale scores from different grade levels for the CAASPP is not appropriate, because the curricula are different across grade levels and the scale scores are not vertically linked between grade levels.

For more details on the criteria for interpreting information provided on the score reports, refer to the CAASPP Starting Smarter website for California assessments or the *2023–24 CAASPP and ELPAC Scoring and Reporting Guide* (CDE, 2024), which was applicable for the 2023–24 CAASPP administration.

### References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Institutes for Research. (2015). *Smarter Balanced scoring specification, 2014–‍2015 administration: Summative and interim assessments: ELA grades 3–8, 11 and mathematics grades 3–8, 11, version 7*. Washington, DC: American Institutes for Research.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

California Department of Education. (2024). *CAASPP and ELPAC scoring and reporting guide*. Sacramento, CA: California Department of Education.

Educational Testing Service. (2015a). *Selection of Smarter Balanced field trial items for operational scoring* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service. (2015b). *Target score reporting* [Memorandum]. Sacramento, CA: Educational Testing Service.

Hambleton, R. K., & Swaminathan, H. (1985). Estimation of Ability. *Item Response Theory: Principles and Applications*, 75–99.

Keith T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–67). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

MetaMetrics. (2020a). *Linking the Smarter Balanced English language arts/literacy summative assessment with the Lexile® framework for reading*. MetaMetrics: Durham, NC.

MetaMetrics. (2020b). *Linking the Smarter Balanced mathematics summative assessment with the Quantile® framework for mathematics*. MetaMetrics: Durham, NC.

Morgan, J., Shermis, M. D., Van Deventer, L., & Vander Ark, T. (2013). *Automated Student Assessment Prize: Phase 1 & Phase 2: A case study to promote focused innovation in student writing assessment*. Seattle, WA: Getting Smart.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–76.

Raczynski, K. R., Cohen, A. S., Engelhard, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame of reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*, 301–18.

Shermis, M. D. (2014). The challenges of emulating human behavior in writing assessment. *Assessing Writing*,*22*, 91–99.

Shermis, M. D., & Hamner, B. (2013). *Contrasting state-of-the-art automated scoring of essays*. *In Handbook of automated essay evaluation* (pp. 335–68). Routledge.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait Ratings for Automated Essay Grading. *Educational and Psychological Measurement*, *62*(1), 5–18.

Smarter Balanced Assessment Consortium. (2014a). *Hand-scoring rules*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2014b). *Interpretation and use of scores and achievement levels*.Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2015). *Reporting achievement level descriptors*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium: 2013–14 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2016b). Smarter Balanced Assessment Consortium: *2014–15 technical report*.

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications, informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications, narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017i). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018a). *Mathematics CAT and performance task (PT) item specifications*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018h). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018i). *Mathematics CAT item specifications, Claim 2, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018j). *Mathematics CAT item specifications, Claim 3, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018k). *Mathematics CAT item specifications, Claim 4, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2023). *Smarter Balanced scoring specifications for summative and interim assessments*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21*, 365–89.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13.

### Accessibility Information

#### Alternative Text for Figure 7.2

In the flow chart, the first rectangle on the top left is for the response. An arrow points to the second rectangle on the right, which is “Preprocess” with the parenthetical “(Create Raw and Canonical Representations).” An arrow on the bottom points to a diamond with the question, “Scorable?” If the response is not scorable, an arrow at the bottom points to a cylinder containing “Algorithm-Based Condition Code.” If the response is scorable, an arrow points to the right, to a group of rectangles titled “Extract Features.” Labels in the rectangles under the title and representing the activities involved in extracting features, from top to bottom, are the following:

* Custom linguistic variables (approximation sign 850)
* LSA [Latent Semantic Analysis] with different weighting schemes
* N-gram detection
* Latent Dirichlet Allocation (topic modeling)

The group of Extra Features rectangles has a right arrow that points to a group of rectangles titled “Compute Component Model Scores.” Labels in the rectangles under the title and representing the activities involved in computing component model scores, from top to bottom, are the following:

* Support vector machine
* Gradient boosted trees
* Linear slash logistic regression
* Kappa regression

The group of Compute Component Model Scores rectangles has a right arrow that points to the rectangle labeled “Ensemble (Weighted Average of CMs [Component Models]).” An arrow at the bottom of this rectangle points to a cylinder beneath it, labeled “Score, Model-Based Condition Code, Error Estimate.” *(Return to figure 7.2.)*

#### Alternative Text for Figure 7.3

The flowchart begins on the top-left with a rectangle representing the response to be scored with the label “Response.” There is a right-facing arrow pointing from this rectangle to diamond #1 with the label “Scoring model?” representing a decision concerning whether a scoring model exists for the prompt that generated the response in question. A right-facing arrow from diamond #1 with the label “Yes” points to flanged rectangle #1 with the label “General prefilter” which represents the predefined process of prefiltering the response.

There is a downward-facing arrow from flanged rectangle #1 to diamond #2 with the label “Machine score?” representing a decision concerning whether the response will be machine scored. If yes, then follow the right-facing arrow from diamond #2 to flanged rectangle #2 with the label “Machine scoring” representing the predefined process of machine scoring the response. If no, then follow the left-facing arrow from diamond #2 to flanged rectangle #3 with the label “Human scoring” representing the process of a human scoring the response rather than a machine.

Diamond #1 with the label “Scoring model?” also has a downward-facing arrow leading to flanged rectangle #3 with the label “Human scoring.” Flanged rectangle #2 with the label “Machine scoring” has an arrow pointing down and to the left toward diamond #3 with the label “Low confidence?” representing a decision concerning whether the machine scoring process assigns a high or a low confidence to the score it gave to the response. If No, representing a low confidence, then follow the left-facing arrow from diamond #3 to flanged rectangle #3 with the label “Human scoring,” indicating that the machine score should be replaced by a human score. If Yes, representing a high confidence, then follow the arrow pointing down and to the left toward the cylinder with the label “Final score” representing the database of final scores.

Flanged rectangle #3 with the label “Human scoring” also has a downward-facing arrow pointing to the database of final scores. *(Return to figure 7.3*.*)*

#### Alternative Text for Equation 7.1

P sub I h of theta sub j equals the numerator exp open parenthesis the sum from v equals 1 to h of D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis and denominator 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis, if score h equals 1, 2, …., n sub i.

P sub I h of theta sub j equals 1 divided by denominator 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c D times a sub i of the quantity open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis, if score h equals 0. *(Return to* [*equation 7.1*](#EQ7_1)*.)*

#### Alternative Text for Equation 7.2

L of the union of theta sub j, U equals the natural logarithm of open parenthesis the product from I equals 1to I times the product from v equals 0 to n sub I of P sub ih of theta sub j to the u sub iv power. *(Return to* [*equation 7.2*](#EQ7_2)*.)*

#### Alternative Text for Equation 7.3

U sub i v equals 1, if the score h on polytomous item i is equal to v, 0, otherwise. *(Return to* [*equation 7.3*](#EQ7_3)*.)*

#### Alternative Text for Equation 7.4

Theta sub t equals theta sub t minus 1 minus the quantity with the numerator L prime sub t minus 1 and the denominator L double prime sub t minus 1. *(Return to* [*equation 7.4*](#EQ7_4)*.)*

#### Alternative Text for Equation 7.5

Theta sub adj equals theta sub min plus open parenthesis theta sub achieved minus theta sub min close parenthesis times PropAdj. *(Return to* [*equation 7.5*](#EQ7_5)*.)*

#### Alternative Text for Equation 7.6

ELA scale score is the sum of 2508.2 and 85.8 times theta. *(Return to* [*equation 7.6*](#EQ7_6)*.)*

#### Alternative Text for Equation 7.7

Mathematics scale score is the sum of 2514.9 and 79.3 times theta. *(Return to* [*equation 7.7*](#EQ7_7)*.)*

#### Alternative Text for Equation 7.8

SEM of Theta sub j equals 1 divided by the square root of I of theta sub j. *(Return to* [*equation 7.8*](#EQ7_8)*.)*

#### Alternative Text for Equation 7.9

I of Theta sub j equals the sum from I equals 1 to n of I sub I of Theta sub j. *(Return to* [*equation 7.9*](#EQ7_9)*.)*

#### Alternative Text for Equation 7.10

I sub i of Theta sub j equals open parenthesis D times a sub I close parenthesis squared times open bracket s sub i2 of theta sub j minus s squared sub I of theta sub j close bracket. *(Return to* [*equation 7.10*](#EQ7_10)*.)*

#### Alternative Text for Equation 7.11

S sub i of Theta sub j equals the sum from h equals 0 to n sub i of h times p sub ih of theta sub j. *(Return to* [*equation 7.11*](#EQ7_11)*.)*

#### Alternative Text for Equation 7.12

S sub i2 of Theta sub j equals the sum from h equals 0 to n sub i of h squared times p sub ih of theta sub j. *(Return to* [*equation 7.12*](#EQ7_12)*.)*

#### Alternative Text for Equation 7.13

Scale score standard error (SE sub scaled) equals a times SE sub theta sub j. *(Return to* [*equation 7.13*](#EQ7_13)*.)*

#### Alternative Text for Equation 7.14

Error band equals open parenthesis scale score (SS) minus scale score standard error (SE sub scaled), comma, scale score (SS) plus scale score standard error (SE sub scaled) close parenthesis. *(Return to* [*equation 7.14*](#EQ7_14)*.)*

#### Alternative Text for Equation 7.15

Lexile measure equals slope sub g times open parenthesis (SB ELA SS) plus intercept. *(Return to* [*equation 7.15*](#EQ7_15)*.)*

#### Alternative Text for Equation 7.16

Quantile measure equals slope sub g times open parenthesis (SB mathematics SS) plus intercept. *(Return to* [*equation 7.16*](#EQ7_16)*.)*

### Appendix 7.A: Overall Theta Score Distribution

*This content is located in a separate file.*

### Appendix 7.B: Scale Scores of Assessments

*This content is located in a separate file.*

### Appendix 7.C: Summary Statistics and Performance Levels of Composite Claims

*This content is located in a separate file.*

### Appendix 7.D: Demographic Student Group Summaries

*This content is located in a separate file.*

### Appendix 7.E: Student Completion Conditions

*This content is located in a separate file.*

## Psychometric Analyses

This chapter summarizes the item- and test-level statistics calculated for the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Summative Assessments administered during the 2023–24 administration. All analyses were conducted on data from students taking the assessments in person and remotely unless specified otherwise.

### Overview

There are five primary statistical analysis procedures presented in this chapter:

1. Item response theory (IRT) parameters
2. Omission and completion analyses
3. Conditional exposure analyses
4. Reliability analyses
5. Analyses in support of validity evidence

#### Summary of the Analyses

This chapter summarizes the item- and test-level statistics calculated for the CAASPP Smarter Balanced Online Summative Assessments administered during the 2023–24 test administration. Each of these sets of analyses is presented in the body of the text and in the listed appendices:

1. **IRT Parameters—**[Appendix 8.D](#_Appendix_8.D:_Item) presents summaries of item difficulty parameter estimates (*b*-values) and item discrimination parameter estimates (*a*-values) for all of the items in each assessment and separate summaries for each composite claim. Also presented for each assessment are conditional distributions of *a*-values and *b*-values for students at specified ability levels (scale-score intervals) and the *a*-values and *b*-‍values of all performance task (PT) items. For polytomous items, partial credit step values (*d*-values) are included.
2. **Omission and Completion Analysis—**[Appendix 8.B](#_Appendix_8.B:_Omission) shows item parameter estimate summaries for items with different omit rates. Statistics are shown for the PTs and computer adaptive test (CAT) items in each assessment. The item parameter estimates are from the field test calibrations. The purpose of these analyses is to examine whether the items with high omit rates are systematically more difficult or more discriminating than items with low omit rates. [Appendix 8.B](#_Appendix_8.B:_Omission) also shows the completion rates for each assessment. A student’s record is considered *complete* when the student answers at least one operational PT and at least 10 CAT items. Table 8.1 and table 8.2 present the minimum number of items for a complete composite claim score in the Smarter Balanced adjusted, shortened-form blueprint, the Smarter Balanced full-form blueprint, and the adjusted, shortened embedded field test form blueprint, respectively.
3. **Conditional Exposure Analyses—**[Appendix 8.C](#_Appendix_8.C:_Item) shows, for each assessment, distributions (in intervals) of item exposure frequency for all items in that assessment, for the items in each composite claim, and for items at different difficulty levels.
4. **Reliability Analyses—**The following results of the analyses are presented:

[Appendix 8.F](#_Appendix_8.F:_Reliability) presents results of the reliability analyses of overall test scores and composite claim scores for the population as a whole and for selected student groups.

Table 8.3 presents the reliability results for the population as a whole.

Table 8.4 shows the conditional standard errors of measurement (CSEMs) at achievement-level scale score cuts.

Tables in [appendix 8.G](#_Appendix_8.G:_Scale) present CSEM distributions for the total test scores.

Figure 8.G.1 through figure 8.G.14 in [appendix 8.G](#_Appendix_8.G:_Scale) present plots of CSEMs conditional on scale scores.

Table 8.5 presents the mean CSEM for each achievement level.

Tables in [appendix 8.H](#_Appendix_8.H:_Analyses) present statistics describing the accuracy and consistency of the performance classifications.

[Appendix 8.I](#_Appendix_8.I:_Interrater) shows interrater reliability statistics for the human-scored items and statistics showing the agreement of human-to-human scoring for the constructed-response items. Note that interrater reliability statistics showing the agreement of automated scoring and human scoring are included in table 7.11 through table 7.16.

1. **Analyses in Support of Validity Evidence—**Validity evidence related to the CAASPP computer-based summative assessments is discussed in section [*8.7 Validity Evidence*](#_Validity_Evidence). [Appendix 8.E](#_Appendix_8.E:_Testing) presents distributions of the time required to complete the total assessment for each content area, including both the PT and CAT portions. Table 8.9 and the tables in [appendix 8.J](#_Appendix_8.J:_Correlations) present correlations between English language arts/‌literacy (ELA) and mathematics scores calculated for all students and for demographic student groups of interest.

#### Samples Used for Analyses

Analyses were conducted on the basis of version 3 of the production data file (“P3”) received on October 7, 2024. The P3 file comprised the full CAASPP computer-based summative assessments’ data for the majority of assessments. All valid student records were used for the technical report analyses.

Students whose records were flagged as “not scored” and students who were enrolled in a different grade level than the one in which they were tested were not included. In addition, the students with paper–pencil test (PPT) scores are excluded from the technical report tables because of low volume. The number of test takers for the PPT is reported in table 11.1.

#### Test-Taking Rates

Most students enrolled in California public schools and eligible for general assessments took the Smarter Balanced Summative Assessments for ELA and mathematics in the 2023–‍24 test administration. In ELA, the testing rate ranged from 96.41 to 97.56 percent in grades three through eight and 94.17 percent in grade eleven. In mathematics, the testing rate ranged from 96.87 to 98.55 percent in grades three through eight and 93.97percent in grade eleven. Refer to [appendix 8.A](#_Alternative_Text_for_38) for the number and percentage of students who completed the Smarter Balanced for ELA and mathematics during the 2023–24 test administration.

### Omission and Completion Analyses

#### Omit Rates

If a student views an item, leaves it unanswered, and then goes on to view and answer another item, the missing response is classified as an “omit.” If the student omits an item—that is, leaves the item unanswered—and does not view additional items, the responses for the successive items are classified as “not seen.”

The adaptive section of the Smarter Balanced assessments required students to provide answers to all items on a page before moving on to the next page; therefore, the possibility of an omission would be very small, and it would only happen to one of the last three items of the PT section that were prefetched by the Test Delivery System (TDS).

For score reporting, missing responses for the machine-scorable items due to “omit” were treated as “incorrect.” Not-reached items were not included in the calculation of student scores.

The percentage of students leaving an item blank can indicate a problem with the time planned for the assessment or with some feature of the item. If students are given an adequate amount of testing time, at least 95 percent of the students should attempt to answer each item. The CAASPP computer-based summative assessments are designed to be untimed, allowing all students to respond to all items. Because there is no time limit for the assessment, a percentage of blank responses that is greater than 5 percent for any single item may be an indication of a problem with an item.

Table 8.B.1 and table 8.B.2 in [appendix 8.B](#_Appendix_8.B:_Omission) present the summary of omit rates, including the number of items in each omit rate interval, for the PT and CAT items, respectively. The tables also contain the average difficulty and discrimination for these items. As shown, the overall omit rates for both CAT and PT items across content areas and grade levels are very low, with omit rates below 0.5 percent for all but one item. This item had an omit rate less than or equal to 1 percent.

##### Rates for Dichotomous and Polytomous Items

For both dichotomous and polytomous items, examining the omit rate is useful for identifying potential problems with test features such as testing time and item or test layout. Items with high omit rates are flagged for further investigation by ETS content specialists to ensure that no issues are found with these items. Omit rates for polytomous items tend to be higher than for dichotomous items.

#### Completion Rates

Completion rates indicate the proportion of students who completed a certain number of items in either the CAT or PT portion of the assessment. A student’s record is considered incomplete if the student did not complete at least 1 PT item and at least 10 CAT items. A student’s record is considered *complete* when the student answers at least 1 PT and at least 10 CAT items.

A student’s record for a composite claim is not considered complete unless the student completed at least the specified minimum number of items for that composite claim—refer to table 8.1 and table 8.2 for the minimum number of operational items in each composite claim, for both ELA and mathematics, in the adjusted, shortened-form blueprint; in the full-form blueprint; and in the adjusted, shortened embedded field test blueprint, respectively. The percentage of students completing each assessment, each composite claim on the assessment, and each of the two parts of the assessment are presented in table 8.B.3 and table 8.B.4 in [appendix 8.B](#_Appendix_8.B:_Omission). The completion rates show a similar pattern across content areas and grade levels. The overall, PT, and non-PT item completion rate reached 100 percent. The composite claim item completion rates reached 99 percent or higher across composite claims.

Table 8.1 presents the minimum number of items a student had to answer to be included in the data used to report aggregated composite claim scores in the adjusted, shortened-form blueprints.

Table 8.1 Minimum Number of Items for a Complete Composite Claim Score in the Adjusted, Shortened-Form Blueprints

|  |  |  |  |
| --- | --- | --- | --- |
| **Content Area and Composite Claim** | **Grades 3–5** | **Grades 6–8** | **Grade 11** |
| ELA Composite Claim 1 | 12 | 14 | 14 |
| ELA Composite Claim 2 | 11 | 11 | 11 |
| Mathematics Composite Claim 1 | 10 | 9 | 11 |
| Mathematics Composite Claim 2 | 9 | 9 | 9 |

Table 8.2 presents the minimum number of items a student had to answer to be included in the data used to report aggregated composite claim scores were the assessment to include field test PT items.

Table 8.2 PT Field Test Minimum Number of Items for a Complete Composite Claim Score If Assessment Includes Field Test PT Items

|  |  |  |  |
| --- | --- | --- | --- |
| **Content Area and Composite Claim** | **Grades 3–5** | **Grades 6–8** | **Grade 11** |
| Mathematics Composite Claim 1 | 10 | 9 | 11 |
| Mathematics Composite Claim 2 | 9 | 9 | 9 |

### Conditional Exposure Rates of Items

Item exposure refers to the frequency of item administration in the student population. Items that are selected too frequently may become known to students in advance of the test administration and, as a result, fail to perform as expected. Table 8.C.1 and table 8.C.2 in [appendix 8.C](#_Appendix_8.C:_Item) show, for each assessment and for each composite claim, the number of items in five intervals of exposure, with the lowest being 1 to 99 student testing events and the highest being greater than or equal to 3,000 student testing events. These tables also show how many items were not administered.

Conditional exposure control refers to the establishment of exposure controls to be applied to the items at a specified level of difficulty (*b*-value). These controls become necessary when items at a particular level of difficulty are especially likely to be used too often. For example, it may be necessary to limit item exposure for very difficult items. Table 8.C.3 through table 8.C.16 present the same information as table 8.C.1 and table 8.C.2, computed separately for items in several intervals of difficulty.

### Item Parameter Estimation

Parameter estimates for the 2023–24 operational items were obtained mainly from the 2013–14 Smarter Balanced field test analyses, and also from the operational assessments from the 2014–15 through 2022–23 test administrations with embedded field tests.

[Appendix 8.D](#_Appendix_8.D:_Item) provides summary statistics describing the distributions of IRT item difficulty and discrimination parameter estimates at each assessment level from the field test calibration and scaling. Note that only operational items from the item pool administered as part of the CAASPP administration are included in this analysis. Summary statistics of these parameter estimates show the difficulty and discrimination of the overall assessment, as well as the difficulty and discrimination of composite claims; distributions of *b*-value and *a*-value parameter estimates provide more detail. The step parameters for all polytomous items are also provided.

For more information regarding the IRT methodology used by Smarter Balanced to form the basis for new item development, test equating, and computer adaptive testing, refer to chapter 9 of the *2013–2014 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

#### All Items

Table 8.D.1 through table 8.D.14 in [appendix 8.D](#_Appendix_8.D:_Item) present summary statistics (mean, standard deviation [SD], minimum, and maximum) of the scaled IRT *a*-values. These statistics for each assessment are presented across all items and across items in each composite claim. Table 8.D.15 through table 8.D.28 present the summary statistics for each assessment of the IRT *b*-values across all items and across items in each composite claim.

#### Computer Adaptive Test Items

Table 8.D.29 through table 8.D.42 in [appendix 8.D](#_Appendix_8.D:_Item) show the distributions of CAT item *a*-‍values across 10 intervals of the ability scale, conditional on six intervals of student ability indicated by ranges of the overall test scale score. Table 8.D.43 through table 8.D.56 present the distributions of CAT items across 16 intervals of *b*-values conditional on six intervals of overall assessment scale scores. The mode of each distribution is highlighted in gray and presented in bold text as well as being indicated with an asterisk.

#### Performance Task Items

Table 8.D.57 through table 8.D.70 in [appendix 8.D](#_Appendix_8.D:_Item) show the conditional distribution of *a*-‍values for the PT items. Table 8.D.71 through table 8.D.84 show the conditional distribution of *b*-values for the PT items. Parameter values of all PT items are presented in table 8.D.85 through table 8.D.98.

For table 8.D.29 through table 8.D.84, the scale score intervals included in the table range from the lowest 100 scale scores containing the lowest obtainable scale score (LOSS) to the highest obtainable scale score (HOSS) with increments of 100 scale score points. For example, “2100–2199” to “2600–2699” for ELA in grade three includes the LOSS of 2115 and the HOSS of 2650.

### Response Time Analyses

Testing times for each administration can be evaluated for consistency, with the expected response processes for the tasks presented to students. The length of time it takes students to take an assessment is recorded and analyzed to build a profile describing what a typical testing event looks like for each content area and grade level. In addition, variability in testing time is investigated to determine whether a student’s testing time should be viewed as unusual or irregular. It should be noted that the Smarter Balanced assessments are untimed.

In these analyses, only students who completed at least 10 CAT items and 1 PT item and had timing records are included. The students having the shortest testing time in the PT portion—1 percent of all the students taking the assessment—and the students with the shortest testing time in the CAT portion—also 1 percent of all the students taking the assessment—are removed from the analysis. The remaining testing population is partitioned into quartiles based on scale scores on the total assessment. These groupings are not the same as the achievement levels.

The descriptive statistics—that is, the number of students, mean, SD, minimum and maximum, and percentiles—of the following time variables are computed for each of the four quartile groups derived from the scale scores for each content area:

* Time required to complete the total assessment
* Time required to complete the CAT section of each assessment
* Time required to complete the PT section of each assessment

Some cases of extremely long testing time may be attributed to students with special needs taking longer to complete the assessments, or the assessment’s not being closed down properly. Therefore, mean testing times may be misleading. The medians (fiftieth percentile) are more meaningful in the interpretation of the time comparisons because medians are less impacted by the extreme values than means. The removal of the 1 percent of the student data with the shortest testing time is a modest exclusion that leaves some students with very short durations in the results for each of the assessments. Similarly, some very long durations are present in the data, which may indicate errors such as the failure to close a testing session. Therefore, the median is a better statistic than the mean for evaluating testing time information.

Table 8.E.1 and table 8.E.2 in [appendix 8.E](#_Appendix_8.E:_Testing) provide descriptive statistics for ELA and mathematics testing time, respectively. These tables include total testing time and percentile information at each ability level. The unit of testing time is minutes; for example, in table 8.E.1, the median of the testing time for the ELA grade three Q1 group is 85 minutes. At every grade level, in both ELA and mathematics, students in the lowest ability level (first quartile, Q1) have shorter median testing times than students in the other groups. The median of total testing time generally increases with ability level from Q1 to Q4 for all grade levels in ELA and the higher grade levels in mathematics. Students at the fiftieth percentile within each ability quartile spent 68 to 167 minutes on ELA assessments across all grade levels and 40 to 90 minutes on mathematics assessments across all grade levels, indicating that students spent more time in ELA than in mathematics.

Table 8.E.3 (for ELA) and table 8.E.4 (for mathematics) provide the descriptive statistics of testing time for the CAT portion and the percentile information at each ability level. The number of CAT items presented to each student is reported in table 4.C.2 in [appendix 4.C](#_Appendix_4.C:_Item).

The median of testing time in the CAT portion generally increases with ability level from Q1 to Q4 for higher grade levels in ELA and all grade levels in mathematics except for grade three. For grades three through six in ELA, students in Q4 spent slightly less time than students in Q3. Students at the fiftieth percentile within each ability quartile spent 32 to 66 minutes on the CAT portion of ELA assessments across all grade levels and 24 to 47 minutes on the CAT portion of mathematics assessments across all grade levels.

Table 8.E.5 and table 8.E.6 in [appendix 8.E](#_Appendix_8.E:_Testing) provide the descriptive statistics for ELA and mathematics testing times for each PT and the percentile information at each ability level, respectively. Overall, students in the lowest ability level (first quartile, Q1) have shorter testing times than students in the other groups. The median of the PT testing time generally increases with ability level from Q1 to Q4 for all grade levels in ELA and the higher grade levels in mathematics. Students at the fiftieth percentile within each ability quartile spent 28 to 104 minutes on the PT portion of ELA assessments across PTs and all grade levels and 14 to 45 minutes on the PT portion of mathematics assessments across PTs and all grade levels. Similar to the overall assessments, students tended to spend more time in ELA than in mathematics on the PTs.

### Reliability Analyses

The reliability for a particular group of students’ test scores is the extent to which the scores would remain consistent if those same students were retested with a parallel version of the same assessment. There are many definitions of reliability (Haertel, 2006) that have their genesis in classical test theory and a variety of methods that can be used to estimate reliability.

The general concept of reliability concerns the extent to which the test scores measure *a particular construct* consistently. The variance in the distribution of test scores—essentially, the observed differences among individuals—is partly due to differences that are consistent and partly due to differences that are not consistent. The measure of variation associated with the first kind of differences—consistent differences—is called “true variance”; this would include actual differences in students’ knowledge. The measure of variation associated with the remaining differences—those that operate essentially at random—is called “error variance.” Error variance includes a variety of underlying differences such as selections of test content, which may cause a student’s test score to be slightly higher in one evaluation and slightly lower in another. Reliability is the proportion of total variance that is due to true variance. The standard error of measurement (SEM) is a statistic that characterizes the error variance.

Reliability coefficients range from zero to one. The higher the reliability coefficient for a set of scores, the more likely individuals are to obtain very similar scores upon repeated testing occasions, if the students do not change in their level of the knowledge or skills measured by the assessment.

#### Sample for Reliability Analyses

The reliability analyses performed for CAASPP require that the sample be screened beyond the requirements listed in subsection [*8.1.2 Samples Used for Analyses*](#_Samples_Used_for). When students’ ability estimates on the overall assessment or a composite claim are lower than the lowest obtainable theta (LOT) for that assessment, they are assigned the LOSS for that assessment. When students’ ability estimates on the overall assessment or a composite claim are higher than the highest obtainable theta (HOT) for that assessment, they are assigned the HOSS for that assessment. When a student is assigned either the LOSS or HOSS, a measure of the student’s true performance is not known, as it would be lower than the LOSS or higher than the HOSS, which ultimately impacts any reliability analyses. Because of this, the reliability analyses in this section further exclude students assigned the LOSS or HOSS from the student data used for general analyses that were described at the beginning of this chapter. (Refer to subsection [*7.3.1.4 Scale Scores for the Total Assessment*](#_7.4.1.2_Scale_Scores) for the definitions of LOSS–LOT and HOSS–HOT.)

#### Reliability Measures

In a specified population of students, the reliability of test scores, *X*, is defined as the proportion of the test score variance that is attributable to true differences in student abilities. It is sometimes operationalized as the correlation between scores on two administrations of the same testing procedure, , where *X* refers to the first administration and *X′* refers to the second administration.

Reliability coefficients may range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain very similar scores if they were retested. In applied settings, the requirement of repeated administrations is impractical, and methodologies estimating reliability from relationships among student performances on items within a single test form are often used.

An IRT-based approach called marginal reliability (Green et al., 1984) can be used to estimate the reliability of the scores. The estimates of reliability coefficients reported here are for IRT-based ability estimates.

This reliability coefficient for theta estimates, , is defined on the basis of a single test administration, as shown in equation 8.1. *Refer to the* [*Alternative Text for Equation 8.1*](#_Alternative_Text_for_34) *for a description of this equation.*

 (8.1)

where,

*J* is the number of students who took the assessment,

 is the measure of variance in ability estimates, and

 is the squared CSEM (that is, error variances) for student *j* with ability estimate .

#### Standard Error of Measurement

The SEM is a measure of how much students’ scores would vary from the scores they would earn on a perfectly reliable assessment. If it were possible to compute the error of measurement for each student’s score in a large group of students, these errors of measurement would have a mean of zero. These SEMs are an indication of how much the errors of measurement affect the students’ scores. The SEM is expressed in the same units as the test score, whether the units are in raw score or scale score metric.

The SEM is the square root of the error variance in the scores (that is, the SD of the distribution of the differences between students’ observed scores and their true scores). The SEM is calculated using equation 8.2. *Refer to the* [*Alternative Text for Equation 8.2*](#_Alternative_Text_for_42) *for a description of this equation.*

 (8.2)

where,

 is the reliability estimated in equation 8.1,

 is the SD of the total test theta score, and

*A* is the slope of the scaling transformation of theta scoresto the reporting scale.

The SEM is useful in determining the confidence interval (CI) that likely captures a student’s true score. A student’s true score can be thought of as the mean of observed scores a student would earn over an infinite number of independent administrations of the assessment. Across those administrations, approximately 95 percent of the time the interval ranging from the student’s observed score minus 1.96 SEMs to the student’s observed score plus 1.96 SEMs would contain that student’s true score (Crocker & Algina, 1986). Therefore, this interval is called a 95 percent CI for the student’s true score. For example, if a student’s observed score on a given assessment equals 2440 points, and the SEM equals 23, one can be 95 percent confident that the student’s true score lies between 2395 and 2485 points (2440  45).

Table 8.3 gives the total score reliability for theta as well as the mean, SD, and SEM of both thetas and scale scores for each of the 14 assessments, along with the number of student results upon which those analyses were performed. Note that in the case of the total test reliability, the reliability is for the whole assessment on the theta score scale; it is calculated using the total test theta score of individual students. Also, student theta scores based on operational items in the embedded field test version of forms were included.

Results from the table show a reasonably high degree of reliability: 0.87 to 0.89 for ELA across grade levels and 0.86 to 0.91 for mathematics assessments across grade levels. Note that the reliability coefficients were calculated using only 24 to 26 total items in ELA for the standard form, as well as 18 to 24 items in the mathematics standard form and 18 to 20 items in the mathematics embedded field test across grade levels based on the adjusted, shortened-form blueprints this test administration year. There were 24 to 26 items in ELA for the standard form and embedded field test, as well as 20 to 24 items in mathematics for the standard form and 18 to 20 items in mathematics for the embedded field test that were used in previous years.

In table 8.3, only students who finished at least 10 CAT items and 1 PT item are included in the analysis.

Table 8.3 Summary Statistics for Scale Scores and Theta Scores, Reliability, and SEMs

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Number of Students** | **Reliability** | **Scale Score Mean** | **Scale Score SD** | **Scale Score SEM** | **Theta Score Mean** | **Theta Score SD** | **Theta Score SEM** |
| ELA 3 | 398,672 | 0.87 | 2409 | 100 | 35.80 | -1.16 | 1.17 | 0.42 |
| ELA 4 | 406,864 | 0.87 | 2450 | 104 | 38.15 | -0.68 | 1.22 | 0.44 |
| ELA 5 | 411,041 | 0.88 | 2489 | 107 | 37.50 | -0.22 | 1.25 | 0.44 |
| ELA 6 | 416,726 | 0.89 | 2515 | 104 | 35.17 | 0.08 | 1.22 | 0.41 |
| ELA 7 | 418,145 | 0.89 | 2538 | 112 | 37.63 | 0.34 | 1.30 | 0.44 |
| ELA 8 | 415,791 | 0.89 | 2553 | 111 | 37.42 | 0.52 | 1.29 | 0.44 |
| ELA 11 | 427,083 | 0.89 | 2595 | 123 | 41.36 | 1.01 | 1.43 | 0.48 |
| Mathematics 3 | 401,389 | 0.91 | 2424 | 90 | 26.97 | -1.15 | 1.13 | 0.34 |
| Mathematics 4 | 411,912 | 0.91 | 2461 | 93 | 28.14 | -0.68 | 1.17 | 0.35 |
| Mathematics 5 | 415,361 | 0.89 | 2483 | 102 | 33.25 | -0.40 | 1.28 | 0.42 |
| Mathematics 6 | 412,775 | 0.90 | 2502 | 112 | 35.14 | -0.17 | 1.41 | 0.44 |
| Mathematics 7 | 416,050 | 0.89 | 2515 | 117 | 39.11 | 0.00 | 1.47 | 0.49 |
| Mathematics 8 | 411,100 | 0.88 | 2527 | 126 | 43.66 | 0.16 | 1.59 | 0.55 |
| Mathematics 11 | 419,288 | 0.86 | 2551 | 129 | 47.73 | 0.46 | 1.63 | 0.60 |

#### Intercorrelations, Reliabilities, and Standard Errors of Measurement for Composite Claim Scores

For each assessment, theta scores and scale scores are computed for composite claims. As described in subsection [*7.1.1 Structure of the Assessments*](#_Structure_of_the) in [*Chapter 7:* *Scoring and Reporting*](#_Scoring_and_Reporting), composite claims identify the set of knowledge and skills being measured. Composite claim scores are scores on the set of items that form the basis for a composite claim.

Intercorrelations, reliability estimates, and theta-based SEMs for the composite claims are presented in table 8.F.1 through table 8.F.14 in [appendix 8.F](#_Appendix_8.F:_Reliability). The reliability estimates vary significantly across composite claims according to both the number of items and the types of content standards that are included in each composite claim. Reliability is higher for the total scores than for the composite claim scores and is inversely related to SEM, because the number of items in each composite claim was much lower than the overall assessment.

Across grade levels in ELA, composite claim 1—Reading and Listening—has reliabilities ranging from 0.70 to 0.76; and composite claim 2—Writing and Research—has reliabilities ranging from 0.77 to 0.81. Composite claim 2 shows higher reliability than composite claim 1, although ELA composite claim 1 has more items than ELA composite claim 2.

In mathematics, composite claim 1—Concepts and Procedures—has reliabilities ranging from 0.75 to 0.80 across grade levels, and composite claim 2—Problem Solving, Communicating Reasoning and Modeling and Data Analysis—has reliabilities ranging from 0.68 to 0.82 across grade levels.

#### Student Group Reliabilities and Standard Errors of Measurement

The reliabilities of the total test scores and the composite claim scores are examined for various student groups within the student population. The reliability analyses are also presented by primary ethnicity within economic status.

Reliabilities and theta-based SEMs for the total test scores and the composite claim scores are reported for each student group analysis. Table 8.F.15 through table 8.F.27 in [appendix 8.F](#_Appendix_8.F:_Reliability) present the overall test reliabilities for student groups defined by student gender, economic status, disability status, accommodations, English language fluency, primary ethnicity, migrant status, military status, homeless status, and foster youth status. Table 8.F.28 through table 8.F.33 present the reliabilities for the student groups based on primary ethnicity within economic status.

The next set of tables, table 8.F.34 through table 8.F.117, present the composite claim–level reliabilities for the student groups as follows:

* Table 8.F.34 through table 8.F.47: gender, economic status, and migrant status
* Table 8.F.48 through table 8.F.61: military status, homeless status, and foster youth status
* Table 8.F.62 through table 8.F.75: disability status and English language fluency
* Table 8.F.76 through table 8.F.89: student primary ethnicity
* Table 8.F.90 through table 8.F.117: primary ethnicity by economic status

Overall, most student groups across grade levels and content areas exhibited reasonably high reliability, at 0.85 or higher. However, several student groups had reliability lower than 0.80 in ELA, such as

* students with accommodations in all grade levels except grade eleven, and
* EL in all grade levels.

Additionally, several student groups had reliability lower than 0.80 in mathematics, such as

* students with disabilities in grades seven, eight, and eleven;
* students with accommodations in grades five, six, seven, eight, and eleven;
* EL in grades five, six, seven, eight, and eleven;
* Black or African American in grade eleven;
* migrant education in grade eleven;
* homeless in grade eleven; and
* foster youth in grades seven, eight, and eleven.

Most of those student groups exhibited moderately high reliability in the range of 0.70 to 0.79. Small sample sizes and smaller variance of the lower scores of those student groups may contribute to the lower reliability.

Note that the reliabilities are not reported for samples that comprise 10 or fewer students based on the data suppression rule. Also, in some cases, score reliabilities are not estimable and are presented in the tables as “N/A.” The reliability estimates for some of the student groups are negative because of small variation in scale scores and large CSEMs for extreme score values. These negative reliabilities and their associated SEMs also are presented as “N/A.”

#### Conditional Standard Errors of Measurement

Classical test theory assumes that the standard error of a test score is constant throughout the score range. While the assumption is probably reasonable in the mid-score ranges, it is less reasonable at the extremes of the score distribution. IRT expands the concept by providing estimates of the standard error at each score point on the distribution.

Refer to subsections [*7.3.4 Theta Score Standard Errors*](#_Theta_Score_Standard) and [*7.3.5 Scale Score Standard Errors*](#_Scale_Score_Standard) for additional information.

##### Results

Because the Smarter Balanced assessments use item pattern scoring, each response pattern can have a unique ability estimate and CSEM. Some response patterns have more uncertainty or random error associated with their ability estimates at the upper or lower ends of the reporting scale, where items administered to students may not be well aligned to a student’s true ability level. For example, if there are not enough difficult items in the item pool, a high-ability student may not be presented with difficult items on every replication of the CAT. Under these circumstances, while the student’s scale score will be high, the student’s CSEM may not be well estimated.

CSEMs vary across the *θ* scale. When an assessment has thresholds, it is important to estimate CSEMs at those thresholds. Table 8.4 presents the scale score CSEMs at the lowest score required for a student to be classified in the *Standard Nearly Met*, *Standard Met*, and *Standard Exceeded* achievement levels for each assessment.

Table 8.4 Scale Score CSEMs at Achievement-Level Thresholds

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Nearly Met Minimum Scale Score** | **Standard Nearly Met CSEM Under Adjusted Blueprint** | **Standard Met Minimum Scale Score** | **Standard Met CSEM Under Adjusted Blueprint** | **Standard Exceeded Minimum Scale Score** | **Standard Exceeded CSEM Under Adjusted Blueprint** |
| ELA 3 | 2367 | 33 | 2432 | 31 | 2490 | 31 |
| ELA 4 | 2416 | 35 | 2473 | 34 | 2533 | 34 |
| ELA 5 | 2442 | 32 | 2502 | 32 | 2582 | 35 |
| ELA 6 | 2457 | 32 | 2531 | 31 | 2618 | 34 |
| ELA 7 | 2479 | 34 | 2552 | 33 | 2649 | 35 |
| ELA 8 | 2487 | 34 | 2567 | 34 | 2668 | 35 |
| ELA 11 | 2493 | 40 | 2583 | 37 | 2682 | 38 |
| Mathematics 3 | 2381 | 25 | 2436 | 23 | 2501 | 22 |
| Mathematics 4 | 2411 | 26 | 2485 | 24 | 2549 | 23 |
| Mathematics 5 | 2455 | 32 | 2528 | 26 | 2579 | 24 |
| Mathematics 6 | 2473 | 33 | 2552 | 29 | 2610 | 27 |
| Mathematics 7 | 2484 | 37 | 2567 | 31 | 2635 | 28 |
| Mathematics 8 | 2504 | 42 | 2586 | 35 | 2653 | 29 |
| Mathematics 11 | 2543 | 42 | 2628 | 35 | 2718 | 29 |

Table 8.5 presents the average CSEMs in each achievement level by content area and grade level. The CSEMs were largest in the achievement level of *Standard Not Met* for all assessments except for ELA in grade five. The pattern of average CSEMs is similar for the assessments in each content area.

Table 8.5 Mean CSEMs for Each Achievement Level

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Content Area and Grade Level** | **Standard Not Met Under Adjusted Blueprint** | **Standard Nearly Met Under Adjusted Blueprint** | **Standard Met Under Adjusted Blueprint** | **Standard Exceeded Under Adjusted Blueprint** |
| ELA 3 | 40.70 | 31.75 | 31.00 | 33.08 |
| ELA 4 | 40.68 | 34.29 | 33.24 | 36.79 |
| ELA 5 | 37.71 | 32.00 | 33.08 | 38.73 |
| ELA 6 | 38.17 | 31.29 | 32.43 | 36.46 |
| ELA 7 | 40.31 | 32.80 | 33.44 | 38.92 |
| ELA 8 | 41.27 | 34.00 | 34.25 | 38.91 |
| ELA 11 | 47.82 | 37.57 | 36.49 | 41.26 |
| Mathematics 3 | 31.53 | 24.00 | 22.42 | 24.80 |
| Mathematics 4 | 33.56 | 24.89 | 23.05 | 25.10 |
| Mathematics 5 | 40.69 | 28.54 | 24.66 | 25.17 |
| Mathematics 6 | 45.16 | 30.71 | 27.53 | 28.52 |
| Mathematics 7 | 48.37 | 34.34 | 29.21 | 28.40 |
| Mathematics 8 | 50.36 | 38.60 | 31.58 | 29.63 |
| Mathematics 11 | 58.45 | 38.61 | 32.37 | 30.07 |

Scale score CSEM distributions are shown in table 8.G.1 through table 8.G.14 in [appendix 8.G](#_Appendix_8.G:_Scale). The plots of the CSEMs conditional on the scale scores are also presented in figure 8.G.1 through figure 8.G.14. In the figures, the vertical axis is defined as the CSEMs and the horizontal axis is designated as scale scores, which is a common metric for assessment within the same content area. Each data point represents an individual student. Typically, for fixed-form assessments, the pattern of the CSEMs tends to be U-shaped, such that the plotted values of CSEMs for the middle scale scores tend to be lower than those for extreme scale scores.

Table 8.4 and table 8.5 and figure 8.G.1 through figure 8.G.14 in [appendix 8.G](#_Appendix_8.G:_Scale) show CSEMs are smallest in the upper-middle portion of the score range, slightly larger for high scores, and much larger for low scores, getting larger as the score gets lower. This is partially due to the impact of the CAT and vertical scales, which, in relation to a fixed-form assessment, is the attenuation of the U-shaped relationship between CSEMs and scale scores.

#### Decision Classification Analyses

When an assessment uses achievement levels as the primary method to report test results, accuracy and consistency of decisions become key indicators of the quality of the assessment.

##### Methodology

The reliabilities of achievement-level classifications, which are criterion referenced, are related to the reliabilities of the test scores on which they are based; however, they are not exactly the same. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) reviewed the topic extensively. While test reliability evaluates the consistency of test scores, decision classification reliability evaluates the consistency of classification.

Decision accuracy is the extent to which students are classified in the same way as they would be if each student’s score were the average over all possible forms of the assessment (the student’s true score). Decision accuracy answers the following question: How closely does the actual classification of students, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores could somehow be known?

Decision consistency is the extent to which students are classified in the same way as they would be on the basis of a single form of an assessment other than the one for which data is available. Decision consistency answers the following question: What is the agreement between the classifications based on two nonoverlapping, equally difficult forms of the assessment?

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995). The necessary input information includes only the maximum and minimum possible scores on the assessment and the observed score distribution and the reliability coefficient for the group of students referenced by the estimates. The method was implemented by the ETS proprietary computer program RELCLASS-COMP (Version 4.14).

Reliability of classification at a threshold is estimated by combining the achievement levels above a particular threshold and combining the achievement levels below that threshold. The result is a two-by-two table indicating whether the students are above or below the threshold. The sum of the entries in the main diagonal is the number of students accurately (or consistently) classified as above or below that threshold.

Table 8.6 and table 8.7 illustrate these two-by-two contingency tables. The proportion of students being accurately classified is determined by summing across the diagonals of table 8.6. The proportion of consistently classified students is determined by summing the diagonals of table 8.7.

Table 8.6 Decision Accuracy for Reaching an Achievement Level

|  |  |  |
| --- | --- | --- |
| **Achievement Level Status** | **Does Not Reach an Achievement Level Based on True Score** | **Reaches an Achievement Level Based on True Score** |
| Does not reach an achievement level | Accurate classification | Inaccurate classification |
| Reaches an achievement level | Inaccurate classification | Accurate classification |

Table 8.7 Decision Consistency for Reaching an Achievement Level

|  |  |  |
| --- | --- | --- |
| **Achievement Level Status** | **Does Not Reach an Achievement Level Based on an Alternate Form** | **Reaches an Achievement Level Based on an Alternate Form** |
| Does not reach an achievement level | Consistent classification | Inconsistent classification |
| Reaches an achievement level | Inconsistent classification | Consistent classification |

##### Results

The results of these analyses are presented in table 8.H.1 through table 8.H.28 in [appendix 8.H](#_Appendix_8.H:_Analyses). Included are the contingency tables for both accuracy and consistency of the various achievement-level classifications. The proportion of students accurately classified is reported in table 8.H.1, table 8.H.3, table 8.H.5, table 8.H.7, table 8.H.9, table 8.H.11, and table 8.H.13. The proportion of students consistently classified is reported in table 8.H.2, table 8.H.4, table 8.H.6, table 8.H.8, table 8.H.10, table 8.H.12, and table 8.H.14.

Two aggregated proportions are reported for each table. One is the proportion for correctly or consistently classified for four levels, which is determined by summing the four numbers in the main diagonal of the tables. The other proportion is based on the combined categories, where the original four achievement levels—Standard Not Met, Standard Nearly Met, Standard Met, and Standard Exceeded—are collapsed into two: Standard Not Met and Standard Nearly Met (combining Levels 1 and 2) and Standard Met and Standard Exceeded (combining Levels 3 and 4). The estimated proportion of exact agreement is then calculated from a derived 2×2 contingency table based on these collapsed categories, using the sum of the diagonal entries where the classifications align (that is, the sum of the shaded cells in each of the tables).

Across all grade levels and content areas, for the categories of *Standard Met* and *Standard Exceeded*, the reliability of classification for accuracy ranges from 0.89 to 0.91, and the reliability of classification for consistency ranges from 0.84 to 0.87, indicating a reasonably high degree of reliability of classification for accuracy and consistency.

#### Interrater Agreement (Constructed-Response Scoring Reliability)

The interrater reliability analyses are performed on approximately 10 percent of the overall testing population, randomly selected from the total population; those students’ responses are scored by two raters. The two sets of ratings are used to compute statistics describing the consistency (or reliability) of the ratings. This interrater consistency is described in three ways:

1. Percentage agreement between two raters
2. Cohen’s Kappa
3. Quadratic-weighted kappa (QWK) coefficient

In some scoring rubrics, zero is a valid score for the responses but is not provided by a rater. Instead, a score of zero is assigned when the student attempted the writing task but did not provide a response. Responses with zero scores should not be included in the calculation of the agreement statistics for these items.

##### Agreement Statistics

###### Percentage Agreement

Percentage agreement between two raters is frequently defined as the percentage of exact score agreement and adjacent score agreement. Exact score agreement means two raters give exact same scores. Adjacent score agreement means agreement between scores that differ by just one point. The percentage of exact score agreement is a stringent criterion, which tends to decrease with an increasing number of possible item score points. The fewer the item score points, the fewer degrees of freedom on which two raters can vary, and the higher the percentage of agreement.

###### Kappa

Interrater reliability or consistency is an indicator of homogeneity and is most frequently measured using Cohen’s Kappa statistic (1960), which takes chance agreement into account. For a human-scored item with *m+1* categories (where *m* is the number of score categories of an item), one can construct an (*m+1)* × *(m+1)* rating table with scores provided by two raters, *X* and *Y*, as shown in table 8.8. Let *nst* denote the number of responses for which rater *X’s* score = *s* and rater *Y’s* score = *t,* *ns+* is the number of responses for which rater *X’s* score = *s*, *n+t* is the number of responses for which rater *Y’s* score = *t*, and *n++* is the number of all responses. An ellipsis (…) signifies that there might be more rows or columns in the table.

Table 8.8 Frequencies of Ratings

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rating** | **Y = 0** | **Y = 1** | **Y = 2** | **…** | **Y = m** |
| X = 0 | n00 | n01 | n02 | … | n0m |
| X = 1 | N10 | n11 | n12 | … | n1m |
| X = 2 | n20 | n21 | n22 | … | n2m |
| … | … | … | … | … | … |
| X = m | nm0 | nm1 | nm2 | … | nmm |

*Refer to the* [*Alternative Text for Equation 8.3*](#_Alternative_Text_for_43) *for a description of this equation.* The kappa statistic is defined as

 (8.3)

*Refer to the* [*Alternative Text for Equation 8.4*](#_Alternative_Text_for_41) *for a description of this equation.*

 (8.4)

*Refer to the* [*Alternative Text for Equation 8.5*](#_Alternative_Text_for_31) *for a description of this equation.*

 (8.5)

where,

*pobs* is the observed agreement, and

*pexp* is the expected agreement between *X* and *Y*.

When *pobs* and *pexp* agree only at the chance level, the value of kappa is 0. When the two measurements agree perfectly, the value of kappa is 1.0.

###### Quadratic-Weighted Kappa

QWK is used because kappa does not take into account the degree of disagreement between raters. It is a generalization of the simple kappa coefficient using weights to quantify the relative difference between categories. The range of the QWK is from 0.0 to 1.0, with perfect agreement being equal to 1.0.

For a human-scored item with *m+1* categories, one can construct an (*m+1)* × *(m+1)* rating table with scores provided by two raters, *X* and *Y,* as described in table 8.8. The weighted kappa coefficient is defined as presented in equation 8.6. *Refer to the* [*Alternative Text for Equation 8.6*](#_Alternative_Text_for_32) *for a description of this equation.*

 (8.6)

For QWK, the weights are calculated using equation 8.7. *Refer to the* [*Alternative Text for Equation 8.7*](#_Alternative_Text_for_33) *for a description of this equation.*

 (8.7)

##### Results of the Human‒Human Interrater Reliabilities

Interrater reliabilities were evaluated during scoring by routing a minimum of 15 percent of the responses that were most different from the training responses to expert raters and assigning them a human score. The results of interrater reliability are presented in table 8.I.1 through table 8.I.14 in [appendix 8.I](#_Appendix_8.I:_Interrater), which present the results of the interrater analyses and descriptive statistics of the ratings by the two human raters on short-answer items, including the following:

* Number of score points in each item
* Number of students for each item
* QWK
* Percentage of exact agreement
* Percentage of adjacent agreement
* Percentage of exact agreement and adjacent agreement

Table 8.I.15 through table 8.I.21 present the results of the interrater analyses on writing extended-response items. The number of items that did not meet the Smarter Balanced interrater agreement standards were flagged and are presented in table 7.3. In addition to the statistics described previously, the dimension name is also identified. Note that tables in this section represent interrater reliability for human-to-human scores. Results of interrater reliability for human-to-automated scores are included in table 7.11 through table 7.16.

Refer to [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting) of this report and the *Smarter Balanced Scoring Guide for Grades 3, 6, and 11: English/Language Arts Performance Task Full-Write Baseline Sets* (Smarter Balanced, 2014) for scoring dimensions.

#### Agreement Between Artificial Intelligence and Human Scoring

In the 2023‒24 test administration, Measurement Incorporated’s automated scoring (artificial intelligence [AI]) was evaluated using interrater reliability between AI-generated and human scores. First, an initial evaluation was conducted for each automated scoring model by comparing the automated scores to the human-scored training data, as well as comparing the performance of automated scoring to the performance of expert raters.

Second, all models associated with items that passed initial validation were subject to a secondary validation. The secondary validation used the first available 500 responses per item, at a minimum. Responses from this sample were scored by both the automated scoring engine and an expert rater. If the interrater reliability between automated scores and the scores assigned by the expert raters was sufficiently large, subsequent student responses for a given item were scored using a hybrid human–automated scoring approach. If not, the item was hand scored. Results of interrater reliability between human and automated scoring for initial validation and secondary validation are reported in table 7.11 through table 7.16. Refer to subsection [*7.2.6 Automated Scoring*](#_Automated_Scoring) for details.

### Validity Evidence

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; ETS, 2014). Concerns about validity drive the development, administration, and scoring of an assessment. Validity evidence also determines the appropriateness of test score interpretations and uses.

Validation is the process of accumulating evidence to support each proposed score interpretation or use. This validation process does not rely on a single study or gathering only one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA, APA, & NCME, 2014; Cronbach, 1971; ETS, 2014; Kane, 2006). It begins with the test design and is implicit throughout the entire assessment process, which includes item development and field testing, analyses of items, standard setting, test scaling and linking, scoring, reporting, and score usage.

In this section, the evidence gathered is presented to support the intended uses and interpretations of scores for the Smarter Balanced Summative Assessments. This section discusses some of the principles prescribed by the AERA, APA, and NCME *Standards for Educational and Psychological Testing* (2014). These *Standards* require a clear definition of the purpose of the assessment, a description of the constructs to be assessed, and the population to be assessed, as well as how the scores are to be interpreted and used.

The *Standards* identify five kinds of evidence that can provide support for score interpretations and uses:

1. Evidence based on test content
2. Evidence based on relations to other variables
3. Evidence based on response processes
4. Evidence based on internal structure
5. Evidence based on the consequences of testing

The next subsection defines the purpose of the Smarter Balanced Summative Assessments, followed by a description and discussion of different kinds of validity evidence that have been gathered.

#### Design of the CAASPP Smarter Balanced Summative Assessments

##### Purpose

The primary purpose of the CAASPP assessment system is to “assist teachers, administrators, students, and parents by promoting high-quality teaching and learning through the use of a variety of assessment approaches and item types” (California Department of Education [CDE], 2024b).

##### Constructs to Be Measured

The CAASPP computer-based summative assessments are designed to show how well students perform relative to the Smarter Balanced Assessment Consortium content standards, which are aligned to the Common Core State Standards (CCSS). These standards describe what students should know and be able to do at each grade level.

Test blueprints define the procedures used to measure the claims and standards. These blueprints, for ELA and mathematics, are provided in [appendix 4.A](#_Appendix_4.A:_Smarter). They also provide an operational definition of the construct to which each set of standards refers (Smarter Balanced, 2021a, 2021b). That is, they define, for each content area, the subject to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score student responses. The test blueprints control as many aspects of the measurement procedure as possible so that the testing conditions will remain the same over test administrations (Cronbach, 1971) to minimize construct-irrelevant score variance (Messick, 1989).

The Smarter Balanced Assessment Consortium also created the content specifications used to create the CAASPP computer-based summative assessments (Smarter Balanced, 2015a, 2015b).

##### Interpretations and Uses of the Scores

Overall student performance is expressed as scale scores and achievement levels, which are generated for both ELA and mathematics assessments, as are strength and weakness levels for each composite claim. An inference is drawn about how much knowledge and skill in the content area the student has, on the basis of a student’s total score. The total score is also used to classify students in terms of their level of knowledge and skill in the content area. These levels are called achievement levels and are labeled *Standard Exceeded*, *Standard Met*, *Standard Nearly Met*, and *Standard Not Met*.

The strength and weakness levels are used to draw inferences about a student’s achievement in each of the claims for each assessment. Starting from the 2023–24 test administration, achievement levels for the composite claim scores are reported at the individual student level. The Smarter Balanced Assessment Consortium conducted a comparability study for the adjusted versus the full blueprint. The results suggest that across all grade levels for both ELA and mathematics, the proportion of these combined claims are well preserved in the adjusted blueprints. Smarter Balanced’s plan to report combined claim scores for the adjusted blueprints aligns strongly with the full blueprints, showing successful preservation of the original content proportions and intent of the test blueprints. This likely means that these adjusted blueprints can be used without greatly disrupting the original intent of the assessments while still providing an equivalent measure of student proficiency in ELA (Smarter Balanced, 2023).

A detailed description of the uses and applications of the CAASPP computer-based summative assessment scores is presented in [chapter 7](#_Scoring_and_Reporting). Parents/Guardians have access to the Starting Smarter website, which describes CAASPP Student Score Reports and how parents/guardians can use the reports to communicate with teachers about a child’s learning (The Regents of the University of California & CDE, 2020). The information provided is available in English and Spanish. Finally, additional information can be found in the *2023–24 CAASPP and English Language Proficiency Assessments for California Scoring and Reporting Guide* (CDE, 2024a).

More detailed descriptions regarding score use can be found in the *Education Code* Section 60602 web page on the California Legislative Information website. Refer also to [*7.5 Reports Produced and Scores for Each Report*](#_Reports_Produced_and).

##### Intended Test Population

Students enrolled in grades three through eight and grade eleven are required to take part in the Smarter Balanced Summative Assessments, unless they are eligible to participate in the alternate assessments, their parents/guardians have opted them out of testing, or they were not tested because of a medical emergency. EL students who were in their first 12 months of attending school in the United States were exempt from taking the ELA portion of the assessments.

#### Content

Evidence based on test content refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as alignment methods for educational assessments that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman et al., 2002; Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009).

The degree to which the Smarter Balanced test specifications captured the CCSS and to which the items adequately represent the domains delineated in the test specifications was demonstrated in the *Alignment Study Report* (Smarter Balanced, 2016c). The major finding presented here is that the knowledge, skills, and abilities measured by the Smarter Balanced assessments are consistent with the ones specified in the CCSS. With computer-adaptive testing, an extra dimension of content validity evidence is to ensure that the item-‍selection algorithm produces forms for individual students that conform to the test blueprint. It was found that across content areas and grade levels, 98 percent or more of the simulated assessments covered the test blueprint (American Institutes for Research [AIR], 2015).

##### Description of the State Standards

As noted in section [*1.1 Background*](#_Background), the Smarter Balanced Summative Assessments are aligned with the CCSS for ELA and mathematics. The purpose of the CCSS is to provide school staff and teachers with the information and tools they need to improve teaching and learning so as to prepare all students for college and career readiness. These content standards describe what students should know and be able to do at each grade level (Smarter Balanced, 2015a, 2015b).

##### Item Specifications

Item specifications describe the characteristics of items that are written to measure each content standard. Specifications were developed for each target, within each composite claim, and at each grade level, and are published by the Smarter Balanced Assessment Consortium for ELA (Smarter Balanced, 2017a through 2017i) and mathematics (Smarter Balanced, 2018a through 2018j).

##### Item Selection Algorithm

The item selection algorithm is designed to cover a standards-based blueprint in the assembly of CAT forms. The general item selection approach is based on an item selection algorithm (refer to [*Chapter 4: Test Assembly*](#_Test_Assembly)) that evaluates an item’s contribution to each of the following measures:

1. A measure of content match to the blueprint
2. A measure of overall test information
3. Measures of test information for each reporting category on the assessment

Details can be found in the *Smarter Balanced Adaptive Item Selection Algorithm Design Report* (AIR, 2014).

##### Assessment Blueprints

The Smarter Balanced summative test blueprints provided in [appendix 4.A](#_Appendix_4.A:_Smarter) describe the content of the ELA and mathematics summative assessments for all grade levels tested and how that content is assessed. The summative computer-based test blueprints reflect the depth and breadth of the performance expectations of the CCSS. The test blueprints have information about the number of items and depth of knowledge for items associated with each assessment target (Smarter Balanced, 2021a, 2021b). Each assessment is described by a single blueprint for each segment of the assessment and identifies the order in which the segments appear.

There are two sets of blueprints for ELA and mathematics available in [appendix 4.A](#_Appendix_4.A:_Smarter). The first set, the adjusted, shortened-form blueprints, was used to create the current computer-based assessments. The second set, the full-form blueprints, was used for the computer-based assessment up until the 2018–19 test administration and for the current PPTs. The Smarter Balanced Assessment Consortium conducted a comparability study for the adjusted blueprint versus the full blueprint. The results revealed that the adjusted and full blueprints are evaluating the same depth of content, keeping the same proportions, and largely using the same distributions of CAT and PT items. This alignment demonstrates the consistent measurement approach employed across different blueprint forms and the capacity of these assessments to capture student performance reliably and accurately (Smarter Balanced, 2023).

##### Item Development Process

A detailed description of the content and psychometric criteria applicable to the construction of the Smarter Balanced item pool is included in *Chapter 4: Test Assembly*, for overall content validity, and *Chapter 3: Item Development*, for item development, of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

##### Alignment Study

A strong alignment between the CCSS and assessments is fundamental to the meaningful measurement of student achievement and instructional effectiveness. Alignment results demonstrate that the assessments represent the full range of the content standards and that these assessments measure student knowledge in the same manner and at the same level of complexity as expected in the content standards. For example, across all grade levels, 64.7 percent of the items are identified in alignment with the ELA grade-level CCSS and 76.7 percent of the items are identified in alignment with the mathematics grade-level CCSS by at least 50 percent of the reviewers (Smarter Balanced, 2016c).

##### Form Assembly Process

The content standards, blueprints, and item-selection algorithm are the basis for choosing items for each assessment. Additional item difficulty and discrimination targets are defined in light of what are desirable statistical characteristics in test items and statistical evaluations. Refer to [*Chapter 4: Test Assembly*](#_Test_Assembly) for additional information.

##### Simulation Study

Simulations are conducted to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm and the scoring algorithm. The simulation tool allows for the manipulation of key blueprint and configuration settings to match the blueprint and minimize measurement error. The report *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm* contains more information about the algorithms used (AIR, 2015).

The findings from the 2016–17 simulation study demonstrate that the Smarter Balanced adaptive TDS administers assessments with items representing the breadth and depth identified in the test specifications and content standards, and that scores are comparable with respect to the targeted content and are measured with good precision across the range of proficiency. Refer to *Smarter Balanced Summative Assessments Simulation Results* for detailed information (AIR, 2016)*.*

#### Response Processes

Validity evidence based on response processes refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers” (AERA et al., 2014, p. 15). This type of evidence generally includes documentation of activities such as

* interviews with students concerning their responses to test items (that is, think alouds);
* systematic observations of test response behavior;
* evaluation of the criteria used by judges when scoring PTs, analysis of student item response-time data, and features scored by automated algorithms; and
* evaluation of the reasoning processes students employ when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009).

This type of evidence is used to confirm that the Smarter Balanced assessments are measuring the cognitive skills that are intended to be the objects of measurement and that students are using these targeted skills to respond to the items.

##### Think Alouds

One way to evaluate response process is through think-aloud protocols (Lewis, 1982). Think-aloud protocols were conducted early in the development of the Smarter Balanced assessments and were described by Smarter Balanced (2015a) in the following way:

“Using the revised item and task specifications, a small set of items was developed and administered in fall 2012 during a small-scale trial. This provided the Consortium with their first opportunity to administer and score the new item types. During the small-scale trials, the Consortium also conducted cognitive laboratories to better understand how students respond to various types of items. The cognitive laboratories used a think-aloud methodology in which students speak their thoughts while working on a test item. The item and task specifications were again revised based on the findings of the cognitive laboratories and the small-scale trial. These revised specifications were used to develop items for the 2013 pilot test, and they were again revised based on 2013 pilot test results and subsequent review by content experts.”

##### Testing Time Analysis

Testing times for each administration can be evaluated for consistency, with the expected response processes for the tasks presented to students. The length of time it takes students to take an assessment is recorded and analyzed to build a profile describing what a typical testing event looks like for each content area and grade level. In addition, variability in testing time is investigated to determine whether a student’s testing time should be viewed as unusual or irregular. It should be noted that the Smarter Balanced assessments are untimed.

In these analyses, only students who completed at least 10 CAT items and 1 PT item and had timing records are included. The students having the shortest testing time in the PT portion—1 percent of all the students taking the assessment—and the students with the shortest testing time in the CAT portion—also 1 percent of all the students taking the assessment—are removed from the analysis. The remaining testing population is partitioned into quartiles based on scale scores on the total assessment. These groupings are not the same as the achievement levels.

The descriptive statistics—for example, the number of students, mean, SD, minimum and maximum, and percentiles—of the following time variables are computed for each of the four quartile groups derived from the scale scores for each content area:

* Time required to complete the total assessment
* Time required to complete the CAT section of each assessment
* Time required to complete the PT section of each assessment

Some cases of extremely long testing time may be attributed to students with special needs taking longer to complete the assessments, or the assessment’s not being closed down properly. Therefore, mean testing times may be misleading.

The medians (fiftieth percentile) are more meaningful in the interpretation of the time comparisons because medians are less impacted by the extreme values than means. The removal of the 1 percent of the student data with the shortest testing time is a modest exclusion that leaves some students with very short durations in the results for each of the assessments. Similarly, some very long durations are present in the data, which may indicate errors such as the failure to close a testing session. Therefore, the median is a better statistic than the mean for evaluating testing time information. Refer to section [*8.5 Response Time Analyses*](#_Testing_Time_Analyses) and [appendix 8.E](#_Appendix_8.E:_Testing) for detailed testing time analyses and results.

#### Internal Structure

Validity evidence based on *internal structure* refers to the statistical analysis of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include factor analysis—both exploratory and confirmatory—or multidimensional IRT scaling. With a vertical scale, a consistent primary dimension across the levels of the assessment should be maintained.

##### Dimensionality

A dimensionality study was conducted during the pilot test phase to determine the factor structure of the assessments and the types of scales developed, as well as the associated IRT models used to calibrate them. In part, that study used the Akaike Information Criterion (Akaike, 1973) to evaluate the fit of potential multidimensional models relative to the unidimensional model. The results suggested that the unidimensional model fit better than the multidimensional model, once model complexity was taken into account. More detailed results for the Smarter Balanced pilot test are available in the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a).

##### Differential Item Functioning

Analysis of item functioning using IRT and differential item functioning (DIF) falls under the internal structure category. For Smarter Balanced assessments, DIF analyses were conducted to assess differences in the item performance of groups of students who differ in their demographic characteristics. DIF analyses were implemented during the pilot test and field test phases when the assessments were delivered in linear fixed-length forms (Smarter Balanced, 2016a, chapter 6; Smarter Balanced, 2016b, chapter 8). For both ELA and mathematics, few items were identified as having significant levels of DIF. In the operational assessment, by virtue of the CAT delivery, non-embedded field test items are not amenable to DIF analyses.

##### Overall Reliability Estimates

The results of reliability analyses on the total test theta scores on each summative assessment are presented in table 8.3. The results indicate that the reliability estimates for all summative assessment total scores are reasonably high, given that there are only 24 to 26 total items for ELA, 18 to 24 items for mathematics, and 18 to 20 for the mathematics embedded field test under the adjusted test blueprint. Theta score SDs and SEMs are increasing with grade level; this is often an artifact of vertical scaling.

##### Composite Claim Reliability Estimates

For each CAASPP Smarter Balanced computer-based summative assessment, theta scores are computed for composite claims. The reliability estimates of these scores are presented in table 8.F.1 through table 8.F.14 in [appendix 8.F](#_Appendix_8.F:_Reliability). The reliability estimates of composite claims are invariably lower than those for the total assessments because they are based on fewer items. Because the reliabilities of scores at the composite claim level are lower than for total scores, and because each composite claim contains a different number of items, educators should supplement the score results with other information when interpreting composite claim scores.

##### Student Group Reliability Estimates

The reliabilities also are examined for various student groups that differ in their demographic characteristics within the student population. The characteristics considered are gender, ethnicity, economic status, accommodations, disability status, migrant status, English language fluency, military status, homeless status, foster youth status, and ethnicity by economic status (refer to table 7.29 for the demographic student groups reported). Reliability estimates and SEM information for the total test theta scores and the composite claim theta scores are reported for each student group. Table 8.F.15 through table 8.F.33 in [appendix 8.F](#_Appendix_8.F:_Reliability) present the reliabilities and SEMs on the overall test theta scores for the various student groups. Table 8.F.34 through table 8.F.117 present the reliabilities and SEMs of theta scores for the composite claims.

Overall, most student groups across grade levels and content areas exhibited reasonably high reliability at 0.85 or higher. The exceptions are noted in subsection [8.6.5](#_Student_Group_Reliabilities).

##### Reliability of Performance Classifications

The methodology used for estimating the reliability of classification decisions is described with the decision classification analyses in subsection [*8.6.7 Decision Classification Analyses*](#_Decision_Classification_Analyses). The results of these analyses are presented in table 8.H.1 through table 8.H.28 in [appendix 8.H](#_Appendix_8.H:_Analyses). When the classifications are collapsed to *Standard Not Met and Standard Nearly Met* versus *Standard Met and Standard Exceeded*, which are the critical categories for accountability analyses, the estimated proportion of students who are classified accurately ranges from 0.89 to 0.91 across all assessments. Similarly, the estimated proportion of students who are classified consistently ranges from 0.84 to 0.87 for students classified into *Standard Not Met and Standard Nearly Met* versus *Standard Met and Standard Exceeded*. These are considered high levels of accuracy and consistency.

##### Interrater Reliability

Interrater reliability is mainly measured by Cohen’s Kappa or the QWK. Those statistics provide evidence of the degree to which a student’s score is consistent from one rater to another. Research has shown values of kappa between 0.41 and 0.60 exhibit moderate levels of agreement between the two ratings (Landis & Koch, 1977; Flack et al., 1988) and that values of QWK greater than 0.70 indicate excellent agreement (Williamson, Xi, & Breyer, 2012). In addition, rater agreement is measured by the percentage of exact agreement, percentage of adjacent agreement, or percentage of the combination of exact agreement and adjacent agreement. Those statistics, including the QWK, percentage of exact agreement, and percentage of combined adjacent agreement and exact agreement, are presented for human–automated scoring and human‒human scoring, respectively.

The human–human interrater reliability results are presented in table 8.I.1 through table 8.I.21 in [appendix 8.I](#_Appendix_8.I:_Interrater). The human–automated scoring interrater reliability results are presented in table 7.11 through table 7.16.

##### Correlations Between the Composite Claims within Content Areas

The distinctiveness and reliability of the composite claim theta scores in each content area are important because CAASPP strength and weakness levels are reported on the basis of composite claim scores. The interrelationships of composite claim scores should be shown to be consistent with the construct being assessed. The correlations between the composite claims within content areas are presented in table 8.F.1 through table 8.F.14.

##### Correlations Between Content-Area Test Scores

The degree to which students’ content-area test scores correlate as expected provides evidence of those scores as measures of the intended constructs. Table 8.9 provides the correlations between scores on the 2023–24 CAASPP ELA and mathematics assessments and the number of students on whom these correlations are based. Results are based on all students with valid scale scores and are provided by grade level.

Table 8.9 Correlations for All Students

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade Level** | **ELA Sample Size** | **Mathematics Sample Size** | **Correlation Sample Size** | **Correlation** |
| 3 | 403,574 | 408,569 | 402,819 | 0.78 |
| 4 | 413,479 | 418,141 | 412,638 | 0.79 |
| 5 | 418,413 | 422,512 | 417,483 | 0.79 |
| 6 | 420,389 | 423,772 | 418,981 | 0.79 |
| 7 | 424,312 | 426,911 | 421,936 | 0.78 |
| 8 | 424,695 | 426,851 | 422,108 | 0.78 |
| 11 | 435,289 | 434,703 | 430,022 | 0.74 |

Results for these students appear to be consistent with expectations. In general, students’ ELA scores correlated moderately with their mathematics scores. They are correlated slightly higher among students in grades three through eight than in grade eleven.

Table 8.J.1 through table 8.J.9 in [appendix 8.J](#_Appendix_8.J:_Correlations) provide the content-area assessment score correlations by gender, ethnicity, English language fluency, economic status, disability status, migrant status, military status, homeless status, and foster youth status. The correlation between students’ ELA and mathematics scores was approximately in the range of 0.60 to 0.84 at all grade levels for nearly all the student groups. One exception was the group composed of EL students, who showed lower correlations at grades five through eight and grade eleven in the range of 0.42 to 0.59. Note that the student group in grade eleven comprising adult ELs showed the lowest correlation, at 0.37 with a sample of 94 students.

Correlations are reported only for groups of more than 10 students. Correlations between scores on any two content-area assessments where 10 or fewer students took the assessments are expressed as “N/A.”

#### Relations to Other Variables

Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait–multimethod studies (Campbell & Fiske, 1959). External variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (for example, test scores) to evaluate the degree to which different assessments actually measure different skills and the utility of test scores for predicting specific criteria (for example, college grades). This type of evidence is essential for supporting the validity of certain inferences based on scores from the Smarter Balanced assessments for certifying college and career readiness, which are the primary test purposes.

A subset of students who took National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) items also took Smarter Balanced CAT items and PTs. A summary of the resulting item performance for NAEP, PISA, and all Smarter Balanced items can be found in chapters 7 and 8 of the *2013–14 Smarter Balanced Technical Report* (Smarter Balanced, 2016a). That study found item-level performance to be similar for NAEP and Smarter Balanced populations. A study relating Smarter Balanced scales to NAEP or PISA scales has not been made.

Another study established the relationship between Smarter Balanced field test scores and the likelihood of achieving “Conditionally Exempt” status based on achieving the required minimum scores for the California State University Early Assessment Program (EAP). During the 2013–14 administration, students in grade eleven took the EAP for ELA, mathematics, or both. The comparison showed a correlation of 0.68 between Smarter Balanced ELA and EAP ELA assessments and correlations from 0.49 to 0.61 between Smarter Balanced mathematics and EAP mathematics assessments (ETS, 2015a, 2015b, 2015c). These correlations indicate that Smarter Balanced Summative Assessments might be measuring different aspects of college readiness than the EAP assessments, which previously provided insight into the readiness of California students in grade eleven for college-level mathematics and ELA courses. Other predictive validity research is being pursued by the Smarter Balanced Assessment Consortium as part of their research agenda.

#### Consequences of Testing

Evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing on issues such as high school dropout rates. With respect to educational assessments, the *Standards* stress the importance of evaluating test consequences. For example, they state the following:

When educational testing programs are mandated…the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user. (AERA et al., 2014, p. 195)

Investigations of testing consequences relevant to the Smarter Balanced goals include analyses of students’ opportunity to learn the CCSS and analyses of changes in textbooks and instructional approaches. Unintended consequences, such as changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging can be evaluated. These sorts of investigations require information beyond what has been available to the CAASPP program to date. Refer to the *Smarter Balanced Assessment Consortium: 2017–18 Technical Report* (Smarter Balanced, 2019) for more validity evidence.

### References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), Proceedings from *2nd International Symposium Information Theory* (pp. 267–81). Budapest, Hungary: Akademia Kiado.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Institutes for Research. (2014). *Smarter Balanced adaptive item selection algorithm design report*. Washington, DC: American Institutes for Research.

American Institutes for Research. (2015). *Smarter Balanced Summative Assessments testing procedures for adaptive item-selection algorithm*. Washington, DC: American Institutes for Research.

American Institutes for Research. (2016). *Smarter Balanced Summative Assessments simulation results*. Washington, DC: American Institutes for Research.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states’ content standards: Methods and issues. *Educational Measurement: Issues and Practice*, *22*, 21–‍29.

California Department of Education. (2024a). *CAASPP and ELPAC scoring and reporting guide*. Sacramento, CA: California Department of Education.

California Department of Education. (2024b). *California Assessment of Student Performance and Progress (CAASPP) System*. California Department of Education website.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, *2*, 179–94.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2015a). *Linking study between Smarter Balanced ELA field test and California State University (CSU) English Placement Test* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service. (2015b). *Linking study between Smarter Balanced ELA field test and CSU entry-level mathematics test* [Memorandum]. Sacramento, CA: Educational Testing Service.

Educational Testing Service. (2015c). *Study of the relationship between the Early Assessment Program and the Smarter Balanced field tests*. Sacramento, CA: Educational Testing Service.

Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–97.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition) (pp. 105–46). New York: Macmillan.

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two rater Kappa statistics. *Psychometrika*, *53*(3), 321–25.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*,18, 519–32.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347–60.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Washington, DC: American Council on Education and National Council on Measurement in Education.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Landis, J. R., & Koch, G. G. (1977). The measurement of interrater agreement for categorical data. *Biometrics*, *33*, 159–74.

Lewis, C. H. (1982). *Using the “thinking aloud” method in cognitive interface design* [Technical report]. IBM. RC-9265.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32*, 179–97.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research*, *4*, 1332–61.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. *CRESST Report 752*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* [Technical Report 566]. Washington, DC: Center for the Study of Evaluation.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*, 299–321.

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced scoring guide for grades 3, 6, and 11 English/language arts performance task full-write baseline sets*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2015a). *Content specifications for the summative assessment of the Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2015b). *Content specifications for the summative assessment of the Common Core State Standards for mathematics*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2016a). *Smarter Balanced Assessment Consortium:* *2013–2014 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2016b). *Smarter Balanced Assessment Consortium: 2014–2015 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2016c). *Smarter Balanced Assessment Consortium: Alignment study report*. Alexandria, VA: Human Resource Research Organization.

Smarter Balanced Assessment Consortium. (2017a). *ELA CAT item specifications, grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017b). *ELA CAT item specifications, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017c). *ELA CAT item specifications, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017d). *ELA PT item specifications, argumentative, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017e). *ELA PT item specifications, explanatory, grades six through eight and grade eleven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017f). *ELA PT item specifications, informative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017g). *ELA PT item specifications; narrative, grades six through eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017h). *ELA PT item specifications, narrative, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2017i). *ELA PT item specifications, opinion, grades three through five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018a). *Mathematics CAT item specifications, Claim 1, grade eight*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018b). *Mathematics CAT item specifications, Claim 1, grade five*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018c). *Mathematics CAT item specifications, Claim 1, grade four*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018d). *Mathematics CAT item specifications, Claim 1, grade seven*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018e). *Mathematics CAT item specifications, Claim 1, grade six*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018f). *Mathematics CAT item specifications, Claim 1, grade three*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018g). *Mathematics CAT item specifications, Claim 1, high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018h). *Mathematics CAT item specifications, Claim 2, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018i). *Mathematics CAT item specifications, Claim 3, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2018j). *Mathematics CAT item specifications, Claim 4, grades three through eight and high school*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: 2017–2018 technical report*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2021a). *ELA/Literacy adjusted form summative assessment blueprint*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2021b). *Mathematics adjusted form summative assessment blueprint*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2023). *Comparability of the adjusted versus full blueprint.* Los Angeles, CA: Smarter Balanced Assessment Consortium.

The Regents of the University of California & California Department of Education. (2020). *Starting smarter*.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13.

### Accessibility Information

#### Alternative Text for Equation 8.1

rho sub theta-hat prime equals 1 minus the fraction with the numerator sum from j equals 1 to J of CSEM squared sub theta-hat sub j divided by the denominator J times s squared sub theta-hat. *(Return to* [*equation 8.1*](#EQ8_14)*.)*

#### Alternative Text for Equation 8.2

SEM sub scaled equals A times s sub theta-hat times the square root of 1 minus rho sub theta-hat prime. *(Return to* [*equation 8.2*](#EQ8_17)*.)*

#### Alternative Text for Equation 8.3

kappa equals the fraction with the numerator p sub obs minus p sub exp the denominator 1 minus p sub exp. *(Return to* [*equation 8.3*](#EQ8_34)*.)*

#### Alternative Text for Equation 8.4

P sub obs equals 1 divided by n times the sum from s equals 0 to m n sub ss. *(Return to* [*equation 8.4*](#EQ8_35)*.)*

#### Alternative Text for Equation 8.5

P sub exp equals 1 divided by n square times the sum from s equals 0 to m n sub s plus times n sub plus s. *(Return to* [*equation 8.5*](#EQ8_36)*.)*

#### Alternative Text for Equation 8.6

K sub ij equals open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub ij divided by n sub plus plus close parenthesis minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub I plus times n sub plus j divided by n squared sub plus plus close parenthesis divided open parenthesis 1 minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub i plus times n sub plus j divided by n squared sub plus plus close parenthesis close parenthesis, K sub ij equals open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub ij divided by n sub plus plus close parenthesis minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub i plus times n sub plus j divided by n squared sub plus plus close parenthesis divided open parenthesis 1 minus open parenthesis the sum from i equals zero to m the sum from j equals zero to m of w sub ij times n sub i plus times n sub plus j divided by n squared sub plus plus close parenthesis close parenthesis. *(Return to* [*equation 8.6*](#EQ8_37)*.)*

#### Alternative Text for Equation 8.7

W sub ij equals 1 minus open parenthesis I minus j close parenthesis squared divided by m squared. *(Return to* [*equation 8.7*](#EQ8_38)*.)*

### Appendix 8.A Test-Taking Rates

*This content is located in a separate file.*

### Appendix 8.B: Omission and Completion Analyses

*This content is located in a separate file.*

### Appendix 8.C: Item Exposure

*This content is located in a separate file.*

### Appendix 8.D: Item Response Theory Parameter Estimates

*This content is located in a separate file.*

### Appendix 8.E: Response Time Analyses

*This content is located in a separate file.*

### Appendix 8.F: Reliability Analyses

*This content is located in a separate file.*

### Appendix 8.G: Scale Score Conditional Standard Error of Measurement Distribution

*This content is located in a separate file.*

### Appendix 8.H: Analyses of Classification

*This content is located in a separate file.*

### Appendix 8.I: Interrater Reliability

*This content is located in a separate file.*

### Appendix 8.J: Correlations Between Content Areas

*This content is located in a separate file.*

## Quality-Control Procedures

The California Department of Education (CDE) and ETS implemented rigorous quality-control procedures throughout the assessment development, administration, scoring, analyses, and reporting processes for the Smarter Balanced Summative Assessments. As part of this effort, ETS staff worked with its Office of Professional Standards Compliance, which publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2014). These *Standards* support the goals of delivering technically sound, fair, and useful products and services; and assisting the public and auditors evaluating those products and services. Quality-control procedures are outlined in this chapter.

### Quality Control of Test Materials

Brief descriptions of the types of materials used for and during testing appear in the following subsections.

#### Developing Assessments

##### Computer-based Assessments

The steps taken to develop and ensure the quality of the computer-based assessments are described in [*Chapter 3: Item Development*](#_Item_Development_1) and [*Chapter 4: Test Assembly*.](#_Test_Assembly)

##### Paper–Pencil Forms

Test forms and response booklets for paper–pencil testing were developed and reviewed by ETS staff to ensure that materials met quality standards. Each document was reviewed for accuracy, completeness, and alignment with supporting materials. Print-ready PDFs underwent a stringent quality-control process to ensure that there was adequate space for student response.

#### Test Administration Manuals

ETS staff verified that test instruction manuals accurately matched the test materials and testing processes. Editors reviewed each document for spelling, grammar, accuracy, and adherence to CDE style. Each document was approved by the CDE before being published to the California Assessment of Student Performance and Progress (CAASPP) & English Language Proficiency Assessments for California (ELPAC) Website. Only nonsecure documents were posted to this website. Secure materials, such as the *CAASPP Directions for Administration*, were made available to designated local educational agency (LEA) staff through the Test Operations Management System (TOMS), which required a secure logon.

The manuals used in the administration of the CAASPP are listed in subsection [*5.2.4 Instructions for Test Administration*](#_Instructions_for_Test_2).

#### Collecting Test Materials

ETS processes ensure the security of assessments delivered using a variety of test modes and delivery methods.

##### Computer-based Assessments

During the 2023–24 CAASPP administration, there were no test materials to be collected as a result of computer-based testing.

##### Paper–Pencil Forms

Once the paper–pencil tests (PPTs) were administered at test sites whose LEAs had received prior approval from the CDE, LEAs were instructed to enter student responses into the CAASPP Data Entry Interface (DEI). LEAs were provided with instructions for secure destruction of materials in the *CAASPP Paper–Pencil Testing Test Administration Manual* (CDE, 2024).

#### Processing Test Materials

The ways in which materials associated with student testing were processed are described in subsequent subsections.

##### Computer-based Assessments

Computer-based assessments submitted by students were transmitted from Cambium Assessment, Inc. (CAI) to ETS each day. Each system checked for the completeness of the student record and stopped records that were identified as having an error. (For example, the system would identify a test part that was missing a content registration ID, which is a unique identifier that matches the student’s available test opportunities.)

Test responses were separated for human scoring between ETS and Measurement Incorporated (MI), and the reader’s ratings were delivered to ETS scoring systems for merging with machine-scored items, final scoring, and scoring quality checks.

##### Paper–Pencil Forms

Once student responses were entered into the DEI, and within five working days after the last day of each test administration period, LEAs returned all scorable and nonscorable materials. The LEAs packed all materials into cartons, applied the labels, and then numbered the cartons prior to returning the materials to ETS by means of their assigned carrier.

### Quality Control of Test Administration

The quality of test administration for the Smarter Balanced Summative Assessments was monitored and controlled through several strategies.

A fully supported Outreach team that includes California Technical Assistance Center phone support and Success Agents supported all LEAs in the administration of the CAASPP. In addition to providing guidance and answering questions, the Outreach team regularly conducted campaigns on particular administration topics to ensure all LEAs understood correct test administration procedures. Outreach was guided by individuals who managed communications to LEAs; provided regional and web-based trainings; and hosted a website, [the](https://www.caaspp.org/) CAASPP & ELPAC Website, that housed a full range of manuals, videos, and other instructional and support materials.

The quality of test administration was further managed through comprehensive rules and guidelines for maintaining the security and standardization of the CAASPP. LEAs received training on these topics and were provided tools for reporting security incidents and resolving testing discrepancies for specific testing sessions.

The ETS Office of Testing Integrity (OTI) reinforced the quality-control procedures for test administration, providing quality assurance (QA) services for all testing programs managed by ETS. The detailed procedures the OTI developed and applied in quality control are described in subsection [*5.6.1 The ETS Office of Testing Integrity*](#_The_ETS_Office).

### Quality Control of Scoring

ETS conforms to high standards of quality and fairness when scoring assessments and reporting scores. These standards dictate that ETS provides accurate and understandable assessment results to the intended recipients. It is also the ETS mission to provide appropriate guidelines for score interpretation and cautions about the limitations in the meaning and use of the test scores. Finally, ETS conducts analyses needed to ensure that the assessments are equitable for various demographic student groups.

#### Human Scoring

##### Quality Control in the Scoring Process

In general, MI allowed raters to score all items within their baseline group during a shift. The items scored were dependent on which responses had the earliest due dates.

Some mathematics performance task (PT) items had scoring dependencies, which means that students based their calculations and responses on the answers to previous items associated with the PT. When these items were human-scored, all the items in the PT, along with the student responses, were provided to the rater. This allowed the rater to evaluate dependent items based on the previous items that served as the basis for the dependent item.

The three traits measured by the extended writing tasks (full-write responses)—organization and purpose, evidence and elaboration or development and elaboration, and conventions—were evaluated together by a single rater. The rater assigned a separate trait score for each of the three traits.

Items were scored by a team of more than 10 raters under the direction of a scoring team leader. Scoring team leaders were supervised by scoring directors. Each scoring director was responsible for multiple teams in a specific content area and grade level. Responses to individual prompts were assigned to teams of no fewer than three raters. Each rater worked individually on the rater’s own device to read each student response and enter a score for each item.

##### Quality Control Related to Raters

MI has developed a variety of procedures to control the quality of ratings and monitor the consistency of scores provided by raters. These procedures specify rater qualifications and procedures for rater certification, as well as rater retraining or removal. Raters were required to demonstrate their accuracy by passing a qualification set before MI assigned them to score a specific assessment. All full-write, brief-write, reading, and research raters had to pass an additional validation stage.

During scoring, rater accuracy was continually assessed using validity responses, by evaluating how closely raters’ scores aligned with the benchmark scores of the validity response. MI used an automated evaluation system that identified raters who required retraining or removal if adequate improvement is not demonstrated. Automated removal of raters and score resets were performed when item and rater performance failed to meet accuracy expectations. In these cases, all responses scored by a rater during a period of poor performance were reset and redistributed to other qualified raters for rescoring. Rater qualification and calibration are key components in maintaining quality and consistency.

Team leaders also monitored raters’ performance. This involved reviewing the rater’s misscored validity responses and spot-checking (that is, reading behind) raters’ work. Results were used to provide individualized feedback to raters based on their performance. Refer to the [*Monitoring Raters*](#_Monitoring_Raters_1) subsection for more information on this process.

###### Rater Qualification

Table 9.1 summarizes the overall active human raters who were trained and prepared for scoring for the 2023–24 test administration. Note that the numbers for Spanish are for all raters who scored for the 2023–24 test administration for MI across Smarter Balanced states. They are not specific to only California educators.

Table 9.1 Summary of Characteristics of MI Human Raters Scoring CAASPP Smarter Balanced Assessments

|  |  |  |
| --- | --- | --- |
| **Characteristic** | **N** | **Percent** |
| Fluent in Spanish and expressed interest in scoring assessments in Spanish | 387 | 9.10% |
| Experience teaching in a kindergarten through grade twelve school | 456 | 10.80% |
| Currently works in a kindergarten through grade twelve school in California | 96 | 2.30% |
| Others—Not meeting any of the previous criteria | 3,296 | 77.80% |
| **Total Raters Scoring in 2023–24:** | **4,235** | **100%** |

California educators with teaching credentials, retired educators, and school administrators were encouraged to apply.

All team leaders and raters were required to qualify before scoring and were informed of what they were expected to achieve to qualify. Refer to [*7.2.2 Training Overview*](#_Training_Overview) for a more complete description of this training.

Training sets are nonconsequential, as the sets provide the raters the opportunity to score sample papers and receive feedback, including the correct score point and rationale associated with that score point and the sample paper. Training sets are a learning tool that the raters are required to complete. Nonadjacent scores may occur in the training sets, as minimum agreement standards are not part of training sets.

Upon completion of the required training sets, raters moved on to a consequential qualification set that determined rater eligibility for operational scoring of a particular item type. Qualification sets have minimum agreement levels that were enforced, and nonadjacent scores were not allowed. All 0–4 and 0–3 point items adhered to the Smarter Balanced recommendation of a 70 percent exact and 0 percent discrepant (nonadjacent) agreement rate to score.

The standards, provided in subsection [*7.2.5.2 Rater Agreement*](#_Rater_Agreement), are minimum qualification expectations for the various score point ranges and the qualification standard in terms of the percentage of exact agreement. This qualification set, like the validity papers discussed in the next subsection ([*Monitoring Raters*](#_Monitoring_Raters_1)), had been scored previously by scoring experts. Raters needed to score the papers in the same manner according to the percentage of agreements listed in table 9.2.

Table 9.2 Rater Qualification Standard for Agreement with Correct Scores

|  |  |
| --- | --- |
| **Score Point Range** | **Qualification Standard (Exact Agreement)** |
| 0–1 | 90% |
| 0–2 | 80% |
| 0–3 | 80% |
| 0–4 | 70% |

The qualification process was conducted through an online system that captured the results electronically for each individual trainee.

An additional validation stage was implemented to support rater qualification for full-write, brief-write, reading, and research items. After training and initial qualification, all prospective raters were required to score a 20-response set of prescored student answers, sourced from the prior test administration. This process, similar to the qualification step, required raters to meet accuracy standards to proceed with operational scoring. Any rater failing to meet these standards was disqualified from scoring that item, even if they had previously qualified.

###### Monitoring Raters

Validity Responses

Validity responses provided the primary tool for monitoring rater performance. Smarter Balanced provides a small set of validity responses for all vendors, which are supplemented by responses selected and approved by MI scoring management. This validity pool includes anchor responses from the field test administration and is designed to be representative of operational responses, with adequate examples of each score point. Raters’ scores on validity responses are compared against benchmark scores to assess scoring accuracy.

Validity responses were inserted randomly into each rater’s scoring queue at a rate of 5 percent of the total responses scored by a rater during a rater’s shift.

Read-Behinds

Any responses scored by a rater are eligible for “read-behind” (score review) to confirm the accuracy of the scores assigned, although the percentage could vary depending on item type and rater performance. Scoring leaders read behind raters throughout a shift and entered their own scores on responses that raters had read. Both first and second readings were eligible for read-behind.

A scoring leader reviewed the randomly selected responses after raters submitted scores. Leaders reviewed rater scoring statistics (that is, interrater reliability, score point distributions, and validity performance) to determine the need for monitoring via read-behinds or additional training. Responses determined to be scored incorrectly during read-behind review could be rescored by leadership and used to inform and instruct raters as a performance improvement strategy.

In addition, MI identified responses that were randomly selected to receive a second score from an expert rater. These “second reads” were a separate monitoring tool that differs from “reading behind.”

Monitoring Reports

MI staff created performance scoring reports so that scoring directors could monitor the daily human-scoring process and plan any retraining activities needed. For monitoring interrater reliability, 10 percent of the student responses that already were scored by the raters were randomly selected to receive a second score and assigned to expert raters by the scoring system. The second rater was unaware of the first rater’s score. The evaluation of the response from the second expert rater was compared to that of the first rater. Team leaders and scoring directors also conducted back-reads during their shifts for additional quality review.

Real-time management tools allowed all members of scoring management to access

* the validity results by item, rater, and response;
* the interrater reliability rate, which measured the percentage of agreement when the scores assigned by raters were compared to the scores assigned by expert raters, including scoring managers;
* the read rate, which is defined as the number of responses read per hour; and
* the projected date for completion of the scoring for a specific prompt or task.

##### Scoring Verification

Automated score verifications were introduced to enhance scoring accuracy. For instance, a “blank” check reset scores if a response marked as “blank” contained characters like spaces or tabs. In such cases, the score was only recorded after three independent raters assigned the “blank” condition code. Similarly, checks were applied when non-“blank” scores were given to responses with no characters. Scores were also reset automatically when two raters assigned nonadjacent scores, mismatched condition codes, or a mix of a condition code and a numeric score. These responses were rescored, and raters’ details were reviewed by scoring directors to identify retraining needs.

#### Quality Control of Automated Scoring

The MI hybrid scoring approach involved several QA steps. First, models were trained exclusively with responses scored by expert raters. Second, each automated scoring model underwent evaluation, including assessing the quality of the training data and comparing the model’s performance to expert raters. Third, for models using prior administration responses, generalizability to 2023–24 operational responses was verified through secondary validation. Lastly, quality during scoring was ensured by routing at least 15 percent of responses, notably those that exhibited characteristics of multiple score points, to expert raters for scoring. Refer also to [*7.2.6.5 Quality Assurance*](#_Quality_Assurance_1) for details of quality control.

#### Machine-Scoring Procedures

To ensure valid item-level scoring for the Smarter Balanced Summative Assessments, quality-control procedures were employed by CAI, the CAASPP subcontractor responsible for providing the Test Delivery System (TDS) and scoring machine-scorable items. CAI staff independently reviewed all Smarter Balanced Summative Assessment forms by producing sample results for assessments. The sample results were compared with the answer keys for each form to confirm the accuracy of scoring keys. The scores for all applicable items were recorded. A final comparison of the test map to each computer-based form as configured in the user acceptance testing (UAT) environment ensured that no changes to the form were introduced prior to operational deployment.

A real-time, quality-monitoring component was built into the TDS. After an assessment was administered to a student, the TDS passed the resulting data to the QA system. QA conducted a series of data integrity checks, ensuring, for example, that the record for each assessment contained information for each item, keys for multiple-choice items, score points in each item, and the total number of operational items. In addition, QA also checked to ensure that the test record contained no data from items that might have been invalidated.

Data passed directly from the Quality Monitoring System to the database of record, which served as the repository for all test information, and from which all test information was pulled and transmitted to ETS in a predetermined results format.

#### Rater Agreement Results

The statistics for interrater reliability for all human-scoring items at all grade levels are presented in table 7.3 and [appendix 8.I](#_Appendix_8.I:_Interrater). The statistics for interrater reliability for all automated scoring (artificial intelligence [AI])–human scoring items are presented in table 7.11 through table 7.16. These statistics include the percentage of perfect agreement and adjacent agreement between the two readers and the quadratic-weighted kappa (QWK) statistic. QWK is a statistic used to measure the degree of association between two ratings with values ranging from 0.0 (indicating no agreement) to 1.0 (indicating perfect agreement). Refer to subsection [*8.6.8.1.3 Quadratic-Weighted Kappa*](#_Quadratic-Weighted_Kappa_3) for detailed information on QWK.

#### Development of Scoring Specifications

A number of measures were taken to ascertain that the scoring keys were applied to the student responses as intended and the student scores were computed accurately. ETS built and reviewed the scoring system models based on the Smarter Balanced Assessment Consortium scoring specifications and CDE requirements (Smarter Balanced, 2014; American Institutes for Research, 2015). Machine-scored item responses and demographic information were collected and provided electronically to ETS in a master student data file. Human-scored item responses were sent electronically to MI scoring centers for scoring by trained, qualified raters. Record counts were verified against the counts obtained during security check-in from the document processing staff to ensure all students were accounted for in the file.

Once the record counts were reviewed, the machine-scored item responses were scored against the appropriate answer key provided by the Smarter Balanced Assessment Consortium. In addition, the student’s original response string was stored for data verification and auditing purposes.

The Smarter Balanced Assessment Consortium provided the specifications for scoring the assessments well in advance of the receipt of student response data. These specifications contained detailed scoring procedures, along with the procedures for determining whether a student had attempted an assessment and whether that student response data should be included in the statistical analyses and calculations for computing summary data. Standard quality inspections were performed on all data files, including the evaluation of each student data record for correctness and completeness. Student results were kept confidential and secure at all times.

#### Paper–Pencil Scoring

If an LEA was approved to administer the PPT version of the Smarter Balanced Summative Assessments, student responses were entered into the DEI and scored electronically and by a rater, depending on the item type.

### Quality Control of Psychometric Processes

#### Score Verification

ETS developed two independent and parallel scoring structures to produce students’ scores: the Enterprise Score Key Management (eSKM) scoring system, which collected, scored, and delivered individual students’ scores to the ETS reporting system; and then the ETS Psychometric Analysis & Research (PAR) team computed individual student scores based on the same scoring specifications as described in subsection [*9.3.5 Development of Scoring Specifications*](#_Toc121308511). The scores from the two sources were then compared for internal quality control. Any differences in the scores were discussed and resolved. All scores complied with the ETS scoring specifications and passed the parallel scoring process. This ensured the quality and accuracy of scoring and supported the transfer of scores into TOMS, the database of the student records scoring.

#### Psychometric Analyses

The psychometric procedures for the Smarter Balanced Summative Assessments were developed, reviewed, and approved prior to the receipt of student response data. The ETS psychometric team also developed specifications for each of the psychometric analyses performed. These specifications contain detailed descriptions of the analysis steps such as sample inclusion, analyses methods, and special handling of the data.

All psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists and psychometric specifications were developed by members of the team for each of the statistical procedures performed on Smarter Balanced Summative Assessment results data, including item analyses, differential item functioning analyses, item response theory (IRT) calibration, equating, and scaling.

Detailed checklists were developed by members of the team for each of the statistical procedures. Classical item analyses were performed to evaluate the performance of the operational items. Classical item statistics included item difficulty and correlations between item scores and total scores. Items that were flagged for questionable statistical attributes were sent to ETS Assessment Development (AD) staff for review; their comments were then reviewed by the psychometricians before the review by the CDE. The ETS AD and PAR teams worked together to evaluate and make recommendations to the CDE about any problematic items that should be removed from IRT calibration.

IRT calibration of field test items included checks to ascertain that the input files were established accurately. Checks were also made on the number of items, number of students with valid scores, IRT item difficulty and discrimination estimates, standard errors for the item difficulty estimates, and the equating and scaling process. Two psychometricians conducted parallel calibration processing and compared the results to check for any inconsistency. Psychometricians also performed detailed reviews of relevant statistics to determine whether the chosen IRT model fits the data. ETS then presented and reviewed the calibration results with the CDE for approval.

Once raw-to-scale-score conversion tables for each form were generated, psychometricians carried out quality-control checks on each scoring table to verify

* all possible raw scores for each form were included in the tables;
* the lowest obtainable scale score and the highest obtainable scale score matched the specifications for each grade level, respectively; and
* the threshold score for the achievement level was correctly identified.

After all quality-control steps were completed and any differences were resolved, one final inspection of scoring tables was conducted prior to uploading the tables to eSKM for score reporting.

### Quality Control of Reporting

To ensure the quality of Smarter Balanced Summative Assessment results, for both individual student and summary reports, three general areas were evaluated:

1. Comparison of report formats with input sources from the CDE-approved samples
2. Validation of the report data through quality-control checks performed by the ETS Data Quality Services and Center of Reporting & Scoring Services teams, as well as running of all Student Score Reports (SSRs) through the ETS patented QC Interrogator software, which compares elements of the SSR to acceptable values to identify errors and is used in conjunction with human review to detect errors on every score report batch as part of quality-control procedures
3. Proofreading of the quality-control and production reports by the CDE and ETS prior to making reports available to the LEA for download in TOMS and the California Educator Reporting System as well as via the LEA student information system

All reports were required to include a single, accurate LEA code, an LEA name, and a school name. All elements conformed to the CDE official county/district/school (CDS) code and naming records. From the start of processing through scoring and reporting, the CDS Master File was used to verify and confirm accurate codes and names. The CDE provided a revised LEA Master File to ETS throughout the year as updates became available.

After the reports were validated in accordance with CDE requirements, a set of reports representing all possible grade levels, content areas, and reporting outcomes was provided to the CDE and ETS for review and approval. Electronic reports were sent on the actual report template to the CDE. The CDE and ETS reviewed and approved the reports after a thorough examination.

Upon CDE approval of the reports generated for the quality-control LEAs, ETS proceeded with the first batch of report production. The reviewed set of reports incorporated CDE-selected LEAs and provided the final check prior to generating all reports and making them available electronically for download in TOMS and for student information systems through an application programming interface.

#### Exclusion of Student Scores from Summary Reports

ETS provided the CDE with reporting specifications that documented when to exclude student scores from summary reports. These specifications included the logic for handling submitted assessments that, for example, indicated the student tested but responded to no items, was absent, was not tested because of parent/guardian request, or did not complete the assessment because of illness. The methods for handling other anomalies were also covered in the specifications. These anomalies are described in more detail in [*7.5.2 Special Cases*](#_Special_Cases_3).

### Quality Control of End-to-End Testing

ETS conducted end-to-end testing prior to the start of the test administration. The purpose of this testing is to verify that all systems, processes, and resources were ready for the operational administration. Once released from processing, the test results were sent through the system for scoring and reporting. SSRs were created, along with data files for subject-matter experts in the teams to review and verify.

#### Computer-based Assessments

ETS employed a number of strategies to verify ongoing systems performance, including monitoring of system availability and system usage. Time was allotted for UAT to confirm that the systems met requirements and to make identified corrections before final deployment. To accomplish system acceptance and sign-off, ETS deployed systems to a staging area, which mirrors the final production environment, for operational testing and UAT. Final approval by the CDE triggered final deployment of the system.

To begin the quality-control process for end-to-end testing of the administration, the ETS program and resolutions teams prepared by entering responses in computer-based assessments for all grade levels and content areas. These responses were entered for fictitious students in selected schools and across several LEAs. Each student’s assessment was completed with responses that were all correct, all incorrect, and combinations of correct and incorrect. These response combinations were the expected results across achievement levels and score ranges. The responses were sent for processing, including for system quality control of computer-based assessments.

Once released from processing, the test results were sent through the system for scoring and reporting. SSRs were created, along with data files for subject-matter experts in the teams to review and verify. Individual SSRs were generated on the basis of the fictitious students when 100 percent quality control was demonstrated by the ETS Resolution staff.

#### Paper–Pencil Tests

The DEI underwent UAT to ensure that the correct test items were available for a grade-level assessment in the DEI. Then, during testing, information technology personnel monitored daily feeds to ensure the completeness and timeliness of records sent for hand scoring.

The processes followed to test the DEI from end to end are described in the previous subsection, [*9.6.1 Computer-based Assessments*](#_Computer-based_Assessments_3).

### References

American Institutes for Research. (2015). *Smarter Balanced scoring specification: 2014–‍2015 administration, version 7*.

California Department of Education. (2024). *CAASPP paper–pencil testing test administration manual*. Sacramento, CA: California Department of Education.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.

Smarter Balanced Assessment Consortium. (2014). *Hand-scoring rules*. Los Angeles, CA: Smarter Balanced Assessment Consortium.

## Historical Comparisons

Historical comparisons are performed to identify the trends in student performance and test characteristics over time. Such comparisons are performed for three of the most recent test administration years of California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced administration—2023–24, 2022–23, and 2021–22. The comparisons include both cross-sectional comparisons for the same grade levels in different years (with different students) and longitudinal comparisons for the same students in different years (in different grade levels).

The indicators of student performance include the mean and standard deviation (SD) of scale scores and the percentage of students classified into achievement levels for an overall assessment and into performance levels for claims. Test characteristics are compared by examining the reliability and standard error of measurement (SEM) for each assessment.

### Student Performance

#### Cross-Sectional Comparisons on the Overall Assessments

In cross-sectional comparisons, cohorts of students from the 2022–‍23 CAASPP administration are compared to students in the same grade levels from the 2023–‍24 CAASPP administration. For example, students enrolled in grade three for the 2022–23 CAASPP administration are compared with students enrolled in grade three for the 2023–24 CAASPP administration.

As noted in table 7.21 in [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting), the reporting scale ranges from 2115 to 2900 for English language arts/literacy (ELA) and from 2190 to 2900 for mathematics. The difference between the two adjacent years in average scale score and percentage of students meeting or exceeding standards is the later year’s values minus the previous year’s values for the same grade level. Therefore, a positive value indicates an increase from 2022–23 to 2023–24 and a negative value indicates a decrease. The achievement level percentage may not sum to exactly 100 or to exactly the combined achievement level percentage because of rounding.

##### Summary Statistics

Table 10.A.1 in [appendix 10.A](#_Appendix_10.A:_Cross-Sectional) contains the number of students assessed; the number of students with valid scores; the means and SDs of students’ scale scores in 2021–22, 2022–‍23, and 2023–24 for each assessment; and the differences in scale scores between 2021–22 and 2022–23 and between 2022–23 and 2023–24.

The performance trend shows that the average scale score increased from 2022–23 to 2023–24 in grade six for ELA and in all grade levels for mathematics; and decreased in grade seven in ELA.

##### Achievement Levels of Overall Students

Scale score thresholds are used to classify each student into one of four achievement levels: *Standard Not Me*t, *Standard Nearly Met*, *Standard Met*, or *Standard Exceeded*. Refer to table 7.21 in [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting) for the achievement level scale score ranges for each assessment.

The percentage of students for each achievement level and qualifying for the *Standard Met* and *Standard Exceeded* levels, as well as the differences in the percentage of the students in *Standard Met* and *Standard Exceeded* levels between 2021–22 and 2022–23 and between 2022–‍23 and 2023–24, are presented in table 10.A.2 in [appendix 10.A](#_Appendix_10.A:_Cross-Sectional). There is an increase in the percentage of students who met or exceeded standards from 2022–23 to 2023–24 in grades four, five, six, eight, and eleven in ELA and in all grade levels in mathematics; and a decrease in this percentage for grades three and seven in ELA. Note that this information may differ slightly from information found on the California Department of Education (CDE) Test Results for California’s Assessments website because of different dates on which the data was accessed.

##### Scale Score Distributions

Table 10.A.3 through table 10.A.6 in [appendix 10.A](#_Appendix_10.A:_Cross-Sectional) show the distribution of scale scores observed in 2021–22, 2022–23, and 2023–24 for each grade level and content area. Frequency counts are provided for each scale score interval of 30. “N/A” indicates that there is no obtainable scale score in the interval. The scale score ranges for each grade on the vertical scale are those defined by the Smarter Balanced Assessment Consortium. Refer to table 7.21 in [chapter 7](#_Scoring_and_Reporting) for the scale score ranges.

##### Achievement Levels of Selected Student Groups

Table 10.A.7 through table 10.A.20 in [appendix 10.A](#_Appendix_10.A:_Cross-Sectional) provide statistics summarizing student achievement by content area and grade level for selected student groups. In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status, disability status, migrant status, assigned designated supports, and assigned accommodations. The tables show, for each demographic student grouping, the number of students with a valid scale score, scale score means and SDs, and the percentage of students in each achievement level for 2021–22, 2022–23, and 2023–24.

The tables also show the differences in the percentage of *Standard Met* or *Standard Exceeded* between 2021–22 and 2022–23, as well as between 2022–23 and 2023–24. From 2022–23 to 2023–24, most student groups show increases in percentages of students meeting or exceeding standards in ELA at most grade levels except grades three and seven, where most student groups showed decreases. For mathematics, most student groups showed increases in all grade levels.

#### Longitudinal Comparisons on the Overall Groups

For longitudinal comparisons, the data is gathered and compared for the same students in 2021–22, 2022–23, and 2023–24. Through vertical scaling, scores on assessments at different grade levels of the same content area are placed on a common scale. For Smarter Balanced Summative Assessments, reporting scores on a vertical scale allows student progress to be tracked for a particular content area across grade levels.

The difference in average scale scores or in the percentage of students meeting or exceeding standards is the later year’s (that is, 2023–24) values minus the previous year’s (that is, 2022–23) values for the same students. Therefore, a positive value indicates an increase in the later year (that is, 2023–24) and a negative value indicates a decrease in the later year (that is, 2023–24). Individual achievement level percentages may not sum to exactly 100 or the combined achievement level percentage because of rounding.

For year-to-year comparisons, the differences between 2021–22 and 2022–23 and the differences between 2022–23 and 2023–24 are presented. The statistics in these tables include only those students who advanced one grade level each year and whose scores are available in all three years. As mentioned previously, only the differences between 2022–23 and 2023–24 are discussed.

##### Summary Statistics

Table 10.B.1 in [appendix 10.B](#_Appendix_10.B:_Longitudinal) shows the number of students assessed, the number of students with valid scores, the means and SDs of students’ scale scores in 2022–23 and 2023–24 for each assessment, as well as the differences in scale scores between 2021–22 and 2022–23. Table 10.B.2 presents the same set of statistics as in table 10.B.1, but for all three administration years (2021–22, 2022–23, and 2023–24), as well as the year-to-year differences in scale scores.

Longitudinal comparisons for the scale scores reveal positive gains for all grade levels in both content areas for the two-year trend. The smallest gain occurs in grades seven to eight in ELA, with an increase in 13 score points. The largest gain occurs in ELA for grades three to four, with an increase of 44 points.

##### Achievement Levels of Overall Students

The percentage of students for each achievement level and qualifying for the *Standard Met* and *Standard Exceeded* levels, as well as the differences in the percentage of the students in *Standard Met* and *Standard Exceeded* between 2022–23 and 2023–24, are presented in table 10.B.3 in [appendix 10.B](#_Appendix_10.B:_Longitudinal). The same information is presented in table 10.B.4 for all three test administration years (2021–22, 2022–23, and 2023–24).

The two-year performance trend (that is, 2022–23 and 2023–24), in terms of the percentage of students meeting or exceeding standards, shows an increase for grades three to four, four to five, and six to seven in ELA, as well as grades five to six and six to seven in mathematics. Conversely, there was a decrease in grades five to six and seven to eight in ELA, as well as grades three to four, four to five, and seven to eight in mathematics.

##### Scale Score Distributions

Table 10.B.5 and table 10.B.8 in [appendix 10.B](#_Appendix_10.B:_Longitudinal) show the distribution of scale scores observed in 2022–23 and 2023–24 for the same students per each grade level in ELA or mathematics, respectively. Frequency counts are provided for each scale score interval of 30. The scale score distributions for 2021–22, 2022–23, and 2023–24 are presented in table 10.B.6 and table 10.B.7 for ELA and table 10.B.9 and table 10.B.10 for mathematics, respectively.

##### Achievement Levels of Selected Student Groups

Table 10.B.11 through table 10.B.20 in [appendix 10.B](#_Appendix_10.B:_Longitudinal) provide statistics summarizing student performance in 2022–23 and 2023–24 for the same students by content area and grade level for selected groups of students. In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status, disability status, migrant status, the assignment of designated supports, and the assignment of accommodations.

The tables show, for each student group, the number of students with valid scale scores in 2022–23 and 2023–24 as well as the scale score means and SDs and the percentage of students in each achievement level. Additionally, the differences in the percentage of *Standard Met* and *Standard Exceeded* between 2022–23 and 2023–24 are shown. The statistics for three years—2021–22, 2022–23, and 2023–24—are presented in table 10.B.21 through table 10.B.28.

In ELA, the two-year trend from 2022–23 to 2023–24 shows mixed results: a positive gain occurring in grades four, five, and seven for all groups, with the exception of Black or African American students in grade four. There is a negative gain occurring in grade eight for all groups, in the percentage of students meeting or exceeding standards. The exceptions are for English learner (EL) students and for students with disability or assigned accommodation in grade eight. In mathematics, the two-year comparison between 2022–23 to 2023–24 indicates that virtually all groups show positive gains in the percentage of students meeting or exceeding standards in grades six and seven but negative gains in grades four, five, and eight. The exceptions are Native Hawaiian or Other Pacific Islander students in grade seven, students in grade seven whose migrant status is “Yes” who showed negative gains, and EL students in grade eight who showed positive gains.

### Test Characteristics

The marginal reliabilities and SEMs expressed in theta score units for each assessment are presented in table 10.C.1 in [appendix 10.C](#_Appendix_10.C:_Comparison).

A reasonably high degree of reliability is consistent over three years, from 2021–22, 2022–‍23, and 2023–24. Reliabilities are stable over 2022–23 and 2023–24. For all grade levels and content areas, reliabilities in 2022–23 and 2023–24 are within the range of 0.86 to 0.89 for ELA and 0.86 to 0.91 in mathematics.

Reliabilities are affected by both item characteristics and student characteristics. Refer to subsections [*8.6.2 Reliability Measures*](#_Reliability_Measures) and [*8.6.3 Standard Error of Measurement*](#_Standard_Error_of_2) for the methods used to calculate marginal reliability and SEM, respectively.

### Appendix 10.A: Cross-Sectional Comparisons of the Overall Group and Student Groups on the Overall Assessments

*This content is located in a separate file.*

### Appendix 10.B: Longitudinal Comparison of the Overall Group and Student Groups on the Overall Assessments

*This content is located in a separate file.*

### Appendix 10.C: Comparison of Test Characteristics

*This content is located in a separate file.*

## Paper–Pencil Forms

This chapter provides a summary of test assembly, test administration, calibration, and scaling procedures that are specifically applied to the paper–pencil tests (PPTs), as well as the results of the analyses performed on the data for students who took PPTs instead of the computer-based assessments. Analyses include score summaries, item response theory (IRT) parameter values, correlations between composite claims and between content areas, and the assignment of designated supports and accommodations.

Because there were no grade-level assessments with more than 17 students in the group, no analysis can be conducted and reported for PPTs administered in 2023–24.

### Overview

The PPT versions of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics in braille, large-print, and standard print versions are made available to students whose need to take a PPT is documented in the student’s individualized education program or Section 504 plan. PPT versions may also be provided to schools experiencing unexpected, temporary technology issues that are beyond the schools’ control, based on California Department of Education (CDE) approval of the request. The PPT versions contain a fixed set of items that includes components of the computer-based assessment such as multiple-choice (MC) items, constructed-response items, and performance tasks (PTs). The assembled PPTs used the Smarter Balanced full-form blueprint that is mentioned in subsection [*2.2.2.1 Operational Items*](#_Operational_Items_1) and included in [appendix 4.A](#_Appendix_4.A:_Smarter).

PPT versions were available only with prior permission from the CDE.

The number of students with valid scores is presented in table 11.1. “Valid score” means the student records were not flagged as “not scored,” and the students were enrolled in the same grade level as they were tested.

Table 11.1 Number of Students with Valid Scores from CAASPP Smarter Balanced Summative PPTs

|  |  |  |
| --- | --- | --- |
| **Grade Level** | **ELA** | **Mathematics** |
| 3 | 14 | 13 |
| 4 | 14 | 14 |
| 5 | 14 | 13 |
| 6 | 17 | 17 |
| 7 | 7 | 7 |
| 8 | 12 | 12 |
| 11 | 7 | 8 |
| **Totals:** | **85** | **84** |

The next sections in this chapter include descriptions of the methodology and procedures of scoring the Smarter Balanced summative PPTs.

### Universal Tools, Designated Supports, and Accommodations

Consistent with the computer-based assessments, designated supports, accommodations (subsections [*5.4.1.1 Universal Tools*](#_Universal_Tools_3), [*5.4.1.2 Designated Supports*](#_Designated_Supports), and [*5.4.1.3 Accommodations*](#_Accommodations)) and unlisted resources (subsection [*5.4.1.4 Unlisted Resources*](#_Unlisted_Resources)) are assigned to individual students based on student need.

### Calibration and Scaling

When a Smarter Balanced Summative Assessment paper form was built for the first time for the 2014–15 test administration, a postequating procedure was conducted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), using testing data from 2014–15 PPTs administered from representative member states of the Smarter Balanced Assessment Consortium (CRESST, 2015). To produce scores for the PPTs that are on the same scale as the computer-based assessments, separate calibrations of the PPT response data were conducted and then scaled to the computer-based item bank. The “new” calibrations for the PPT versions were established by calibrating samples of item response data from the PPT administrations; the “reference” calibrations were based on the Smarter Balanced Online Summative Assessment item bank that was established during the field test.

After the resulting equated item parameters for the summative PPTs were determined, the Smarter Balanced member states used those item parameters to estimate student theta scores using the pattern scoring procedures, and then the theta scores were converted to the scale scores using the same transformation constants that the computer-based assessments used. In summary, a postequating was conducted in the first year for a new paper form to scale the item parameters to the computer-based item pool scale based on testing data. After the equated item parameters were obtained in the first year, and when the form was reused in the subsequent years, the same procedures as used for computer-based assessments were used to estimate and produce the theta scores and scale scores. Refer to section [*7.3 Student Test Scores*](#_Student_Test_Scores) for details.

For the purpose of linking the PPT forms to the official reporting scale derived from the computer-based test mode, the PPT item parameter estimates were placed on the reference scale by using a set of anchor items that were not modified. Specifically, these unmodified items indicated these items could appear in either test delivery mode as is without altering the construct; that is, the item parameter estimates should be invariant across the delivery mode.

The procedure used for equating the Smarter Balanced summative PPTs involved three parts: initial item calibration, anchor item evaluation, and final item calibration. Each of those procedures, as described in the next subsection, was applied to all assessments. The calibrations were performed with the flexMIRT® item response modeling software (Cai, 2017).

#### Initial Calibration

The following steps are involved in the initial calibration to obtain item parameter estimates and model goodness-of-fit indices. The generalized partial credit model (GPCM) was applied to both MC items and polytomously scored items. Refer to subsection [*7.3.1 Total Test Scores*](#_Total_Test_Scores) in [*Chapter 7: Scoring and Reporting*](#_Scoring_and_Reporting) of this report for the mathematical formula of the GPCM:

1. The parameters of all unmodified items are fixed to the parameter values obtained from the computer-based item pool.
2. The parameters of all modified items are freely estimated.
3. The latent variable density is estimated as an empirical histogram (refer, for example, to Woods, 2007; Houts & Cai, 2013) with estimated mean and variance from the “all” student population, including students taking computer-based assessments.

#### Anchor Item Evaluation

The purpose of anchor item evaluation is to select items that function similarly across both computer-based and PPT modes as anchors. By linking assessments through these anchor items, PPT results are placed onto the computer-based assessment scale and scores from the two modes should be comparable.

A series of calibrations identical to the “initial” calibration were performed but with the parameters of one unmodified item at a time freely estimated. The parameters of all other unmodified items were fixed to their prior estimates from the computer-based item pool. As in the initial calibration, the parameters of all modified items were freely estimated, along with the population distribution’s mean, variance, and shape.

To decide whether each unmodified item should be retained or rejected as an anchor in the final calibration for the PPT forms, the parameter estimates from the computer-based item pool administration and the parameter estimates from the initial calibration were used to compute the expected score functions for the two modes of test administration. The two expected score functions—for the computer-based and PPT administrations—were plotted, and differences in item functioning across the two modes were quantified by computing a weighted Area Between the Curves (wABC; refer to Hansen et al., 2014). Any items with a wABC value greater than 0.150 were rejected as anchors.

#### Final Calibration

For assessments in which any unmodified item was rejected as an anchor, a final calibration was conducted using the approach described in subsection *11.6.1 Initial Calibration* of the *2016–‍17 CAASPP Smarter Balanced Technical Report* (CDE, 2018), except that the parameters of all rejected anchor items were freely estimated. Parameters of the modified items also were freely estimated. The parameter estimates from this final calibration were used in scoring the PPT forms. In this way, PPT version scores were placed on the computer-based assessment scale.

### Scoring

As in the CAASPP Smarter Balanced computer-based assessments, student item responses in the PPT forms were scored and individual student scores were calculated (that is, overall scale scores and composite claims and subscores) based on the scored item responses. The same scoring specifications and procedures as in the computer-based assessments were followed. However, because of the small student sample sizes, caution should be taken when interpreting some of the summary statistics.

#### Number of Students Who Took the 2023–24 Smarter Balanced Paper–‍Pencil Assessments

Table 11.1 presents the number of students who took the 2023–24 Smarter Balanced summative PPTs. There were 169 PPTs administered across grade levels and content areas in the 2023–24 administration, with 85 students testing in ELA and 84 students testing in mathematics. The largest group testing was in grade six, where 17 students took both the ELA and mathematics PPTs.

#### Item Response Theory Parameter Values

Parameter estimates for the PPT versions of the 2023–24 CAASPP Smarter Balanced operational items were obtained using the procedure described in section [*11.3 Calibration and Scaling*](#_Calibration_and_Scaling). Summary statistics of these parameter estimates are calculated to show the difficulty and discrimination for the overall assessment, as well as the difficulty and discrimination for the composite claims; distributions of *b*-value and *a*-value parameter estimates are created to provide more detail. The step parameters for all polytomous items are also presented.

Table 11.A.1 through table 11.A.14 in [appendix 11.A](#_Appendix_11.A:_Item) present univariate statistics (mean, standard deviation, minimum, and maximum) of the scaled IRT *a*-values. For each assessment, the results are presented for all items in the assessment and for the items in each composite claim. Table 11.A.15 through table 11.A.28 present the univariate statistics of the IRT *b*-‍values for all items in the assessment and for the items in each composite claim.

Table 11.A.29 through table 11.A.35 show PT item statistics in ELA. Table 11.A.36 through table 11.A.42 show PT item statistics in mathematics.

### References

Cai, L. (2017). *FlexMIRT: Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

California Department of Education. (2018). *2016–17 California Assessment of Student Performance and Progress Smarter Balanced technical report*. California Department of Education website.

CRESST. (August, 2015). *Initial report on the calibration of paper and pencil forms*. Los Angeles, CA.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. S., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine and Tobacco Research*, *16*, *Supplement 3,* S175-S189.

Houts, C. R., & Cai, L. (2013). *FlexMIRT user’s manual version 2: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and psychological measurement*, *67*, 73–87.

### Appendix 11.A: Item Response Theory Parameter Estimates

*This content is located in a separate file.*

## Continuous and Systematic Improvement

The tenth operational administration of the Smarter Balanced Summative Assessments for English language arts/literacy (ELA) and mathematics occurred in 2023–24. Throughout the past 10 years, continuous efforts have been made to improve the assessments. This chapter summarizes accomplishments and ongoing improvements for the Smarter Balanced Summative Assessments as well as strategies to implement possible future improvements.

Because the Smarter Balanced Assessment Consortium owns the test design and item development of these assessments, the focus of ETS continuous improvement is limited to test administration, scoring and reporting, and analyses.

### 2023–24 Feedback for Continuous Improvement Survey

The California Assessment of Student Performance and Progress (CAASPP) program annually solicits feedback from educators through the Feedback for Continuous Improvement Survey. Local educational agency (LEA) and test site staff, as well as test administrators and test examiners, were invited to participate in the 2023–24 Feedback for Continuous Improvement Survey. Its goal was to highlight successes and identify areas for improvement. A total of 3,162 survey respondents participated in this survey for the 2023–‍24 administration, compared to 3,869 respondents for the previous year. The California Department of Education (CDE) and ETS use key recommendations from educators to implement positive changes in the following administration year.

Educators provided valuable feedback for potential improvements to the future administration of the CAASPP and the English Language Proficiency Assessments for California (ELPAC) by reporting some lessons they learned in 2023–24. Based on those lessons and suggestions for improvement, the *CAASPP and ELPAC Feedback for Continuous Improvement Survey and Focus Groups Report* (CDE, 2024) presents recommendations for the CDE, with the goal of enhancing the administrative support provided to LEAs and schools for future CAASPP and ELPAC test administrations. Refer also to subsection [*5.3.4 Feedback for Continuous Improvement Survey*](#_Feedback_for_Continuous_3) for assessment-specific results.

### Communications

#### Updated Program Website

The CAASPP and ELPAC program websites will be combined into one website, the CAASPP & ELPAC Website. This site will feature responsive design, improved navigability, and a simplified layout.

### Reporting

End-of-year data files will be completed earlier to support the CDE goal of delivering statewide reporting earlier than in previous years.

### Accessibility Resources

Like all CAASPP assessments, the Smarter Balanced for ELA and mathematics are administered using the Test Delivery System (TDS) created by Cambium Assessment, Inc.As such, implementation of new computer-based universal tools, designated supports, and accommodations are aligned with the TDS.

The following changes will be implemented during the 2024–25 CAASPP Smarter Balanced administration:

* Text-to-speech will be added to the student sign-in workflow and be available for all students who select the [**Read Page**] button on the *Student Sign-In* and subsequent pages.
* The embedded American Sign Language accommodation will be enabled after the student selects an icon instead of a context menu item.
* Color contrast with non-embedded print-on-demand will no longer be listed as a non-embedded accommodation.
* The embedded text-to-speech accommodation and text-to-speech in Spanish mathematics designated support will be enhanced in the following ways:
* Will include line tracking, which causes the entire line to be highlighted lightly and each spoken word to be highlighted in a contrasting color as it is read aloud
* Will be enabled after the student selects an icon instead of a context menu item
* The definition of non-embedded alternate response options accommodation will be expanded to include augmentative and alternative communication devices.
* The definition of non-embedded speech-to-text will note that Spanish is available for the Smarter Balanced for Mathematics.
* The embedded and non-embedded word prediction accommodation will be renamed “word completion.”
* The non-embedded braille hybrid adaptive test (HAT) will have a raised-line graphics option (in addition to embossed graphics) for students assigned the braille accommodation for whom HAT has been requested.

### Reference

California Department of Education. (2024). *2023–24 CAASPP and ELPAC feedback for continuous improvement survey and focus groups report* [Unpublished manuscript]. Sacramento, CA: California Department of Education.

1. This definition was retrieved from the Child Care Reporting--Child is English Learner web page on the CDE website. [↑](#footnote-ref-2)
2. Refer to the CAASPP Student Data File Layout (CDE, 2024a) for the complete list of include indicators and condition codes. [↑](#footnote-ref-3)
3. Data for 2023–24 was retrieved from the *CalEdFacts* web page on the CDE website. [↑](#footnote-ref-4)
4. This definition was retrieved from the CDE California Longitudinal Pupil Achievement Data System (CALPADS) web page on the CDE website. [↑](#footnote-ref-5)
5. While California does not administer the Smarter Balanced assessments to students in grade ten, this was included because the simulation was run for all Smarter Balanced Assessment Consortium member states, some of which do test students in grade ten. [↑](#footnote-ref-6)
6. This technical report is based on the versions of the CDE Accessibility Matrix and the *Usability, Accessibility, and Accommodations Guidelines* that were available during the 2023–24 CAASPP administration. [↑](#footnote-ref-7)
7. Smarter Balanced blueprints specify that each WER item is initially scored out of 10 points: one dimension worth 2 points (conventions) and two dimensions each worth 4 points (organization and purpose and evidence and elaboration). However, the two 4-point dimensions are averaged together and then the convention dimension is added, resulting in a final score range of 0–6 points. [↑](#footnote-ref-8)
8. Responses and results of the 2014–15 Smarter Balanced field test administration were used to derive the base scale to which subsequent item parameters are aligned. [↑](#footnote-ref-9)
9. QWK is a measure used to assess the agreement between two raters, accounting for the possibility of agreement occurring by chance and giving more weight to larger discrepancies between ratings. [↑](#footnote-ref-10)
10. Responses and results of the 2014–15 Smarter Balanced field test administration were used to derive the base scale to which subsequent item parameters are aligned. [↑](#footnote-ref-11)
11. The LOTs and HOTs were revised slightly by the CDE in 2020 based on the LOTs and HOTs set by Smarter Balanced in 2015 to avoid the accumulation of LOTs and HOTs. [↑](#footnote-ref-12)
12. Note that the information in this technical report may differ slightly from information found on the CDE Test Results for California’s Assessments website because of different dates on which the data was accessed. [↑](#footnote-ref-13)