

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



# California Assessment of Student Performance and Progress California Spanish Assessment 2020–‍2021 Technical Report

**Submitted May 12, 2023**

**By ETS**



**Contract #CN150012**

Table of Contents

[Chapter 1: Introduction 1](#_Toc102976205)

[1.1. Background 1](#_Toc102976206)

[1.2. Test Purpose 2](#_Toc102976207)

[1.3. Test Design 2](#_Toc102976208)

[1.4. Intended Population 2](#_Toc102976209)

[1.5. Intended Use and Purpose of Test Scores 2](#_Toc102976210)

[1.6. Testing Window 3](#_Toc102976211)

[1.7. Impact of the Novel Coronavirus Disease 2019 Pandemic 4](#_Toc102976212)

[1.7.1. Remote Testing Flexibility 4](#_Toc102976213)

[1.7.2. Test-Taking Rates in the 2020–2021 Administration 4](#_Toc102976214)

[1.8. Significant Developments in 2020–2021 5](#_Toc102976215)

[1.8.1. Remote Testing 5](#_Toc102976216)

[1.8.2. Accessibility Resources 6](#_Toc102976217)

[1.8.3. Reuse of the 2019–2020 Test Forms 6](#_Toc102976218)

[1.8.4. High School Postequating Procedures 7](#_Toc102976219)

[1.9. Groups and Organizations Involved with the CSA 7](#_Toc102976220)

[1.9.1. California State Board of Education 7](#_Toc102976221)

[1.9.2. California Department of Education 7](#_Toc102976222)

[1.9.3. California Educators 7](#_Toc102976223)

[1.9.4. Contractors 8](#_Toc102976224)

[1.10. Systems Overview and Functionality 9](#_Toc102976225)

[1.10.1. Test Operations Management System 9](#_Toc102976226)

[1.10.2. Test Delivery System 9](#_Toc102976227)

[1.10.3. Practice and Training Tests 10](#_Toc102976228)

[1.10.4. California Educator Reporting System 10](#_Toc102976229)

[1.11. Overview of the Technical Report 11](#_Toc102976230)

[References 12](#_Toc102976231)

[Chapter 2: An Overview of CSA Processes 13](#_Toc102976232)

[2.1. Item Development 13](#_Toc102976233)

[2.1.1. Item Format 13](#_Toc102976234)

[2.1.2. Item Specifications 13](#_Toc102976235)

[2.1.3. Item Banking 14](#_Toc102976236)

[2.2. Test Assembly 14](#_Toc102976237)

[2.3. Test Administration 15](#_Toc102976238)

[2.3.1. Test Security and Confidentiality 15](#_Toc102976239)

[2.3.2. Procedures to Maintain Standardization 16](#_Toc102976240)

[2.4. Fairness and Accessibility 16](#_Toc102976241)

[2.4.1. Universal Tools, Designated Supports, and Accommodations 17](#_Toc102976242)

[2.4.2. Description of Differential Item Functioning 18](#_Toc102976243)

[2.5. Scores 19](#_Toc102976244)

[2.5.1. Estimating Ability Scores 19](#_Toc102976245)

[2.5.2. Score Reporting 19](#_Toc102976246)

[2.5.3. Aggregation Procedures 19](#_Toc102976247)

[2.6. Psychometric Analyses 20](#_Toc102976248)

[2.7. Standard Setting 20](#_Toc102976249)

[References 21](#_Toc102976250)

[Chapter 3: Item Development and Form Assembly 22](#_Toc102976251)

[3.1. Overview 22](#_Toc102976252)

[3.2. Test Blueprint 22](#_Toc102976253)

[3.3. Guidelines 23](#_Toc102976254)

[3.3.1. Item Specifications 23](#_Toc102976255)

[3.3.2. Item Format 24](#_Toc102976256)

[3.3.3. Item Types and Features 24](#_Toc102976257)

[3.4. Item Writing Process 26](#_Toc102976258)

[3.4.1. Item Development Plan 26](#_Toc102976259)

[3.4.2. Selection of Item Writers 27](#_Toc102976260)

[3.4.3. Item Writer Training 27](#_Toc102976261)

[3.5. ETS Item Review Process 28](#_Toc102976262)

[3.5.1. Overview 28](#_Toc102976263)

[3.5.2. ETS Content Review 28](#_Toc102976264)

[3.5.3. ETS Accessibility Review 29](#_Toc102976265)

[3.5.4. ETS Editorial Review 29](#_Toc102976266)

[3.5.5. ETS Sensitivity and Fairness Review 29](#_Toc102976267)

[3.6. California Department of Education Review 30](#_Toc102976268)

[3.7. California Educator Review 30](#_Toc102976269)

[3.7.1. Composition of Item Review Panels 30](#_Toc102976270)

[3.7.2. Item and Passage Review 32](#_Toc102976271)

[3.7.3. Data Review 33](#_Toc102976272)

[3.8. Test Assembly and Length 33](#_Toc102976273)

[3.9. Test Production Process 33](#_Toc102976274)

[3.9.1. Psychometric Criteria and Identification of Eligible Items 33](#_Toc102976275)

[3.9.2. Selection of Items 34](#_Toc102976276)

[3.9.3. Forms Review 35](#_Toc102976277)

[3.9.4. Configuration of the Test Delivery System 38](#_Toc102976278)

[References 39](#_Toc102976279)

[Appendix 3.A: CSA Blueprint Overview—Operational Forms 40](#_Toc102976280)

[Chapter 4: Test Administration 44](#_Toc102976281)

[4.1. Overview 44](#_Toc102976282)

[4.1.1. Remote and In-Person Testing 45](#_Toc102976283)

[4.1.2. Test-Taking Summaries 45](#_Toc102976284)

[4.1.3. Demographic Student Group Summaries 47](#_Toc102976285)

[4.2. Procedures to Maintain Standardization 48](#_Toc102976286)

[4.2.1. Local Educational Agency CAASPP Coordinator 48](#_Toc102976287)

[4.2.2. CAASPP Test Site Coordinator 49](#_Toc102976288)

[4.2.3. Test Administrators 49](#_Toc102976289)

[4.2.4. Instructions for Test Administrators 50](#_Toc102976290)

[4.3. Local Educational Agency Training 52](#_Toc102976291)

[4.3.1. Workshops, Virtual Training, and Webcasts 52](#_Toc102976292)

[4.3.2. Videos and Guides 53](#_Toc102976293)

[4.3.3. Training for Proper Identification and Assignment of Designated Supports and Accommodations 53](#_Toc102976294)

[4.4. Accessibility Resources 54](#_Toc102976295)

[4.4.1. Accessibility Resource Categories 54](#_Toc102976296)

[4.4.2. Identification 57](#_Toc102976297)

[4.4.3. Assignment 57](#_Toc102976298)

[4.4.4. Usage of Designated Supports and Accommodations 58](#_Toc102976299)

[4.5. Practice and Training Tests 60](#_Toc102976300)

[4.6. Test Security and Confidentiality 61](#_Toc102976301)

[4.6.1. ETS’ Office of Testing Integrity 61](#_Toc102976302)

[4.6.2. Procedures to Maintain Standardization of Test Security 62](#_Toc102976303)

[4.6.3. Test Security Monitoring 62](#_Toc102976304)

[4.6.4. Security of Electronic Files Using a Firewall 63](#_Toc102976305)

[4.6.5. Transfer of Scores via Secure Data Exchange 63](#_Toc102976306)

[4.6.6. Data Management in the Secure Database 63](#_Toc102976307)

[4.6.7. Statistical Analysis on Secure Servers 64](#_Toc102976308)

[4.6.8. Student Confidentiality 64](#_Toc102976309)

[4.6.9. Student Test Results 64](#_Toc102976310)

[4.6.10. Security and Test Administration Incident Reporting System Process 65](#_Toc102976311)

[4.6.11. Appeals 67](#_Toc102976312)

[References 70](#_Toc102976313)

[Appendix 4.A: Demographic Summaries 71](#_Toc102976314)

[Appendix 4.B: Special Service Summaries 85](#_Toc102976315)

[Chapter 5: Standard Setting 91](#_Toc102976316)

[5.1. Background 91](#_Toc102976317)

[5.2. Achievement Level Descriptors 91](#_Toc102976318)

[5.3. Standard Setting Methodology 92](#_Toc102976319)

[5.3.1. Bookmark Method 92](#_Toc102976320)

[5.4. Standard Setting Procedures 92](#_Toc102976321)

[5.4.1. Panelists 92](#_Toc102976322)

[5.4.2. Materials 93](#_Toc102976323)

[5.4.3. Process (Including Articulation) 93](#_Toc102976324)

[5.5. Results of the Standard Setting Process 95](#_Toc102976325)

[References 96](#_Toc102976326)

[Chapter 6: Scoring and Reporting 97](#_Toc102976327)

[6.1. Student Test Scores 97](#_Toc102976328)

[6.1.1. Scoring of Incomplete Cases 97](#_Toc102976329)

[6.1.2. Theta Scores 98](#_Toc102976330)

[6.1.3. Scale Scores for the Total Assessment 99](#_Toc102976331)

[6.1.4. Score Reporting Ranges 100](#_Toc102976332)

[6.2. Overview of Score Aggregation Procedures 100](#_Toc102976333)

[6.2.1. Individual Student Score Distributions and Summary Statistics 101](#_Toc102976334)

[6.2.2. Demographic Student Group Summaries 102](#_Toc102976335)

[6.3. Reports Produced and Scores for Each Report 102](#_Toc102976336)

[6.3.1. Online Reporting 103](#_Toc102976337)

[6.3.2. Special Cases 103](#_Toc102976338)

[6.3.3. Types of Score Reports 104](#_Toc102976339)

[6.3.4. Score Report Applications 105](#_Toc102976340)

[6.3.5. Criteria for Interpreting Individual Test Scores 105](#_Toc102976341)

[6.3.6. Criteria for Interpreting Score Reports 106](#_Toc102976342)

[References 107](#_Toc102976343)

[Appendix 6.A: Raw Score, Theta Score, and Scale Score Distributions of Students Taking Each Test 108](#_Toc102976344)

[Appendix 6.B: Demographic Summary of Students in Each Reporting Score Range 122](#_Toc102976345)

[Chapter 7: Psychometric Analyses 137](#_Toc102976346)

[7.1. Overview 137](#_Toc102976347)

[7.1.1. Summary of the Analyses 137](#_Toc102976348)

[7.1.2. Sample Used for the Analyses 138](#_Toc102976349)

[7.2. Classical Item Analyses 139](#_Toc102976350)

[7.2.1. Classical Item Difficulty Indices 139](#_Toc102976351)

[7.2.2. Item-Total Correlation 140](#_Toc102976352)

[7.2.3. Distractor Analyses 140](#_Toc102976353)

[7.2.4. Omission and Completion Rates 141](#_Toc102976354)

[7.2.5. Distribution of Item Scores 141](#_Toc102976355)

[7.2.6. Summary of Classical Item Analyses Flagging Criteria 141](#_Toc102976356)

[7.2.7. Classical Item Analyses Results Summary 142](#_Toc102976357)

[7.3. Differential Item Functioning Analyses 144](#_Toc102976358)

[7.3.1. Differential Item Functioning Procedure for Dichotomous Items 144](#_Toc102976359)

[7.3.2. Differential Item Functioning Procedure for Polytomous Items 145](#_Toc102976360)

[7.3.3. Classifications 145](#_Toc102976361)

[7.3.4. Items Exhibiting Significant Differential Item Functioning 148](#_Toc102976362)

[7.4. Item Response Theory Analyses 148](#_Toc102976363)

[7.4.1. Models 148](#_Toc102976364)

[7.4.2. Equating 149](#_Toc102976365)

[7.5. Classical Equating Analysis for High School 159](#_Toc102976366)

[7.5.1. Methodology of the Classical Linear Equating 160](#_Toc102976367)

[7.5.2. Empirical Analysis with the Classical Linear Equatings 161](#_Toc102976368)

[7.5.3. Evaluation of the Classical Equating Results 162](#_Toc102976369)

[7.6. Testing Time Analyses 162](#_Toc102976370)

[7.7. Reliability Analyses 163](#_Toc102976371)

[7.7.1. Marginal Reliability 164](#_Toc102976372)

[7.7.2. Standard Error of Measurement for Raw Scores 164](#_Toc102976373)

[7.7.3. Student Group Reliabilities and Standard Errors of Measurement 165](#_Toc102976374)

[7.7.4. Theta Scores Standard Error 166](#_Toc102976375)

[7.7.5. Scale Scores Conditional Standard Errors 167](#_Toc102976376)

[7.7.6. Decision Classification Analyses 168](#_Toc102976377)

[7.8. Validity Evidence 170](#_Toc102976378)

[7.8.1. Evidence in the Design of the CSA 170](#_Toc102976379)

[7.8.2. Evidence Based on the Test Content 172](#_Toc102976380)

[7.8.3. Evidence Based on Response Processes 173](#_Toc102976381)

[7.8.4. Evidence Based on Internal Structure 174](#_Toc102976382)

[References 176](#_Toc102976383)

[Accessibility Information 179](#_Toc102976384)

[Alternative Text for Equation 7.1 179](#_Toc102976385)

[Alternative Text for Equation 7.2 179](#_Toc102976386)

[Alternative Text for Equation 7.3 179](#_Toc102976387)

[Alternative Text for Equation 7.4 179](#_Toc102976388)

[Alternative Text for Equation 7.5 179](#_Toc102976389)

[Alternative Text for Equation 7.6 179](#_Toc102976390)

[Alternative Text for Equation 7.7 179](#_Toc102976391)

[Alternative Text for Equation 7.8 180](#_Toc102976392)

[Alternative Text for Equation 7.9 180](#_Toc102976393)

[Alternative Text for Equation 7.10 180](#_Toc102976394)

[Alternative Text for Equation 7.11 180](#_Toc102976395)

[Alternative Text for Equation 7.12 180](#_Toc102976396)

[Alternative Text for Equation 7.13 180](#_Toc102976397)

[Alternative Text for Equation 7.14 180](#_Toc102976398)

[Alternative Text for Equation 7.15 180](#_Toc102976399)

[Alternative Text for Equation 7.16 180](#_Toc102976400)

[Alternative Text for Equation 7.17 180](#_Toc102976401)

[Alternative Text for Equation 7.18 180](#_Toc102976402)

[Alternative Text for Equation 7.19 180](#_Toc102976403)

[Alternative Text for Equation 7.20 180](#_Toc102976404)

[Alternative Text for Equation 7.21 181](#_Toc102976405)

[Alternative Text for Equation 7.22 181](#_Toc102976406)

[Alternative Text for Equation 7.23 181](#_Toc102976407)

[Alternative Text for Equation 7.24 181](#_Toc102976408)

[Alternative Text for Equation 7.25 181](#_Toc102976409)

[Appendix 7.A: Classical Item Analyses 182](#_Toc102976410)

[Appendix 7.B: Item Response Theory Analyses Results 196](#_Toc102976411)

[Appendix 7.C: Classical Equating Analysis for High School 208](#_Toc102976412)

[Appendix 7.D: Testing Time Analysis 211](#_Toc102976413)

[Appendix 7.E: Reliability Analyses 215](#_Toc102976414)

[Chapter 8: Quality Control Procedures 222](#_Toc102976415)

[8.1. Quality Control of Item Development 222](#_Toc102976416)

[8.2. Quality Control of Test Form Development 222](#_Toc102976417)

[8.3. Quality Control of Test Materials 223](#_Toc102976418)

[8.3.1. Developing Computer-based Assessments 223](#_Toc102976419)

[8.3.2. Developing Test Administration Instructions 223](#_Toc102976420)

[8.3.3. Processing Test Materials 223](#_Toc102976421)

[8.4. Quality Control of Test Administration 223](#_Toc102976422)

[8.5. Quality Control of Scoring 224](#_Toc102976423)

[8.5.1. Development of Scoring Specifications 224](#_Toc102976424)

[8.5.2. Quality Control of Machine-Scoring Procedures 224](#_Toc102976425)

[8.5.3. Enterprise Score Key Management System Processing 225](#_Toc102976426)

[8.6. Quality Control of Psychometric Processes 225](#_Toc102976427)

[8.6.1. Development of Scoring Specifications 225](#_Toc102976428)

[8.6.2. Psychometric Analyses 225](#_Toc102976429)

[8.7. Quality Control of Reporting 226](#_Toc102976430)

[8.7.1. Exclusion of Student Scores from Summary Reports 227](#_Toc102976431)

[8.8. Quality Control of End-to-End Testing 227](#_Toc102976432)

[Reference 228](#_Toc102976433)

[Chapter 9: Continuous and Systematic Improvement 229](#_Toc102976434)

[9.1. Feedback for Continuous Improvement Survey 229](#_Toc102976435)

[9.2. Administration and Test Delivery 229](#_Toc102976436)

[9.2.1. Training and Communication 229](#_Toc102976437)

[9.3. Accessibility 230](#_Toc102976438)

[9.4. Item and Form Development 230](#_Toc102976439)

List of Tables

[Acronyms and Initialisms Used in the *California Spanish Assessment Technical Report* x](#_Toc103234908)

[Table 1.1 Total Number Tested in the 2020–2021 Administration 5](#_Toc103234909)

[Table 3.1 Number of CSA Items to Administer per Form 23](#_Toc103234910)

[Table 3.2 Item Types for the CSA 24](#_Toc103234911)

[Table 3.3 Number of Items Developed per Grade Level and Grade Band for the CSA 26](#_Toc103234912)

[Table 3.4 CSA Item Reviewer Qualifications 31](#_Toc103234913)

[Table 3.5 Status of Items After the 2019 Item Review Panel Meetings 32](#_Toc103234914)

[Table 3.6 ETS Operational Assessment Forms Review Process 35](#_Toc103234915)

[Table 3.7 Number of Forms and Items Reviewed Psychometrically 37](#_Toc103234916)

[Table 3.A.1 CSA Test Blueprint for Grade Span Three Through Five 41](#_Toc103234917)

[Table 3.A.2 CSA Test Blueprint for Grade Span Six Through Eight 42](#_Toc103234918)

[Table 3.A.3 CSA Test Blueprint for Grade Span Nine Through Twelve (High School) 43](#_Toc103234919)

[Table 4.1 Test-Taking Rates by California Region 46](#_Toc103234920)

[Table 4.2 CSA Test-Taking Rates by Grade Level and Grade Band 46](#_Toc103234921)

[Table 4.3 Demographic Student Groups to Be Reported 47](#_Toc103234922)

[Table 4.4 Summary of Accommodations and Designated Supports Used by Students, by Grade Level or Grade Band 59](#_Toc103234923)

[Table 4.5 Types of Appeals 66](#_Toc103234924)

[Table 4.6 Number and Types of Incidents Submitted in STAIRS in the 2020–2021 Administration 68](#_Toc103234925)

[Table 4.7 Number of Appeals Approved in STAIRS in the 2020–2021 Administration 69](#_Toc103234926)

[Table 4.8 Number of Appeals Rejected in STAIRS in the 2020–2021 Administration 69](#_Toc103234927)

[Table 4.A.1 Demographic Summary, Grade Three 71](#_Toc103234928)

[Table 4.A.2 Demographic Summary, Grade Four 73](#_Toc103234929)

[Table 4.A.3 Demographic Summary, Grade Five 75](#_Toc103234930)

[Table 4.A.4 Demographic Summary, Grade Six 77](#_Toc103234931)

[Table 4.A.5 Demographic Summary, Grade Seven 79](#_Toc103234932)

[Table 4.A.6 Demographic Summary, Grade Eight 81](#_Toc103234933)

[Table 4.A.7 Demographic Summary, High School 83](#_Toc103234934)

[Table 4.B.1 Special Services Summary, Grades Three Through Five 85](#_Toc103234935)

[Table 4.B.2 Special Services Summary, Grades Six Through Eight 87](#_Toc103234936)

[Table 4.B.3 Special Services Summary, High School 89](#_Toc103234937)

[Table 5.1 SSPI Recommendations for the Preliminary Threshold Scores and Three Scale Score Ranges on the CSA 95](#_Toc103234938)

[Table 6.1 Rules for Incomplete Tests 98](#_Toc103234939)

[Table 6.2 CSA Score Reporting Ranges by Grade Level and Grade Band 100](#_Toc103234940)

[Table 6.3 Mean and SD of Theta Scores and Scale Scores 101](#_Toc103234941)

[Table 6.4 Numbers and Percentages of Students in Score Reporting Ranges 102](#_Toc103234942)

[Table 6.A.1 Overall Raw, Theta, and Scale Score Distribution Grade Three 108](#_Toc103234943)

[Table 6.A.2 Overall Raw, Theta, and Scale Score Distribution for Grade Four 110](#_Toc103234944)

[Table 6.A.3 Overall Raw, Theta, and Scale Score Distribution for Grade Five 112](#_Toc103234945)

[Table 6.A.4 Overall Raw, Theta, and Scale Score Distribution for Grade Six 114](#_Toc103234946)

[Table 6.A.5 Overall Raw, Theta, and Scale Score Distribution for Grade Seven 116](#_Toc103234947)

[Table 6.A.6 Overall Raw, Theta, and Scale Score Distribution for Grade Eight 118](#_Toc103234948)

[Table 6.A.7 Overall Raw, Theta, and Scale Score Distribution for High School 120](#_Toc103234949)

[Table 6.B.1 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Three 122](#_Toc103234950)

[Table 6.B.2 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Four 125](#_Toc103234951)

[Table 6.B.3 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Five 127](#_Toc103234952)

[Table 6.B.4 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Six 129](#_Toc103234953)

[Table 6.B.5 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Seven 131](#_Toc103234954)

[Table 6.B.6 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Eight 133](#_Toc103234955)

[Table 6.B.7 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for High School, All Grade Levels 135](#_Toc103234956)

[Table 7.1 Sample Size by Test Form 138](#_Toc103234957)

[Table 7.2 Item Difficulty Distributions by Grade Level 142](#_Toc103234958)

[Table 7.3 Item-Total Correlation Distributions by Grade Level 143](#_Toc103234959)

[Table 7.4 Flagged Items Summary in Each Form by Grade Level 143](#_Toc103234960)

[Table 7.5 DIF Categories for Dichotomous Items 146](#_Toc103234961)

[Table 7.6 DIF Categories for Polytomous Items 146](#_Toc103234962)

[Table 7.7 Gender DIF Classifications Summary by Grade Level and Grade Band 147](#_Toc103234963)

[Table 7.8 Item Difficulty Parameter Distribution by Grade Level 151](#_Toc103234964)

[Table 7.9 Final Linking Summary 154](#_Toc103234965)

[Table 7.10 Evaluation of Anchor Set (Common Items) Between 2018–2019 and 2020–‍  
2021 155](#_Toc103234966)

[Table 7.11 Linked Item Parameter Results 155](#_Toc103234967)

[Table 7.12 IRT Summary *b*-value Estimates for All CSA Operational Items 156](#_Toc103234968)

[Table 7.13 Conversion of Theta Score to Reporting Scores by Grade Level or Grade Band Based on the 2018–2019 Administration (the Baseline) 158](#_Toc103234969)

[Table 7.14 Test Reliability of the Total Scores 165](#_Toc103234970)

[Table 7.15 Scale Score CSEMs at the Reporting Score Range Thresholds 168](#_Toc103234971)

[Table 7.16 Decision Accuracy for Reaching a Score Reporting Range Threshold 169](#_Toc103234972)

[Table 7.17 Decision Consistency for Reaching a Score Reporting Range Threshold 169](#_Toc103234973)

[Table 7.A.1 Item Difficulty Distributions by Domain 182](#_Toc103234974)

[Table 7.A.2 Item-Total Correlation Distributions by Domain 183](#_Toc103234975)

[Table 7.A.3 Item Difficulty and Item-Total Correlation by Item for Grade Three 184](#_Toc103234976)

[Table 7.A.4 Item Difficulty and Item-Total Correlation by Item for Grade Four 186](#_Toc103234977)

[Table 7.A.5 Item Difficulty and Item-Total Correlation by Item for Grade Five 188](#_Toc103234978)

[Table 7.A.6 Item Difficulty and Item-Total Correlation by Item for Grade Six 190](#_Toc103234979)

[Table 7.A.7 Item Difficulty and Item-Total Correlation by Item for Grade Seven 192](#_Toc103234980)

[Table 7.A.8 Item Difficulty and Item-Total Correlation by Item for Grade Eight 194](#_Toc103234981)

[Table 7.B.1 IRT Item Difficulty for Grade Three 196](#_Toc103234982)

[Table 7.B.2 IRT Item Difficulty for Grade Four 198](#_Toc103234983)

[Table 7.B.3 IRT Item Difficulty for Grade Five 200](#_Toc103234984)

[Table 7.B.4 IRT Item Difficulty for Grade Six 202](#_Toc103234985)

[Table 7.B.5 IRT Item Difficulty for Grade Seven 204](#_Toc103234986)

[Table 7.B.6 IRT Item Difficulty for Grade Eight 206](#_Toc103234987)

[Table 7.C.1 Summary of High School Forms in 2018–2019 and 2020–2021 208](#_Toc103234988)

[Table 7.C.2 Standard Error of Classical Equating 208](#_Toc103234989)

[Table 7.D.1 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Three 211](#_Toc103234990)

[Table 7.D.2 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Four 212](#_Toc103234991)

[Table 7.D.3 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Five 212](#_Toc103234992)

[Table 7.D.4 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Six 213](#_Toc103234993)

[Table 7.D.5 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Seven 213](#_Toc103234994)

[Table 7.D.6 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Eight 214](#_Toc103234995)

[Table 7.D.7 Total Testing Time (in Minutes) at Each Raw Score Interval—High School, All Grade Levels 214](#_Toc103234996)

[Table 7.E.1 Reliabilities and SEMs by Gender 215](#_Toc103234997)

[Table 7.E.2 Reliabilities and SEMs by Economic Status 215](#_Toc103234998)

[Table 7.E.3 Reliabilities and SEMs by Special Education Services 216](#_Toc103234999)

[Table 7.E.4 Reliabilities and SEMs by Attendance in US Schools 216](#_Toc103235000)

[Table 7.E.5 Reliabilities and SEMs by Spanish Instruction Status 217](#_Toc103235001)

[Table 7.E.6 Reliabilities and SEMs by English Proficiency 217](#_Toc103235002)

[Table 7.E.7 Reliabilities and SEMs by English Proficiency (Continued) 218](#_Toc103235003)

[Table 7.E.8 Reliability of Classification for Grade Three: Decision Accuracy 218](#_Toc103235004)

[Table 7.E.9 Reliability of Classification for Grade Three: Decision Consistency 218](#_Toc103235005)

[Table 7.E.10 Reliability of Classification for Grade Four: Decision Accuracy 218](#_Toc103235006)

[Table 7.E.11 Reliability of Classification for Grade Four: Decision Consistency 219](#_Toc103235007)

[Table 7.E.12 Reliability of Classification for Grade Five: Decision Accuracy 219](#_Toc103235008)

[Table 7.E.13 Reliability of Classification for Grade Five: Decision Consistency 219](#_Toc103235009)

[Table 7.E.14 Reliability of Classification for Grade Six: Decision Accuracy 219](#_Toc103235010)

[Table 7.E.15 Reliability of Classification for Grade Six: Decision Consistency 220](#_Toc103235011)

[Table 7.E.16 Reliability of Classification for Grade Seven: Decision Accuracy 220](#_Toc103235012)

[Table 7.E.17 Reliability of Classification for Grade Seven: Decision Consistency 220](#_Toc103235013)

[Table 7.E.18 Reliability of Classification for Grade Eight: Decision Accuracy 220](#_Toc103235014)

[Table 7.E.19 Reliability of Classification for Grade Eight: Decision Consistency 221](#_Toc103235015)

[Table 7.E.20 Reliability of Classification for High School: Decision Accuracy 221](#_Toc103235016)

[Table 7.E.21 Reliability of Classification for High School: Decision Consistency 221](#_Toc103235017)

List of Figures

[Figure 3.1 Sample CSA practice test item 24](#_Toc102976550)

[Figure 3.2 Example TIF 36](#_Toc102976551)

[Figure 7.C.1 Standard error of classical equating 210](#_Toc102976552)

Acronyms and Initialisms Used in the *California Spanish Assessment Technical Report*

|  |  |
| --- | --- |
| Term | Definition |
| 1PL | one-parameter logistic |
| 1PL-IRT | one-parameter logistic item response theory |
| ADEL | adult English learner |
| AERA | American Educational Research Association |
| AIS | average item score |
| ALD | achievement level descriptor |
| ALTRD | Assessment and Learning Technology Research & Development |
| APA | American Psychological Association |
| 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 |
| CCCSSeE | California Common Core State Standards en Español |
| *CCR* | *California Code of Regulations* |
| CDE | California Department of Education |
| CDS | county/district/school |
| CERS | California Educator Reporting System |
| COVID-19 | novel coronavirus disease 2019 |
| CR | constructed response |
| CRT | Content Review Tool |
| CSA | California Spanish Assessment |
| CSEM | conditional standard error of measurement |
| *DFA* | *Directions for Administration* |
| DIF | differential item functioning |
| *EC* | *Education Code* |
| EL | English learner |
| ELA | English language arts/literacy |
| ELPAC | English Language Proficiency Assessments for California |
| eSKM | Enterprise Score Key Management |
| GPCM | generalized partial credit model |
| HOSS | highest obtainable scale score |
| IBIS | Item Banking Information System |
| IEP | individualized education program |
| IFEP | initial fluent English proficient |
| IRT | item response theory |
| ISAAP | Individual Student Assessment Accessibility Profile |
| JAWS® | Job Access With Speech |

Table of Acronyms and Initialisms *(continuation)*

|  |  |
| --- | --- |
| Term | Definition |
| K | kindergarten |
| LEA | local educational agency |
| LOSS | lowest obtainable scale score |
| MC | multiple choice |
| MH | Mantel-Haenszel |
| MLE | maximum likelihood estimation |
| NCME | National Council on Measurement in Education |
| NEAT | nonequivalent groups with anchor test |
| OIB | ordered item booklet |
| OTI | Office of Testing Integrity |
| PAR | Psychometric Analysis & Research |
| PCM | partial credit model |
| QA | quality assurance |
| RFEP | reclassified fluent English proficient |
| RMSEA | root mean square error of approximation |
| RSD | ratio of standard deviations |
| SBE | State Board of Education |
| SCOE | Sacramento County Office of Education |
| SD | standard deviation |
| SEM | standard error of measurement |
| SFTP | secure file transfer protocol |
| SMD | standardized mean difference |
| SSID | Statewide Student Identifier |
| SSPI | State Superintendent of Public Instruction |
| SSR | Student Score Report |
| STAIRS | Security and Test Administration Incident Reporting System |
| TCC | test characteristic curve |
| TDS | test delivery system |
| TEI | technology-enhanced item |
| TIF | test information function |
| TOMS | Test Operations Management System |
| TVI | teacher of students with visual impairment |
| UAT | user acceptance testing |
| USC | United States Code |

**This page is intentionally left blank.**

## Introduction

This chapter provides an overview of the California Spanish Assessment (CSA), including background information, purpose of the assessment, intended population, testing window, organizations and systems involved, and an overview of the CSA technical report. Additionally, this chapter provides a summary of the impact of the novel coronavirus disease 2019 (COVID-19) pandemic to the test administration.

### Background

In October 2013, Assembly Bill 484 established the California Assessment of Student Performance and Progress (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 tests provide the foundation for the state’s school accountability system.

During the 2020–2021 administration, the CAASPP System comprised the following assessments:

* Smarter Balanced assessments and tools:
* Summative Assessments—Computer-based assessments for English language arts/literacy (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 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 grades
* California Alternate Assessments (CAAs) for ELA and mathematics in grades three through eight and grade eleven for students with the most significant cognitive disabilities
* Science assessments in grades five, eight, and high school (grade ten, eleven, or twelve; these are the California Science Test and the CAA for Science)
* The CSA, optional for eligible students in grades three through eight and high school (grades nine through twelve) and designed to measure a student’s Spanish competency in reading, writing mechanics, and listening

As part of the CAASPP System of assessments, the CSA was approved as California’s primary language assessment by the California State Board of Education (SBE). It was developed as an optional Spanish reading language arts assessment.

Development of the CSA started in September 2016 with the SBE’s approval of the high-level test design. Following the 2018 fall CSA field test, the first CSA operational administration was administered optionally to students seeking a measure of their Spanish reading language arts skills during spring 2019. The spring 2020–2021 administration marks the third operational year of the CSA, including the 2019–2020 administration, which was suspended in the middle of testing window because of the COVID-19 pandemic.

More background information about the CAASPP System can be found on the *CalEdFacts* web page on the California Department of Education (CDE) website.

### Test Purpose

This computer-based assessment for students in grades three through eight and high school was designed to measure a student’s Spanish reading language arts skills in reading, writing mechanics, and listening for the purposes of providing:

* student-level data in Spanish competency,
* aggregated data that may be used for evaluating the implementation of Spanish language arts programs at the local level, and
* a high school measure suitable to be used, in part, for the California State Seal of Biliteracy. Currently, the CSA does not meet the requirements identified in California *Education Code (EC)*, Section 51460(a) for the State Seal of Biliteracy.

The CSA is aligned with the translated and linguistically augmented version of the Common Core ELA standards (i.e., California Common Core State Standards en Español [CCCSSeE]). The CSA measures a student’s Spanish reading language arts skills in reading, writing mechanics, and listening based on the CCCSSeE.

### Test Design

The CSA is administered to eligible students in grades three through twelve. The CSA was delivered online with fixed forms—one general form, one accommodated form, and one braille form. The general form was delivered in three versions, with each having the same operational items but different embedded field test items. The accommodated and braille forms had accommodated resources specific to the targeted population for these forms. All items were machine scored.

### Intended Population

The population for the CSA comprises all California students in grades three through twelve who receive instruction in Spanish and who seek a measure that recognizes their Spanish-specific reading, writing mechanics, and listening skills.

During the first operational CSA administration in 2018–2019, the number of students taking the CSA varied significantly across different grade levels, from approximately 10,000 in grade three to fewer than 1,500 in grade twelve. This trend may be affected by the course of the COVID-19 pandemic on local educational agencies (LEAs); the emergence of Spanish-language programs within LEAs; and how LEAs exercise the option to assign students to take this optional assessment.

### Intended Use and Purpose of Test Scores

The results of tests within the CAASPP System, including the CSA, are used for two primary purposes as described in *EC* sections 60602.5(a) and (a)(4). (This was 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.”

In other words, results for tests within the CAASPP System are used for two primary purposes:

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 CSA for grades three through eight and high school are administered within a testing window pursuant to *California Code of Regulations*, Title 5 [5 *CCR*]*,* sections 855(a)(1), 855(a)(2), 855(b), and 855(c). The typical testing window starts in the middle of January and ends in the middle of July; however, LEAs can begin testing only once 66 percent of the LEA’s instructional year has been completed. For the 2020–2021 CSA administration, the state testing window opened on January 14, 2021, and ended on July 30, 2021. The July 30, 2021, date was an approved extension to the legislated testing window and offered LEAs more flexibility in testing, allowing schools more time for students to test.

A student could take the CSA within the testing window over as many days as required to meet the student’s needs (5 *CCR,* Education, Division 1, Chapter 2, Subchapter 3.75, Article 2*,* Section 855[a]). Similar to other CAASPP assessments, the CSA was untimed for test takers. The estimated duration for the assessment was approximately two hours, on average. Refer to section [*3.8 Test Assembly and Length*](#_Test_Assembly_and) for more information.

### Impact of the Novel Coronavirus Disease 2019 Pandemic

#### Remote Testing Flexibility

When the 2020–2021 school year began, LEAs were offering varying instructional options, with a substantial percentage offering only distance learning options. When the US Department of Education notified states that they should not expect waivers of the 2020–2021 annual state assessments, the CDE began exploring options for delivering its annual summative assessments. The approach taken by the CDE was to “allow flexibility for LEAs to utilize multiple test administration options to best meet the needs of students in response to the local context and to ensure the safety and health of students and LEA staff” (CDE, 2021b, p. 4). That flexibility offered LEAs two options for testing students using the CSA:

1. Test in person, with both students and test administrators co-located in the same room at a school or other secure location and following physical distancing guidelines.
2. Test remotely, with students and test administrators located at different physical locations. The test administrator would monitor students’ progress throughout the test by using remote monitoring tools connected to the test delivery system (TDS).

For some of the larger-volume assessments, ETS investigated test score validity and, in particular, whether the test options used in the 2020–2021 administration impacted student testing performance and student testing experience that, in turn, impacted the test score interpretation. The CSA was not included in the analyses because of the small testing sample.

Refer to subsection [*4.1.1 Remote and In-Person Testing*](#_Remote_and_In-Person) for detailed information on the administration.

#### Test-Taking Rates in the 2020–2021 Administration

The impact of the COVID-19 pandemic on the school year, and whether students returned to in-person learning, varied greatly across the state. The COVID-19 pandemic also impacted how students were tested, as remote administration was introduced and widely used.

Table 1.1 lists the total number of students who took the CSA during the 2020–2021 administration as well as the totals by test location. In-person testing involved arranging both students and test administrators to be co-located in the same room at a school or other secure location while following physical distancing guidelines. Remote testing involved students and test administrators being present at different physical locations; the test administrator monitored students’ progress throughout the test by using remote monitoring tools connected to the TDS.

Table 1.1 Total Number Tested in the 2020–2021 Administration

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Total Registered | Total Started | Total Completed | Completed Remotely | Completed in Person | Completed Mixed |
| Grade 3 | 3,066 | 1,223 | 1,213 | 609 | 582 | 22 |
| Grade 4 | 2,778 | 1,054 | 1,052 | 568 | 474 | 10 |
| Grade 5 | 2,371 | 919 | 916 | 459 | 442 | 15 |
| Grade 6 | 1,662 | 952 | 948 | 590 | 325 | 33 |
| Grade 7 | 1,292 | 735 | 729 | 521 | 159 | 49 |
| Grade 8 | 1,137 | 572 | 572 | 440 | 84 | 48 |
| High school | 1,542 | 262 | 253 | 213 | 38 | 2 |
| **Total** | **13,848** | **5,717** | **5,683** | **3,400** | **2,104** | **179** |

Completion means a student answered at least 10 CSA questions and was eligible to receive test scores. If a student answered at least 1 question but less than 10 questions, the student was assigned a lowest obtainable scale score. Note that, in table 1.1, the *Completed Mixed* column includes students who tested in a combination of remote and in-‍person locations. Results show that about 38 to 40 percent of students registered completion for the CSA in grades three through five, 50 to 57 percent in grades six through eight, and 16 percent in high school.

The number of students registered to take the 2020–2021 administration was much lower than in 2018–2019. The number of students who took the CSA during the 2020–2021 administration was about 13 percent of the number of students who took the CSA in 2018–‍2019 for grades three through five, 20 percent for grades six through eight, and 6 percent for high school. Results also show that a higher percentage of students tested remotely than tested in person. As the grade level increases, the percentage of students who tested remotely is higher.

### Significant Developments in 2020–2021

#### Remote Testing

LEAs were heavily impacted by the school and business shutdowns as a result of the COVID-19 pandemic and the impacts carried over into the beginning of the 2020–2021 CAASPP test administration. The CDE and ETS worked together to develop a way to allow schools to administer the CSA to students remotely that included the following updates:

* The Student Testing Interface was enhanced to enable students and test administrators to access the tests securely when the student could not be tested at the school or LEA in person.
* The secure browser was updated with the following new features:
* One-on-one chatting or voice or video calls with the test administrator
* A raise-hand feature to signal the test administrator for attention
* Ability to approve a request from the test administrator to share a screen
* The Test Administrator Interface was updated with the following new features:
* Selection of an option to indicate whether the test session was in person or remote (for tracking and analysis purposes)
* One-on-one chatting or voice or video calls with a student
* One-way broadcast of messages to the students in the test session
* Multiple monitoring options (gallery or list views of testing students)
* Ability to request the student to share a screen
* Remote testing instructions and scripts were developed for test administrators, providing remote logon instructions.
* Remote testing administration videos were created to show test administrators how to give a remote test; how to monitor a remote test session; and how to schedule, start, and stop a remote test session.
* A video was created for students and parents/guardians to introduce them to remote testing.

#### Accessibility Resources

The following accessibility resource–related updates were made:

* CAASPP Matrix One was combined with the English Language Proficiency Assessments for California Matrix Four to create the California Assessment Accessibility Resources Matrix that serves both testing programs.
* Text-to-speech can be established independently of assigned tests
* An increased number of prefetched items for braille embossing were sent to the embosser prior to the student’s reaching the item when the auto emboss feature was enabled for tests presented in braille. This feature enhancement allowed for items in fixed-form tests to be printed or embossed prior to the student’s reaching the item in the test, thus speeding up the testing time for students with the braille accommodation.
* A break was considered both an embedded and a non-embedded universal tool.
* The description of the scratch paper resource indicated that students could create graphic organizers.

#### Reuse of the 2019–2020 Test Forms

Because of the suspension of the 2019–2020 testing window resulting from the COVID-19 pandemic, not enough testing samples were collected to support the item analysis activities for items that were to be field-tested in 2019–2020. Consequently, no new forms could be developed on the basis of the incomplete data from 2019–2020 for the 2020–2021 administration. All 2019–2020 forms, including the general forms, accommodated forms, and braille forms, as well as the embedded field test items, were reused for the 2020–2021 administration.

In addition, the embedded field-tested items were implemented in the reused 2019–2020 operational forms for the 2020–2021 administration. Note that no embedded field test items had been included in the 2018–2019 operational forms.

#### High School Postequating Procedures

Beginning in the 2018–2019 administration, the one-parameter logistic item response theory (IRT) measurement model was implemented in the CSA psychometric analyses for item calibration, equating, and scaling. For the 2020–2021 administration, IRT postequating procedures were used for grade three through grade eight but not for high school because of the insufficient testing sample (262 students in total took the CSA for high school). Instead, the classical linear equating method was used for the high school grade band to link the 2020–2021 data to the 2018–2019 base scale. Refer to section [*7.5 Classical Equating* *Analysis for High School*](#_Classical_Equating_Analysis) for details.

### Groups and Organizations Involved with the CSA

#### California State Board of Education

The 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 as well as the state’s Public School Accountability Act that measures the academic performance and progress of schools on a variety of academic metrics (CDE, 2021d).

#### California Department of Education

The CDE oversees California’s public school system, which is responsible for the education of more than 6,160,000 children and young adults in more than 10,500 schools.[[1]](#footnote-2) 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, 2021c).

#### California Educators

A variety of California educators and content experts, including teachers and school administrators—who were selected on the basis of their qualifications, experience, demographics, and geographic locations—were invited to participate in the various aspects of the assessment process prior to the current administration. This included defining the purpose and scope of the assessment, assessment design, item development, and standard setting.

#### Contractors

##### Primary Contractor—ETS

The CDE and the SBE contract with ETS to develop and administer the CSA and report CSA results. As the primary contractor, ETS has the overall responsibility for working with the CDE to implement and maintain an effective assessment system and coordinating 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
* Providing a tiered help desk support system for LEAs
* Developing all CSA items
* Constructing, producing, and controlling the quality of CSA forms and related test materials, including *Directions for Administration*
* 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 the California Educator Reporting System (CERS)
* Processing student test assignments
* Producing and distributing score reports
* Developing a summary score reporting website that can be viewed by the public
* Completing all psychometric procedures

##### 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;
* 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 direct funded charter schools
* Developing informational materials
* Recruiting and providing logistics for educator meetings

### 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, which, for the CSA, included but were not limited to the following:

* Managing test administration windows
* Assigning and managing the CSA test administrator user role
* Managing student test assignments and accessibility resources
* Viewing and downloading reports
* Providing a platform for authorized user access to secure materials such as 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, including LEA and school hierarchy data, from the California Longitudinal Pupil Achievement Data System (CALPADS) via a 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.”[[2]](#footnote-3) LEA staff involved in the administration of the CAASPP, such as LEA CAASPP coordinators, test site 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 set up the LEA’s test administration window; a test administrator cannot download student reports. A description of user roles is explained more extensively in the *2020–2021 CAASPP Online Test Administration Manual* (CDE, 2021a).

#### 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 tests and monitor student testing;
* the Student Testing Interface, on which students take the test using the secure browser; and
* the secure browser, the web-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

The publicly available practice and training tests were provided to LEAs to prepare students for administration of the CSA. The practice tests, offered for each of grades three through eight and the high school grade band; and training tests, offered by grade band (grades three through five, grades six through eight, and high school), were provided to LEAs to prepare students and LEA staff for the summative assessment. These tests simulated the experience of the computer-based CSA. Unlike the summative assessments, the practice and training tests did not assess standards, gauge student success on the operational assessment, or produce scores. Students, teachers, and the public could access them using a web browser, although accessing them through the secure browser permitted students to take the tests using the text-to-speech embedded accommodation and to test assistive technology. Both the practice and training tests were offered in general versions and accommodated versions for students with visual impairment. Test administrators could access scoring guides that described related scoring considerations.

The purposes of the practice and training tests are to

* allow students and administrators to become familiar with the user interface and components of the TDS as well as with the process of starting and completing a testing session;
* allow students and administrators to experience a grade-level assessment, grade-specific items and difficulty levels, and the format and structure of an operational assessment; and
* 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.

Details on practice and training tests are presented in section [*4.5 Practice and Training Tests*](#_Practice_and_Training).

#### California Educator Reporting System

CERS is the system used by LEAs to view preliminary student results from the CAASPP assessments. The primary purpose of CERS is to provide educators and administrators with access to timely test results data for individual students and groups of students.

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

### Overview of the Technical Report

This technical report addresses the characteristics of the CSA administered in spring 2021 and contains eight additional chapters as follows:

* [Chapter 2](#_An_Overview_of) presents an overview of the processes involved in a testing cycle for the CSA. This chapter includes item development, test assembly, test administration, scoring, reporting, psychometric analyses, and standard setting. The details on each stage in the testing process will be presented in the subsequent chapters.
* [Chapter 3](#_Item_Development_and_1) discusses the test blueprint, item development, and detailed procedures of test assembly for the 2020–2021 administration.
* [Chapter 4](#_Chapter_4:_Test) details the processes involved in the administration of the CSA. It also describes the procedures followed by ETS to maintain test security throughout the test administration process.
* [Chapter 5](#_Standard_Setting) summarizes the standard setting process that established the base-year score reporting ranges. Details include the achievement level descriptors, an overview of the standard setting methodology, and the process to establish the threshold scores that define the score reporting ranges for the CSA. These standard setting processes were based on student testing results from the 2018–2019 administration.
* [Chapter 6](#_Scoring_and_Reporting) summarizes the types of scores and score reports that are produced at the end of each administration of the CSA.
* [Chapter 7](#_Psychometric_Analyses_and) summarizes the results of the psychometric analyses for CSA administration, including classical item analyses, testing time analyses, test completion analyses, and differential item functioning analyses, as well as IRT calibration, postequating, and scaling. Test reliability and reliability analysis results are also reported.
* [Chapter 8](#_Quality_Control_Procedures) highlights the quality control processes used at various stages of administration of the CSA.
* [Chapter 9](#_Continuous_and_Systematic) discusses the various procedures used to gather information to improve the CSA 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 855.

California Department of Education. (2021a). *CAASPP online test administration manual.* Sacramento, CA: California Department of Education.

California Department of Education. (2021b). *California State Board of Education January 2021 agenda item #3.* California Department of Education website.

California Department of Education. (2021c, August). *Organization.* California Department of Education website.

California Department of Education. (2021d, November). *State Board of Education responsibilities.* California Department of Education website.

## An Overview of CSA Processes

This chapter provides a brief description of the processes implemented by ETS during a typical, full testing cycle for the California Spanish Assessment (CSA), including item development, test design, test administration, and scoring. The details on each step will be presented in the subsequent chapters.

### Item Development

ETS reused the 245 field test items developed in 2019–2020 across the six grade levels and the grade band (i.e., grades three through eight and high school, 35 items per grade) for the 2020–2021 administration—the third operational administration—and delivered them to the California Department of Education (CDE) via the ETS Item Banking Information System (IBIS). The total number of machine-scorable items for operational use (245) was greater than the number to be field-tested (228) in the 2020–‍2021 administration because overage was built in.

The developed items were designed to be engaging to the student population and represented a wide variety of item types. All items for the CSA field tests were developed in accordance with the *ETS Standards for Quality and Fairness* (ETS, 2014) across all phases of item and test development. While under initial development, the assessment materials, including items, passages, and listening stimuli, were stored on password-protected ETS computers and secure internal network drives. Audio recordings were produced as electronic audio files and delivered securely to the CDE for review.

All secure documents needed for CDE review that were not available in IBIS were delivered to the CDE via a secure file transfer protocol server.

#### Item Format

The CSA includes the following primary computer-based item formats:

* **Selected-response items—**Students are instructed to select one or more choices. Most CSA items have three or four options; a few items have more than four options.
* **Technology-enhanced items (TEIs)—**Technology beyond simple option selection is incorporated in some items.

Detailed information on item format is included in subsection [*3.3.3 Item Types and Features*](#_Item_Types_and) in [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1). All items included in the CSA 2020–‍2021 forms were machine-scorable.

#### Item Specifications

The CSA item specifications provide descriptions of item characteristics that are intended to measure each content standard consistently. They were developed based on the California Common Core State Standards en Español guidelines. During item development, item writers were provided CSA item specifications and a CSA style guide that contained detailed information about the consistency in item development and item review processes. Refer to subsection [*3.3.1 Item Specifications*](#_Item_Specifications) in [chapter 3](#_Item_Development_and_1) for detailed information about item specifications.

#### Item Banking

The CSA test forms administered in 2019–2020 were reused during the 2020–2021 administration. Following a typical operational administration of the CSA, the operational forms across all grades are refreshed for future administrations. To support the proposed refresh rates of 20 percent for grades three through eight and 35–50 percent for high school, it is necessary to build an item bank where the content and statistical attributes of each item are included. All the items in the item bank need to be calibrated and linked onto the base scales that were established after the CSA’s first operational administration, which occurred during the 2018–2019 California Assessment of Student Performance and Progress (CAASPP) administration.

The test forms used for the 2020–2021 CSA administration included operational and field test items. As a standardized procedure, after each CSA administration, initial item analyses are implemented, and the results are reviewed by ETS Psychometric Analysis & Research (PAR) and Assessment and Learning Technology Research & Development (ALTRD) staff, who provided recommendations to the CDE on whether the items should be included or excluded from the calibrations. Decisions were made in consultation with the CDE; details of this process are in section [*7.2 Classical Item Analyses*](#_Classical_Item_Analyses).

Next, the operational items are calibrated and postequated to the base scales established in the 2018–‍2019 administration. Refer to section [*7.4 Item Response Theory Analyses*](#_Item_Response_Theory) for calibration and linking. Final item analyses are usually conducted following the calibration and linking step after the testing window is closed.

Content experts from ETS and the CDE, as well as selected California educators, review the associated item statistics and evaluate the performance of items during the annual data review meeting. Particularly, they pay attention to reviewing the flagged items—those whose statistics fell beyond expected ranges—and work to provide plausible explanations for these particular items based on their knowledge of the student population.

With the CDE’s approval, the operational items and field test items, together with their statistical information, are entered into the item bank for form assembly in future administrations. It is expected that more new items will be developed, field-tested, and entered into the item bank for future administrations. Over time, the item bank will expand gradually to support the rate of refresh.

### Test Assembly

For the 2019–2020 operational forms that were reused for 2020–2021 CSA testing, the ETS ALTRD team used 52 items that were previously administered during the 2018–2019 operational administration and the fall 2018 field test. The CDE reviewed the 2020–2021 forms in IBIS before they were configured in the test delivery system (TDS) by Cambium Assessment, Inc.

Ten field test items were embedded into each of the 2020–2021 forms for grades three through eight, so each grade level’s test form was composed of 62 items. Similarly, 16 field test items were embedded into the high school 2020–2021 operational form, so that test form was composed of 68 items. Additional information about CSA test assembly can be found in [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1).

Note that test assembly activities were not conducted for the 2020–2021 administration, because the sample size in the 2019–2020 administration was too small to conduct item analysis. As a result, the 2019–2020 forms were reused in the 2020–2021 administration.

In a typical test administration year, when forms are assembled and not reused, psychometric criteria are specified for the test form review before the test administration. The psychometric guidelines of item selection and form building are developed and shared with ALTRD in the previous spring, before the start of the test assembly process for the next test administration.

ETS content staff and PAR staff review the assembled forms thoroughly in regard to the following aspects of the operational forms:

* Coverage of blueprints
* Overall test design and statistical properties
* Statistical properties of individual items

Details of the psychometric criteria of form review are included in section [*3.9 Test Production Process*](#_Test_Production_Process). Note that test assembly activities were not conducted for the 2020–‍2021 administration because the sample size for tests administered in 2019–2020, which were reused in 2020–2021, was too small for analysis.

### Test Administration

The CSA was administered online using the secure browser and TDS, ensuring a secure, confidential, standardized, consistent, and appropriate administration for students. Additional information about the administration of the CSA can be found in [*Chapter 4: Test Administration*](#_Chapter_4:_Test).

#### Test Security and Confidentiality

All tests within the CAASPP System are secure. For the CSA, every person with access to test materials maintained the security and confidentiality of the tests. ETS’ internal Code of Ethics requires that all test information, including tangible materials (such as test questions and test results), confidential files, processes, and activities were kept secure. To ensure security for all tests 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 [*4.6.1 ETS’ Office of Testing Integrity*](#_ETS’_Office_of).

In the pursuit of enforcing secure practices, ETS strives to safeguard the various processes involved in a test development and administration cycle. The practices related to each of the following security processes are discussed in detail in section [*4.6 Test Security and Confidentiality*](#_Test_Security_and):

* [Procedures to maintain standardization of test security](#_Procedures_to_Maintain)
* [Test security monitoring](#_Test_Security_Monitoring)
* [Security of electronic files using a firewall](#_Security_of_Electronic_1)
* [Transfer of scores via secure data exchange](#_Transfer_of_Scores)
* [Data management in the secure database](#_Data_Management_in)
* [Statistical analysis on secure servers](#_Statistical_Analysis_on)
* [Student confidentiality](#_Student_Confidentiality)
* [Student test results](#_Student_Test_Results)

#### Procedures to Maintain Standardization

ETS takes all necessary measures to ensure the standardization of CSA administration. The measures for standardization include, but are not limited to, the aspects described in these subsections.

##### Test Administrators

The CSA is 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 subsection [*4.6.2 Procedures to Maintain Standardization of Test Security*](#_Procedures_to_Maintain).

Staff at local educational agencies (LEAs) involved in CSA administration include LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators. The responsibilities of each of the staff members are described in the *CAASPP Online Test Administration Manual* (CDE, 2021c).

##### Test Directions

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 is a web-based manual that provides test administration procedures and guidelines for LEA CAASPP coordinators and CAASPP test site coordinators, as well as the script and *Directions for Administration* to be followed exactly by test administrators during a testing session (CDE, 2021c). (Refer to [*4.2.4.2 CAASPP Online Test Administration Manual*](#_5.4.4.2_CAASPP_Online) in [chapter 4](#_Chapter_4:_Test) for more information.)
* ***Spring Administration Information for Educators—***This is a web-based manual that was developed in response to the need to test students remotely. It supplements the *CAASPP Online Test Administration Manual* (CDE, 2021d). (Refer to [*4.2.4.3 Spring Administration Information for Educators*](#_Spring_Administration_Information) in [chapter 4](#_Chapter_4:_Test) for more information.)
* ***CAASPP and English Language Proficiency Assessments for California (ELPAC) Test Operations Management System (TOMS) User Guide*—**This is a web-based manual that provides instructions for TOMS, allowing LEA staff, including LEA CAASPP coordinators and CAASPP test site coordinators, to perform a number of tasks, including setting up test administrations, adding and managing users, assigning tests, and configuring computer-based student test settings (CDE, 2021b). (Refer to [*4.2.4.4 CAASPP and English Language Proficiency Assessments for California Test Operations Management System User Guide*](#_CAASPP_and_English) in [chapter 4](#_Chapter_4:_Test) for more information.)

### Fairness and Accessibility

There are several procedures in place to ensure that the CSA is fair and accessible to all test takers. This section provides information on the available accessibility resources to use with the CSA. Additionally, the differential item functioning (DIF) analysis used to identify items that may function differently across groups of examinees (e.g., gender) is also discussed briefly.

California public school students in grades three through twelve participate in the CAASPP System of assessments, including students with disabilities and English learner students. Additional resources are sometimes needed for these students. The CDE provides a full range of assessment resources for all students.

#### Universal Tools, Designated Supports, and Accommodations

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, 2020).

**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 CSA on the basis of student preference and selection.

**Designated supports** are available 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 CSA 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 only on approval by the CDE.

[Appendix 4.B](#_Appendix_4.B:_Special) presents counts and percentages of students assigned designated supports, accommodations, and unlisted resources for the 2020–2021 CSA administration. Table 4.B.1 presents data for grades three through five, table 4.B.2 presents data for grades six through eight, and table 4.B.3 presents data aggregated data for high school. Table 4.4 summarizes the number of students who were assigned and actually used the accessibility resources.

##### Selection

The full list of the universal tools, designated supports, and accommodations that are used in CAASPP computer-based assessments, including the CSA, is documented in the Accessibility Matrix (CDE, 2020). Most embedded and non-embedded universal tools, designated supports, and accommodations listed in parts 1, 2, and 3 of the Accessibility Matrix are available 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 that are available. School-level personnel, IEP teams, and Section 504 teams used the Accessibility Matrix when deciding how best to support a student’s or students’ test-taking experience.

In the selection of universal tools, designated supports, and accommodations deemed necessary for individual students, the CDE Accessibility Matrix follows the guidelines outlined in the Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines* (“*Guidelines*”) (Smarter Balanced, 2020).[[3]](#footnote-4) The *Guidelines* apply to all students and promote an individualized approach to the implementation of assessment best practices. The *Guidelines* are intended to provide policy regarding universal tools, designated supports, and accommodations. Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014),provides suggestions for implementation of these resources.

##### 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 CAASPP test site coordinator, either through individual assignment in 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 Individual Student Assessment Accessibility Profile (ISAAP) Tool. The ISAAP Tool could be used by LEAs in conjunction with the *Guidelines* and the CAASPP and ELPAC Accessibility Guide for Online Testing (CDE, 2021a), 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 [*1.10 Systems Overview and Functionality*](#_Systems_Overview_and) in [*Chapter 1: Introduction*](#_Chapter_1:_Introduction) for more details regarding the TDS.

##### Delivery

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 the computer-based CAASPP assessments. Examples of embedded resources include the braille language resource, color contrast, and closed-captioning for listening items.

Non-embedded resources are not part of the technology platform for the computer-administered CAASPP tests. Examples of non-embedded resources include magnification, noise buffers, and the use of a scribe.

Refer to subsection [*4.4.1 Accessibility Resource Categories*](#_Universal_Tools,_Designated) for a detailed description of the accessibility resources available to students taking the CSA.

#### Description of Differential Item Functioning

DIF analyses are conducted to detect possible test bias by locating items for which one group of students performs significantly better than another group. DIF is a collection of statistical methods used to recognize whether performance varies across different groups of examinees (e.g., male vs. female). If an item performs differentially across student groups when students are matched on ability, the item may be measuring something other than the intended construct. Therefore, it is important to identify items flagged for DIF. Content experts and bias and sensitivity experts from diverse backgrounds review these DIF-flagged items and determine the sources and meanings of performance differences. Refer to section [*7.3 Differential Item Functioning Analyses*](#_Differential_Item_Functioning) for additional information about DIF.

### Scores

Individual student scores were reported for the 2020–2021 CSA administration. The CSA contained traditional multiple-choice (MC) items and TEIs. The MC items and TEIs were machine-scored through the TDS. The CSA total test raw scores equal the sum of students’ scores on the operational test items.

In a typical test administration, total test raw scores on each CSA are converted to three-digit scale scores using the scaling process described in [*Chapter 7: Psychometric Analyses*](#_Psychometric_Analyses_and).

#### Estimating Ability Scores

The item response theory (IRT) inverse test characteristic curve method (Stocking, 1996)—where the student’s ability value is estimated to be the value for which the expected number-correct score is equal to the student’s number-correct score—is used to estimate students’ overall ability parameters. For the purpose of reporting, students’ ability estimates (theta scores) are then expressed in three-digit scale scores by applying an appropriate linear transformation for each grade level in a typical test administration.

Student performance on the reporting scale is designated into one of three score reporting ranges. For information regarding score specifications and the establishment of score-reporting scales, refer to [*Chapter 6: Scoring and Reporting*](#_Scoring_and_Reporting)*.* For information regarding CSA score reporting ranges, refer to [*Chapter 5: Standard Setting*](#_Standard_Setting) for a description of the process used to develop preliminary threshold scores.

#### 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. It also provides a secure means for LEA CAASPP coordinators to download Student Score Reports as PDF files.

CAASPP 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 the CSA at the student, school, and LEA levels. CERS also provides an individual score report. Refer to subsection [*6.3.1 Online Reporting*](#_Online_Reporting) for details about TOMS and CERS and subsection [*6.3.3 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, CSA scores for a given grade were aggregated and generated 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 aggregated scores were presented for all students or selected demographic student groups.

Aggregated scores were generated by combining student scores. They could be created by combining results at the state, LEA or direct funded charter school, or school level; combining for all students; or by combining results for students who represent selected demographic student groups.

Aggregation procedures used to present CSA results are described in section [*6.2 Overview of Score Aggregation Procedures*](#_Overview_of_Score). Aggregated scores that summarize student performance by grade for selected groups of students are provided in a typical administration. The tables show the number of students with valid scores in each group, scale score means and standard deviations, and percentage in the score reporting ranges.

Aggregated results by demographic variables are described in table 4.A.1 through table 4.A.7 in [appendix 4.A](#_Appendix_4.A:_Demographic), where students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, length of enrollment in US schools reported in the California Longitudinal Pupil Achievement Data System, self-reported Spanish-language program type, and self-reported percentage of instruction in Spanish. Definitions for the demographic student groups included in these tables are provided in table 4.1.

### Psychometric Analyses

In a typical administration, psychometric analyses are conducted on the data from the CSA, including classical item analyses, DIF analyses, IRT calibration and linking, testing time analyses, and reliability analyses. The results of these analyses support understanding of item performances and internal structure and provide validity evidence for both response processes and scoring. Detailed descriptions of these analyses are presented in [*Chapter 7: Psychometric Analyses*](#_Psychometric_Analyses_and). In the 2020‒2021 administration, all psychometric analyses were conducted and are reported in this technical report.

### Standard Setting

Standard setting was conducted and score reporting ranges were established after the 2018–‍2019 administration, which was the first operational administration of the CSA. Student performance on the reporting scale was designated into one of three ranges as follows:

1. Score Reporting Range 1
2. Score Reporting Range 2
3. Score Reporting Range 3

For information regarding development of the CSA threshold score recommendations, refer to [*Chapter 5: Standard Setting*](#_Standard_Setting).

### References

California Department of Education. (2020). *California assessment accessibility resources matrix.* California Department of Education website.

California Department of Education. (2021a). *CAASPP and ELPAC* *accessibility guide for online testing.* Sacramento, CA: California Department of Education.

California Department of Education. (2021b). *CAASPP and ELPAC Test Operations Management System user guide*. Sacramento, CA: California Department of Education.

California Department of Education. (2021c). *CAASPP online test administration manual.* Sacramento, CA: California Department of Education.

California Department of Education. (2021d). *Spring administration information for educators.* 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). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes.

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. 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.

## Item Development and Form Assembly

This chapter describes the detailed procedures of item development and test form assembly for the California Spanish Assessment (CSA). Those processes include the ones that are entirely internal to ETS and those that are conducted in coordination with the California Department of Education (CDE) and Cambium Assessment, Inc. (CAI). In particular, development of new item types and features that differ from traditional item types are described.

CSA item development incorporated innovations and best practices from national language assessments. California Spanish-language teachers assisted in creating these items, and item review meetings with California educators were instrumental in determining the complexity of content and grade-level appropriateness. This section describes how California educators were selected and the process used to develop new items.

Note that some item development activities were scaled back for the 2020–2021 administration as an impact of the ongoing novel coronavirus disease 2019 (COVID-19) pandemic.

### Overview

The 2020–2021 forms were reused from the 2019–2020 administration. ETS chose 364 items, which were from the 2018–2019 forms or previously field-tested items, for use on operational assessments across the seven grade levels—52 items on each general form for grades three through eight and high school—and submitted them to the CDE by way of the ETS Item Banking Information System (IBIS). Approximately 20 percent of the items were refreshed from the 2018–2019 spring administration for grades three through eight. Thirty-five percent of high school items were refreshed. This allowed items that were field-tested earlier in fall 2018 to be used operationally for the first time while maintaining reliability via repeating operational items from the 2018–2019 spring administration.

The developed items were designed to be engaging to the student population and represented a wide variety of item types. All items for the operational CSA were developed in accordance with the *ETS Standards for Quality and Fairness* (ETS, 2014) across all phases of item and test development. While under initial development, the assessment materials, including items, passages, and listening stimuli, were stored on password-protected ETS computers and secure internal network drives. Audio recordings were produced as electronic audio files and delivered securely to the CDE for review.

All secure documents needed for CDE review that were not available in IBIS were delivered to the CDE via the secure file transfer protocol server.

### Test Blueprint

Each operational assessment form contained items that approximate the proportions in the test blueprint. The test blueprint for the CSA provides the proposed number of items to be included in an operational assessment for each language arts domain assessed in grades three through eight and high school (CDE, 2017).

Table 3.1 shows the distribution of the operational assessment items by domain and grade level or grade band. [Appendix 3.A](#_Appendix_3.A:_CSA) presents the overview of the CSA blueprint by grade span.

Table 3.1 Number of CSA Items to Administer per Form

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Domain | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | High School |
| Listening | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| Reading | 24 | 24 | 24 | 24 | 24 | 24 | 24 |
| Writing Mechanics | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| **Total Items** | **52** | **52** | **52** | **52** | **52** | **52** | **52** |

### Guidelines

Each item for the CSA was developed through a comprehensive development cycle and designed to conform to principles of item writing defined by ETS. Each item in the CSA operational item bank was developed to measure a specific California Common Core State Standards en Español (CCCSSeE). In addition, guidelines for style, fairness, and bias and sensitivity helped item developers and reviewers to ensure consistency across the item development process.

#### Item Specifications

ETS maintained item specifications for the CSA that describe the characteristics of items written to measure the CCCSSeE that, in turn, provide evidence for the CSA’s reading, writing mechanics, and listening domains. Using the item specifications helped ensure that all items developed for the CSA measured standards consistently. Item writing assignments were guided by the CSA blueprints, developed in consultation with the CDE.

The specifications include

* a description of best practices for item writing:

universal design,

bias and sensitivity avoidance,

cognitive level,

anatomy of an item,

item types and characteristics,

a general list of elements to avoid, and

stand-alone items;

* information about passages used to assess CSA domains;
* a description of standards used for items associated with reading passages, writing mechanics passages, and listening passages;
* a full statement of each standard featured on the CSA blueprint; and
* sample item stems at each grade level for some standards.

#### Item Format

CSA items were developed with the understanding that students who are able may select responses using a mouse, touchscreen, or other supported input device. The majority of items were presented in a split-screen format, with a “stimulus” on the left side of the screen and the item to be answered on the right. The stimulus was usually a passage. This is shown in figure 3.1.

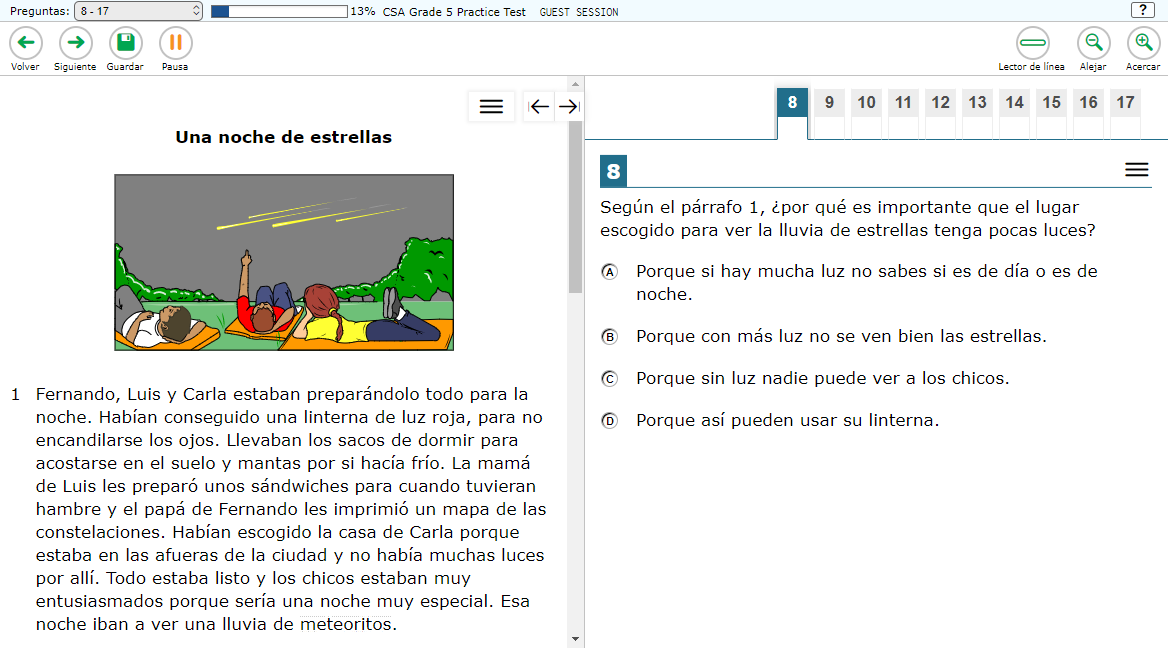


Figure 3.1 Sample CSA practice test item

A selected number of items had a multimedia stimulus, either a short audio file, or, for students with visual impairment, alternative text.

Items developed for the CSA were scored as being worth one point or two points.

#### Item Types and Features

ETS developed a variety of technology-enhanced item (TEI) types that required the student respond to a question in different ways from typical selected-response items. Items could contain a stimulus (e.g., a passage, audio, or image).

Students responded to TEIs by typing an answer, completing a grid, dragging a response to a designated area, using a drop-down list selection, or selecting multiple areas in a graphic (also known as “hot spots”). As currently configured, all CSA item types were machine-scored in the test delivery system (TDS); there was no human scoring of CSA test questions.

Table 3.2 lists item types used in the CSA. Response types indicated with an asterisk (\*) are TEIs.

Table 3.2 **Item Types for the CSA**

|  |  |  |
| --- | --- | --- |
| Response Type | Item Type | Description |
| Multiple choice (MC) single select | MC | The item generally consists of a stem and list of choices; the test taker can select only one choice to respond. This may also include a stimulus. |

Table 3.2 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Response Type | Item Type | Description |
| MC multiple select | MC | The item generally consists of a stem and list of choices; the test taker can select two or more choices to respond. This may also include a stimulus. |
| Zones single select\* | Hot Spot | This is an item where the answer choices are predefined “hotspots” on an image. When the test taker selects (clicks) on the spot, the selection is highlighted, shaded, or outlined in red. The test taker selects one zone to respond. |
| Zone multiple select\* | Hot Spot | This is an item where the answer choices are predefined “hotspots” on an image. When the test taker selects (clicks) on the spot, the selection is highlighted, shaded, or outlined in red. The test taker selects two or more zones to respond. |
| Inline choice list single select\* | MC | The stem contains a single blank, and the test taker must fill the blank by selecting a choice from its corresponding choice list. |
| Inline choice list multiple select\* | MC | The stem contains two or more blanks, and the test taker must fill each blank by selecting a choice from the corresponding choice lists. |
| Text choices single select\* | MC | The test taker responds by selecting only one of several underlined words or phrases embedded in a larger section of text. |
| Text choices multiple select\* | MC | The test taker responds by selecting two or more underlined words or phrases embedded in a larger section of text. |
| Numeric | Numeric constructed response (CR) | The test taker responds by filling in a blank entry box with a numeric value. |
| Grid multiple select\* | MC | The test taker responds by marking two or more cells in a table grid. |
| Match single select\* | Drag & Drop | The test taker responds by dragging and dropping a single choice (“source”) into the appropriate location (“target”). |
| Match multiple select\* | Drag & Drop | The test taker responds by dragging and dropping one or more choices (“sources”) into the appropriate locations (“targets”). |
| Composite\* | All item types except CR | The test taker completes multiple tasks based on a combination of machine-scored items. |

### Item Writing Process

Throughout the item writing process, ETS adhered to its foundational guidelines for quality item writing. According to these guidelines, item developers ensured items conformed to the following list of attributes:

* + - 1. The question is clearly and concisely presented.
      2. There is an absence of clueing in the item stem and supporting stimuli.
      3. The supporting stimulus or stimuli is presented clearly and is construct-relevant.
      4. There is a single correct answer (for selected-response items only).
      5. Distractors are plausible, but incorrect (for selected-response only).
      6. The answer key is correct.
      7. Item format and content adhere to the principles of universal design.

#### Item Development Plan

The items developed for use as operational and field test items in the 2020–2021 CSA forms closely reflected the distribution of domains in the blueprint. The total number of machine-scorable items developed (245) was greater than the number used as field test items (228) in the CSA 2020–2021 forms because overage was built in. ETS developed overage to account for the potential rejection of items during item review and data review meetings. If item reviewers at the item review meeting determined that certain items were not appropriate for operational testing, the overage ensured that the minimum item counts for the assessment forms would be satisfied.

Similarly, if item reviewers at the data review meeting determined that certain items were not performing well enough for operational use, the overage ensured that the blueprint for the test forms would still be satisfied.

For the general forms in the assessment, there was substantial overage built in. However, for the accommodated forms, there was little overage when accommodations, such as text-to-speech, closed-captioning, audio transcript, and braille were applied to the items. In contrast, the current policy is for every item testing a standard that is eligible to be tested on accommodated forms to be created (“born”) accessible, removing the need to later apply accommodations to such items.

Table 3.3 shows the number of items developed for the 2019–2020 administration and field-tested in each of the domains of reading, writing mechanics, and listening for the CSA in the 2019–2020 administration. Those items were field-tested again in the 2020–2021 administration in the reused forms. The total number of items developed across the six grade levels and one grade band is 245.

Table 3.3 **Number of Items Developed per Grade Level and Grade Band for the CSA**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Domain | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | High School | Total Items by Domain |
| Listening | 8 | 8 | 8 | 8 | 8 | 8 | 16 | 64 |
| Reading | 27 | 27 | 27 | 27 | 27 | 27 | 39 | 201 |
| Writing Mechanics | 18 | 18 | 18 | 18 | 18 | 18 | 42 | 150 |
| **Total:** | **53** | **53** | **53** | **53** | **53** | **53** | **97** | **415** |

All items created for the CSA adhere to the *ETS* *Standards for Quality and Fairness* (2014)across all phases of item and test development. Each CSA item was developed through a comprehensive development cycle and designed to conform to the principles of quality item writing as defined by ETS.

#### Selection of Item Writers

Senior ETS content staff screened applications for item writers for the CSA, and ETS approved only those with strong content and teaching backgrounds for the item writing training program. ETS selected item writers after the training, but not all recipients of the training became an item writer.

Because some of the participants were current or former California educators, they were particularly knowledgeable about the standards assessed by the CSA. All item writers met the following minimum qualifications:

* Possession of a bachelor’s degree in a relevant field of education (An advanced degree in the relevant content was desirable.)
* Biliterate and bilingual (Spanish and English)
* Previous experience or training in writing items for standards-based assessments, including knowledge of the many considerations that are important when developing items for special student populations

#### Item Writer Training

Item writer training is a vital part of establishing the validity chain for item development. In addition to contracting item writing experts for the CSA, ETS recruited and trained California educators to enrich the range of ideas brought to the pool of items.

ETS assessment specialists provided item writer training to California educators and ETS contractors. The in-person meeting trained California educators on how to write items for the computer-based CSA. ETS led educators through the CCCSSeE, detailed how to write a strong item, and described the functionality of the internet-delivered item types used on this new assessment.

The three primary goals for the training were to

1. provide teachers with knowledge, via professional development on writing items, that they can use to help develop or refine their own classroom teaching and assessments;
2. ensure that teachers who successfully completed the training were ready to develop high-quality items for the CSA; and
3. leverage the experiences, perspectives, and expertise of the teachers in writing items for the CSA.

ETS held item writer training workshops to provide prospective item writers with professional development in several areas. A review of the general assessment development process gave trainees a sense of the total life cycle of an item.

Participants learned best practices in item writing to provide clarity within the item and avoid bias or sensitivity concerns, learned how to review a passage for item opportunities, and were introduced to how the new, innovative item types work.

Given that the trainees were California educators and educational leaders, ETS also emphasized incorporation of current effective teaching practices and instructional activities. Small-group and individual work generated sample items that the ETS facilitators then used in a large-group discussion to analyze and ascertain overall item quality. The ETS team also provided post hoc feedback via email and phone calls to train item writers on further item samples and ideas submitted ahead of contractual item submissions.

### ETS Item Review Process

After items were drafted, ETS placed items developed for the CSA through an extensive internal item review process designed to provide the best standards-based assessments possible. This section summarizes the item review process that confirmed the quality of CSA items.

#### Overview

Once an item was accepted for authoring, ETS employed a series of internal reviews. These reviews used established criteria to judge the quality of item content and to ensure that each item measured what it was intended to measure. These internal reviews also examined the overall quality of the items before presentation to the CDE and item review meetings, which are described in more detail in section [*3.7 California Educator Review*](#_Content_Expert_Review).

All items were entered into IBIS with corresponding artwork and metadata. Within IBIS, items received ETS internal content, fairness, and editorial reviews.

The CDE reviewed proposed changes to items in response to reviews by the participants of the Item and Passage Review meetings to ensure the quality of the item pool. The CDE then gained access to operational CSA items and conducted reviews in IBIS. ETS revised items in response to comments from the CDE prior to using them in the assessment forms.

The ETS review process for the CSA includes the following; these tasks are described in the next subsections:

1. Content review
2. Accessibility review
3. Editorial review
4. Sensitivity and fairness review

Throughout this multistep item review process, the lead content-area assessment specialists and development team members at ETS continually evaluated the activities and items for adherence to the rules for item development.

#### ETS Content Review

On all items ETS developed, content-area assessment specialists conducted three reviews on items and stimuli. These assessment specialists verified thatthe items and stimuli were in compliance with ETS’ written guidelines for clarity, style, accuracy, and appropriateness for California students and were also in compliance with the approved item specifications, the *CAASPP and ELPAC Item Review Acceptance Criteria* (ETS, 2019), and other ETS-produced guidelines for clarity, style, accuracy, and appropriateness for California students. Assessment specialists reviewed each item in terms of the following characteristics:

* Relevance to the purpose of the test
* Match of each item to the item specifications, including the tier of item complexity
* Match of each item to the principles of quality item writing
* Match of each item to the identified standard or standards
* Difficulty of the item
* Accuracy of the content of the item
* Readability of the item or passage
* Grade-level appropriateness of the item
* Appropriateness of any illustrations, graphs, or figures

Assessment specialists checked each item against its classification codes, both to evaluate the correctness of the classification and to confirm that the task posed by the item was relevant to the outcome it was intended to measure. The reviewers could accept the item and classification as written, suggest revisions, or recommend that the item be discarded. These steps occurred prior to the CDE’s review.

#### ETS Accessibility Review

The ETS Accessible Content & Inclusive Solutions team advised on accessibility of items and item types during the ETS content review. These experts on alternate test formats reviewed all items, with a focus on accessibility for all student populations, and provided potential refinement solutions to improve the accessibility.

#### ETS Editorial Review

After content-area assessment specialists and researchers reviewed each item, a group of specially trained editors also reviewed each item in preparation for consideration by the CDE and the item review panelists. The editors checked items for clarity, correctness of language, appropriateness of language for the grade level assessed, adherence to the style guidelines, and conformity with accepted item-writing practices.

#### ETS Sensitivity and Fairness Review

ETS assessment specialists who were specially trained to identify and edit or eliminate questions that contained content or wording that could be construed to be offensive to, or biased against, members of specific student groups (e.g., ethnicity, race, or gender) conducted the next level of review (ETS, 2014 and 2016). These trained staff members reviewed every item before the CDE and item review meetings.

The review process promoted a general awareness of, and responsiveness to, the following:

* Cultural diversity
* Diversity of background, cultural tradition, and viewpoints to be found in the test-taking populations
* Changing roles and attitudes toward various groups
* Role of language in setting and changing attitudes toward various groups
* Topics that may be unsettling or otherwise distract the student from the content being measured, such as natural disasters, disease, or family discord
* Contributions of diverse groups (including ethnic and minority groups, individuals with disabilities, and women) to the history and culture of the United States and the achievements of individuals within these groups
* Item accessibility for language learners of diverse backgrounds

### California Department of Education Review

After ETS reviews of items were completed, the items were reviewed by the CDE content teams. CDE content experts reviewed the items using the same criteria used in the ETS reviews. After CDE reviews occurred, ETS made edits to the items based on the CDE feedback, and the items were then finalized for item review meetings with California educators.

### California Educator Review

In a typical year, ETS holds meetings with California educators in person at the end of the item review process as the final content-expert review that items must undergo before being placed on the CSA. The California educators fill an advisory role to the CDE and ETS and provide guidance on matters related to item development for the CSA.

However, because of the COVID-19 pandemic, the suspension of testing in 2019–2020, and the decision prior to the conclusion of the item development process to reuse test forms from 2019–2020 for the 2020–2021 administration, the educator review meeting was postponed to late 2020 and then held virtually.

ETS convened the meeting with California educators in Sacramento, California, from May 21 through May 23, 2019,[[4]](#footnote-5) to

* review Spanish passages and items for the 2019–2020 assessment (that was reused in the 2020–2021 administration) for grade-level appropriateness, content, bias and sensitivity, readability, and overall interest for the test taker; and
* obtain feedback from California educators about the passages and items to inform ETS on the appropriateness of their use on future test forms for the CSA.

The meeting with California educators was held at the end of the item review process as the final content expert review that items must undergo before being placed on an operational assessment. The California educators filled an advisory role to the CDE and ETS and provided guidance on matters related to item development for the CSA. These educators were responsible for reviewing all newly developed items for alignment to the CCCSSeE. Meeting participants also reviewed the items for accuracy of content, clarity of phrasing, and quality. In their examination of test items, participants could raise concerns related to age or grade-level appropriateness and gender, racial, ethnic, or socioeconomic bias.

#### Composition of Item Review Panels

The panelists for the CSA Item Review Meeting included current and former teachers, resource specialists, administrators, curriculum experts, and other education professionals. To qualify to participate in the item review meeting, educators had to self-assess their written and spoken Spanish as fluent. Preferred qualifications included

* currently being assigned to teach Spanish language arts;
* currently working in Advanced Placement Spanish Language and Culture, on an International Baccalaureate in Spanish, or both;
* currently working in dual immersion or bilingual programs;
* currently serving heritage speakers;
* having a Spanish focus in the postsecondary studies background; and
* having studied or taught in a Spanish-speaking country.

Every effort was made to ensure that groups of item reviewers included a wide representation of genders, geographic regions, and ethnic groups in California. Efforts also were made to ensure representation by members with experience serving California’s diverse Spanish-learning population.

Table 3.4 shows the educational qualifications; present, self-reported occupation; and credentials of the individuals who participated in CSA item review.

Table 3.4 CSA Item Reviewer Qualifications

|  |  |  |
| --- | --- | --- |
| Qualification Type | Qualification | Total |
| **Occupation** | Spanish Teacher | 13 |
| **Occupation** | General Education Teacher | 5 |
| **Highest Degree Earned** | Bachelor’s Degree | 7 |
| **Highest Degree Earned** | Master’s Degree | 11 |
| **Kindergarten (K)–12 Teaching Credential** | Elementary Teaching (multiple subjects) | 9 |
| **K–12 Teaching Credential** | Secondary Teaching (single subject) | 9 |
| **K–12 Teaching Credential** | Spanish | 7 |
| **K–12 Teaching Credential** | English Learner (CLAD, BCLAD) | 2 |
| **K–12 Teaching Credential** | Administrative | 0 |
| **K–12 Teaching Credential** | Other | 10 |

**Note:** Numbers may not match the totals because item reviewers may have multiple occupations or teaching credentials or are currently working toward earning their highest degree.

During a typical item review process, item reviewers are recruited through an online application process. Recommendations are solicited from LEAs and county offices of education as well as from CDE and California State Board of Education staff. ETS assessment directors review applications and confirm that an applicant’s qualifications meet the specified criteria. Applicants who meet the criteria have their information forwarded to CDE staff for further review and agreement before invitations to participate are distributed.

Table 3.5 provides the status of the items after the 2019 item review meetings, since the embedded field test items were a reuse of the 2019–2020 forms. In addition to 245 items developed in the 2019–2020 cycle, some items developed previously were included in the pool for committee review.

Table 3.5 Status of Items After the 2019 Item Review Panel Meetings

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level or Grade Band | Approved As Is | Approved with Revisions | Total Number of Items (\*Including Passages) |
| Grade 3 | 51 | 7 | 58 |
| Grade 4 | 42 | 15 | 57 |
| Grade 5 | 48 | 9 | 57 |
| Grade 6 | 46 | 11 | 57 |
| Grade 7 | 45 | 12 | 57 |
| Grade 8 | 47 | 10 | 57 |
| High school | 86 | 18 | 104 |
| **Total** | 365 | 82 | 447 |

#### Item and Passage Review

After an introductory presentation, an ETS assessment specialist led the participants through a comprehensive training for reviewing passages and items. This training included the structure of an item, the best practices for passage and item reviewing, an explanation of item types and functionality, and a discussion of the metadata accompanying items. This metadata—aligned with the CCCSSeE, depth of knowledge levels, difficulty levels, etc.—was available for each item on a comment sheet.

Following the training, ETS specialists facilitated the review of grade-six passages and items by projecting the items on-screen with printed copies of passages associated with the items, while using the *CAASPP and English Language Proficiency Assessments for California Item Review Acceptance Criteria* (CDE, 2019). The participants were asked to read a passage. When all participants finished, the facilitators projected each item associated with that passage one at a time. The facilitators read each item aloud and displayed any technology-enabled functions.

The educators reviewed grade-six passages and items as a group. Grade six was chosen as a starting point because it is a grade level in the middle of the range of grade levels and it requires neither the extra training in foundational reading for grades three through five nor the secondary consideration of the State Seal of Biliteracy. Then, upon completion of the grade six review, participants were divided into two groups to continue the review process. One group focused on grades three through five, and the other focused on grade seven, grade eight, and high school.

The group discussed each passage together, considering grade-level appropriateness, content, bias and sensitivity, readability, and overall interest for the test taker. The group also discussed the passage’s items together, reviewing for grade-level appropriateness, content, bias and sensitivity, depth of knowledge, standard alignment, and the correct answer or answers as indicated in the metadata. ETS summarized comments, captured any recommended edits, and reached consensus from the group before moving forward to the next item. The group continued in this manner until all passages and items were reviewed. The CDE made decisions separately from the group, as needed, and gave the final approval after requested edits had been applied. Passages and items were then placed on the test forms.

#### Data Review

In a typical year, after items have been included in the CSA forms and administered to students, ETS conducts data review meetings with California teachers and the CDE after the data analysis is complete. Reviewers examine items that are flagged for item difficulty, item-total correlation, item response distribution, and differential item functioning (DIF) according to predefined criteria. The ETS facilitator leads discussions about each flagged item and reviews the content of the item to reach consensus on whether items should be accepted as is or rejected.

The 2020–2021 CSA Data Review Meeting was held in June 2021, with a 94 percent acceptance rate. However, because a sufficient sample size was not available to conduct meaningful item analysis for high school, the data review meeting did not include high school for the 2020–2021 administration.

### Test Assembly and Length

Following the item review process, ETS assessment specialists worked closely with the CDE to select items and assemble test forms. The test forms were assembled to cover a variety of item types, item difficulties, cognitive levels, and key distributions.

ETS developed two test forms per grade. Grades three through eight each had one general form with 62 items per form, including 10 field test items; high school had one general form with 68 items per form, including 16 field test items. Each grade level also had one form with accessibility features. The accommodated forms had 52 operational items and an additional 10 field test items for grades three through eight and 16 items for high school; this form was assigned to students with an individualized education program or Section 504 plan. While many of the items on the accommodated form were identical to, or close variants of, selected items on the general test form, a few items were unique items not appearing on the general form operationally but were field-tested earlier, in fall 2018.

The estimated duration for the CSA was approximately two hours, depending on the student’s grade level. For more detailed information about student testing time, refer to [appendix 7.D](#_Appendix_7.D:_Response).

### Test Production Process

The final steps in production of the CSA are to identify, select, and review items. These are discussed in the following subsections.

#### Psychometric Criteria and Identification of Eligible Items

In addition to the CSA blueprint, statistical guidelines were developed by the ETS Psychometric Analysis & Research (PAR) team to assist in test assembly. The guidelines included the following:

* All items were operationally ready, with item statistics.
* All items conformed to the specifications in the test blueprint.
* Items with *p*-values between 0.2 and 0.95 were used.
* Items that were too difficult or too easy—indicated by low or high *p*-values—were not be used, as they served little purpose in evaluating test takers’ abilities.
* For polytomous items with a maximum of more than one point, the *p*-values could be obtained by dividing the average item score by the maximum score points.
* Items with polyserial correlations greater than 0.2 were to be used. However, given the limited number of CSA items in the item bank, for the 2020–2021 administration, a few items with slightly lower than 0.2 polyserial correlations were included to ensure complete test content coverage. Such items passed the data review and were confirmed with no content issues. The lower polyserial correlations might be due to the small sample size, with limited variance from the field test.
* Category C (large) DIF items were not to be included in the operational form.
* If, for content coverage reasons, it was necessary to include C-DIF items in the form, those items were reviewed by a DIF panel that included members of the focal groups that were affected. The members of the panel confirmed that the items were not biased. The panelists were not to have a vested interest in the outcome of the decision.
* If C-DIF items were to be selected, then a balance with regard to the direction of the C-DIF items was considered; that is, not all C-DIF items should be C- or C+ items.
* The CDE needed to sign off on any C-DIF items before they appear on a test. Refer to section [*7.3 Differential Item Functioning Analyses*](#_Differential_Item_Functioning) for additional information about this criterion.
* Items with *b*-parameters were in the range of ±4.0.
* Items with standard error of *b*-parameters were 0.4 or less.

Because the one-parameter item response theory model was applied to the CSA, the *a*‑parameterwas fixed without estimating.

#### Selection of Items

From the eligible item pool, assessment specialists selected items that, as a whole

* met the coverage specifications of the test blueprint,
* met the form-building guidelines developed by the ETS PAR team,
* represented a wide variety of item types, and
* provided a wide variety of item context.

#### Forms Review

The test forms were evaluated prior to CDE review using the ETS review process presented in table 3.6 and reviewed and approved by the CDE. The details of the ETS review process are included in this subsection.

Table 3.6 **ETS Operational Assessment Forms Review Process**

|  |  |
| --- | --- |
| Step | Task |
| 1. Test Assembly | Assessment specialists select test items that meet the specifications, are fair, and reflect appropriate content coverage. These items are collected in the item bank so they can be tracked as a unit. |
| 1. Senior Review | An assessment specialist with content-area expertise, who did not assemble the test, reviews all of the items and checks for content-related issues (e.g., incorrect keys, overlapping content, clueing of one item by another) and other concerns (e.g., confirming that the items match the test framework). The assessment specialist also verifies that the test meets content and statistical specifications. |
| 1. Psychometric Review | A psychometrician reviews items, operational forms, and field test forms to examine whether all items and forms meet the statistical specifications and whether all information of items and forms related to psychometric analysis are accurate (e.g., the item ID, keys, item sequence, maximum score points, domain category, item statistics, etc.). The psychometrician also reviews the alignment of a form with the blueprint, as well as examines whether items and form statistics meet all psychometric standards. |
| 1. Senior Fresh-Perspective Review | Every new test form goes through a senior fresh-perspective review. During this review, a senior-level content expert, who has never reviewed the form, reads it carefully for any content errors that may have been missed during earlier stages of review. |
| 1. Certification | Once these reviews are completed and the test form is judged to be free from errors, ETS certifies the test form and sends it to be packaged for device delivery. |

##### ETS Psychometric Review

ETS assessment specialists sent the proposed assessment to the ETS PAR team for approval. The proposed assessment was reviewed to ensure that all statistical guidelines were met for both individual items and the assessment as a whole.

Forms should meet the following three statistical specifications:

1. **Statistical properties:** The average *b*-value of the form should be within the range of ±0.35. Also, the test information function (TIF) is used to choose items to produce a test that has the desired precision of measurement at all ability levels. An example of TIF for the operational form is shown in figure 3.2, which uses data from a hypothetical table. In this sample, the y-axis shows the TIF values from 0 to 12 in intervals of 2, and its x-axis shows corresponding ability (theta) values from -3 to 3 in intervals of .5. The curves appear in a reversed U shape. At the peak of the curve, the TIF (test information) is at the maximum, which differentiates students well, while at the bottom of the curve, the TIF is near zero. The solid curved line represents a new form that is based on the hypothetical current administration. The dashed curved line represents a test form for a hypothetical previous administration. Those curved lines summarize how well an assessment differentiates students at which range of ability. This figure indicates that, when TIF is in the peak—0 to 0.25—the assessment differentiates students well. For the example, the TIF value for both administrations is near zero when the theta is -3. The TIF value then increases as theta increases until the TIF value for the hypothetical current administration reaches its peak at about 11, where theta is about .25; or, for the hypothetical previous administration, when it reaches its peak at about 10.5, where the theta is about zero. After the peak TIF value, the TIF value decreases as theta increases until the TIF value is near zero, where the theta is 3.

Figure 3.2 Example TIF

1. **Item sequence:** The sequence of items used in the operational form should be as close as possible to the sequence in the field test form when they were field-tested or their positions in the operational forms when they were tested in the previous operational administration. Whenever possible, the positional difference should be no more than five items.
2. **Other considerations:** Keys of all MC, single-select items should be of roughly equal distribution across all options (i.e., the number of items with keys as A, B, C, or D should be approximately equal).

The psychometric review of forms uses the aforementioned acceptable psychometric criteria to verify item statistics in the forms. To pass this psychometric review, all items should meet the psychometric standards without being flagged.

In addition, all fields related to scoring, such as the maximum scoring point, the sequence of items in the form, the domain ID, and the item ID must undergo psychometric review. This review is designed to ensure that the forms are free from errors, particularly from all problematic items that were identified prior to finalizing the forms. Any flagged items are discussed with ETS Assessment and Learning Technology Research & Development for a final decision.

Psychometric review results for the forms used in the 2020–2021 CSA administration, including the number of forms and number of items, are presented in table 3.7.

Table 3.7 Number of Forms and Items Reviewed Psychometrically

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | # General Form | # Accommodated Form | # Operational Items in General Form | # Operational Items in Accommodated Form | # Field Test Items | Total Items (Operational + Field Test) |
| Grade 3 | 3 | 2 | 52 | 52 | 30 | 112 |
| Grade 4 | 3 | 2 | 52 | 52 | 30 | 123 |
| Grade 5 | 3 | 2 | 52 | 52 | 30 | 122 |
| Grade 6 | 3 | 2 | 52 | 52 | 30 | 121 |
| Grade 7 | 3 | 2 | 52 | 52 | 30 | 123 |
| Grade 8 | 3 | 2 | 52 | 52 | 30 | 125 |
| High school | 3 | 2 | 52 | 52 | 46 | 139 |
| **Overall** | **21** | **14** | **364** | **364** | **226** | **865** |

**Note:** The items that were overlapped between general and accommodated forms were counted just one time in the *Total Items (Operational + Field Test)* column.

##### ETS Content Review of Forms

After psychometric approval, the proposed assessment underwent two additional content reviews and one editorial review. The form reviewers are content specialists who work on testing programs other than the CSA for ETS and who are able to bring a fresh perspective to the review. They were given the appropriate materials to complete the following tasks:

* Verification of item keys
* Identification of possible clueing across the items
* Verification that individual items meet the standard
* Verification of coverage of the standards
* Identification of any possible grammatical or production errors

##### California Department of Education Review of Forms

Following the ETS content review, all proposed assessments were sent to the CDE for review to ensure the proposed assessments met CSA blueprint requirements and to verify there was no clueing between items or statistical issues. The CDE was provided with the following materials:

* Access to items in the item banking system
* A cover sheet indicating key details for the CDE review of the proposed forms in IBIS
* Modified form planners
* Comment sheets

Comments from the CDE were resolved during a virtual meeting with the ETS test development team.

#### Configuration of the Test Delivery System

Once all the test reviews were completed and concerns, if any, had been resolved, the official ordered item sequence of the proposed forms was sent to CAI for configuration of the TDS. Unlike other stages of the test production process, this stage must occur prior to every administration of the CSA even in the case of a form reuse. Therefore, the configuration of the TDS was done prior to the 2020–2021 administration.

Each item underwent an extensive platform review on different operating systems, such as Windows, Linux, and iOS, to ensure that the item looked consistent across all platforms.

The platform review was conducted by a team at CAI consisting of a team leader and several team members. The team leader presented the item as it was approved in ETS and CAI item banks. Each team member was assigned a different platform—hardware device and operating system—and reviewed the item to verify that it rendered as expected. This platform review meeting ensured that all items were presented consistently to all students regardless of testing device or operating system for standardization of the test administration.

Prior to operational deployment, the testing system and content were deployed to a staging server where they were subject to user acceptance testing (UAT) by both ETS and CAI staff. The TDS UAT served to function as both software evaluation and content approval. The UAT procedures followed by ETS staff included reviewing all items for the CSA.

Following the UAT by ETS and CAI staff, separate UAT cycles were conducted by the CDE. The UAT review provided the CDE with an opportunity to interact with the exact test that would be administered to the students. The CDE had to approve the CSA UAT before the test could be released for administration to students.

### References

California Department of Education. (2017). *California Spanish Assessment blueprint.* Sacramento, CA: California Department of Education.

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

Educational Testing Service. (2016). *ETS guidelines for fair tests and communications*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2019). *CAASPP and ELPAC item review acceptance criteria*. [Unpublished manuscript]. Princeton, NJ: Educational Testing Service.

### Appendix 3.A: CSA Blueprint Overview—Operational Forms

The test blueprint for the CSA provides the proposed numbers of items and points to be included in an operational assessment for each Spanish reading/language-arts domain assessed in grades three through eight and high school.

All items are aligned with the translated and linguistically augmented version of the CCCSSeE.

Each grade has more than 50 testable standards at its disposal. There are three overview tables provided—grades three through five, six through eight, and high school (grades nine through twelve)—to clarify the overall proportions of the blueprint. After the overview tables, specifics for each tested grade level are given, enumerating further subdivisions of the content and specific groups of testing standards. Note that high school grades are tested together in one level using the CCCSSeE designated as “9–10” and “11–12” and uses the designation “high school” (HS).

The blueprint is represented in tables. Each overview table is organized by the three domains assessed: reading, writing (mechanics), and listening—referred to as claim/score reporting category—and are provided in the first column. Other columns in the blueprint are as follows:

* *Second column:* Content Category
* *Third column:* Number of items representing the content category on an operational assessment
* *Fourth column:* Number of points for the given content category
* *Remaining columns:* Aggregated item counts, points, and percentages by claim

Item counts and point values may be adjusted further during future stages of the CSA design and development effort to take into consideration the evaluation of pilot test results as well as the analyses of statistics of both the first field test and the first operational administration of the CSA.

Finally, for all tables, some items are anticipated to be polytomously scored (maximum of two points), so the number of items is smaller than the number of score points.

Table 3.A.1 CSA Test Blueprint for Grade Span Three Through Five

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Claim/Score Reporting Category** | **Content Category** | **Total Items by Content Category** | **Total Score Points by Content Category** | **Total Items by Claim** | **Percentage of Items by Claim** | **Total Score Points by Claim** | **Percentage of Score Points by Claim** |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Literary | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Informational | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Vocabulary and Meaning | 8–10 | 10–13 | 24 | 46% | 27–35 | 40–58% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Foundational Mechanics and Conventions | 8–10 | 10–13 | 16 | 31% | 19–22 | 28–37% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Revising and Editing | 5–7 | 6–9 | 16 | 31% | 19–22 | 28–37% |
| **Listening Claim:** Students can comprehend spoken Spanish in a range of contexts. | Listening Comprehension | 12 | 15–17 | 12 | 23% | 15–17 | 22–28% |
| N/A | N/A | N/A | **TOTALS:** | **52** | **100%** | **61–66** | **100%** |

Table 3.A.2 CSA Test Blueprint for Grade Span Six Through Eight

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Claim/Score Reporting Category** | **Content Category** | **Total Items by Content Category** | **Total Score Points by Content Category** | **Total Items by Claim** | **Percentage of Items by Claim** | **Total Score Points by Claim** | **Percentage of Score Points by Claim** |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Literary | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Informational | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Vocabulary and Meaning | 8–10 | 10–13 | 24 | 46% | 27–35 | 40–58% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Mechanics and Conventions | 7–9 | 8–11 | 16 | 31% | 19–22 | 28–37% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Revising and Editing | 7–9 | 8–11 | 16 | 31% | 19–22 | 28–37% |
| **Listening Claim:** Students can comprehend spoken Spanish in a range of contexts. | Listening Comprehension | 12 | 15–17 | 12 | 23% | 15–17 | 22–28% |
| N/A | N/A | N/A | **TOTALS:** | **52** | **100%** | **61–66** | **100%** |

Table 3.A.3 CSA Test Blueprint for Grade Span Nine Through Twelve (High School)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Claim/Score Reporting Category** | **Content Category** | **Total Items by Content Category** | **Total Score Points by Content Category** | **Total Items by Claim** | **Percentage of Items by Claim** | **Total Score Points by Claim** | **Percentage of Score Points by Claim** |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Literary | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Informational | 6–9 | 7–11 | 24 | 46% | 27–35 | 40–58% |
| **Reading Claim:** Students can read, analyze, and interpret a variety of texts and genres through Spanish. | Vocabulary and Meaning | 8–10 | 10–13 | 24 | 46% | 27–35 | 40–58% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Mechanics and Conventions | 7–9 | 8–11 | 16 | 31% | 19–22 | 28–37% |
| **Writing Mechanics Claim:** Students can revise writing products that accurately and convincingly present, describe, and explain ideas for a range of purposes and audiences through Spanish. | Revising and Editing | 7–9 | 8–11 | 16 | 31% | 19–22 | 28–37% |
| **Listening Claim:** Students can comprehend spoken Spanish in a range of contexts. | Listening Comprehension | 12 | 15–17 | 12 | 23% | 15–17 | 22–28% |
| N/A | N/A | N/A | **TOTALS:** | **52** | **100%** | **61–66** | **100%** |

## Test Administration

This chapter provides an overview of the California Spanish Assessment (CSA) administration. It includes a system functionality overview, descriptions of the efforts and measures to ensure test security, and procedures for implementation of test accommodations based on the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, Chapter 6). In particular, because of the novel coronavirus disease 2019 (COVID-19) pandemic, remote and in-person testing options are discussed.

### Overview

The CSA was available to all students in grades three through twelve in spring 2021 in conjunction with the other tests that comprise the California Assessment of Student Performance and Progress (CAASPP) System.

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

The testing window for the spring 2021 administration of the CSA was scheduled for January 14 through July 30, 2021. Specific test administration schedules within that window were determined locally pursuant to the *California Code of Regulations*, Title 5 (5 *CCR),* sections 855(a)(1), 855(a)(2), 855(b), and 855(c).

For the 2020–2021 administration year, 13,848 students were registered to take the CSA. Of those registered, 5,683 students, or 41 percent, started and completed the CSA, as table 1.1 shows. Only 34 students, or 0.25 percent, across all grade levels started but did not complete the CSA.

The CSA used the same secure browser and online testing platform as all the CAASPP assessments. The students received initial directions in Spanish from the test administrator as well as item-level directions, as needed. At the beginning of each assessment, there were three additional questions, administered to collect information on whether the student received instruction in Spanish, the Spanish-language program type, and the percentage of instruction in Spanish. Results of the survey provide information on different types of Spanish-language programs in LEAs, which is helpful in identifying the CSA test-taking population. [Appendix 4.A](#_Appendix_4.A:_Demographic) contains information based on the survey questions.

#### Remote and In-Person Testing

When the 2020–2021 school year began, LEAs offered varying instructional options because of the COVID-19 pandemic, with a substantial percentage offering only distance learning. This resulted in the need for the California Department of Education (CDE) to explore different options for delivering its annual summative assessments. Two means of testing students were offered to LEAs for all CAASPP assessments (with the exception of the alternate assessments):

1. Test in person, with both students and test administrators co-located in the same room at a school or other secure location and following physical distancing guidelines.
2. Test remotely, with students and test administrators located at different physical locations. The test administrator would monitor students’ progress throughout the test by using remote monitoring tools connected to the test delivery system (TDS).

After 2020–2021 testing, ETS conducted internal studies on the potential impact of the options provided for the 2020–2021 administration. (Because of the insufficient sample size, the CSA was not included in the analyses.) Findings of the analyses using other, larger-volume California assessments provide useful perspectives for CSA 2020–2021 administration and score interpretation. The results of the analyses support the suggestion that the remote test can be viewed as reasonably comparable to in-person testing for the large-volume assessments, such as the Smarter Balanced for English Language Arts/Literacy assessments and for the lower grades (i.e., grade three through grade six) of the Smarter Balanced for Mathematics assessments.

#### Test-Taking Summaries

Although student participation was voluntary for the CSA, all LEAs in California were invited to administer this assessment.

LEAs were given the following guidelines to determine whether a student should take the CSA when either of these conditions applied:

* The student is receiving instruction in Spanish in the state of California.
* The student is seeking a measure that recognizes the student’s Spanish reading, writing mechanics, and listening language arts skills (CDE, 2021e).

Table 1.1 in [chapter 1](#_Chapter_1:_Introduction) presents the test-taking summary by grade level in the 2020–2021 administration. What follows are the findings from this analysis:

* The completion rates for the CSA were about 38 to 40 percent of registered students in grades three through five, 50 to 57 percent in grades six through eight, and 16 percent in high school.
* The number of students registered to test in the 2020–2021 administration is much lower than the number of students registered in 2018–2019.
* The number of students who took the CSA during the 2020–2021 administration is about 13 percent of the number of students who took the CSA in 2018–2019 in grades three through five, 20 percent in grades six through eight, and 6 percent in high school.

Results also show that a higher percentage of students tested remotely than in person. As grade level increases, the percentage of students tested remotely is higher.

A total of 78 LEAs participated in the CSA. More than one half of the total number of LEAs—41 out of 78 LEAs—are from the southern region. The mean completion rate is the average of the completion rates from all LEAs in the region.

Table 4.1 presents the test-taking rates of each region in California. Note that the numbers in this table show whether LEAs actually have students tested after they registered their students for testing. It did not show, however, the participation rates that would have occurred under a typical administration. Because of the impact of the COVID-19 pandemic, not all LEAs had registered their students for testing as they would in a typical year.

Table 4.1 Test-Taking Rates by California Region

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Region | No. of LEAs | Total Students Registered | Total Students Completed | Mean Completion Rate | Minimum Completion Rate | Maximum Completion Rate |
| North | 12 | 214 | 201 | 74.74 | 0.00 | 100.00 |
| Central | 25 | 5,297 | 3,255 | 55.43 | 0.00 | 100.00 |
| South | 41 | 8,337 | 2,227 | 32.16 | 0.00 | 100.00 |

Table 4.2 presents the test-taking rates by grade level. The data reveals that about 50 percent or more of students in grades six through eight who were registered for the test actually took it. Slightly more than one third of students who registered in grades three through five actually took the test. High school grades had the lowest test-taking rate, at 16.41 percent.

Table 4.2 CSA Test-Taking Rates by Grade Level and Grade Band

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level or Grade Band | Number of Registered Students | Number of Students Tested | Percent of Students Tested |
| Grade 3 | 3,066 | 1,213 | 39.56 |
| Grade 4 | 2,778 | 1,052 | 37.87 |
| Grade 5 | 2,371 | 916 | 38.63 |
| Grade 6 | 1,662 | 948 | 57.04 |
| Grade 7 | 1,292 | 729 | 56.42 |
| Grade 8 | 1,137 | 572 | 50.31 |
| High school—Grade 9 | 422 | 114 | 27.01 |
| High school—Grade 10 | 385 | 72 | 18.70 |
| High school—Grade 11 | 450 | 43 | 9.56 |
| High school—Grade 12 | 285 | 24 | 8.42 |
| High school—All grades | 1,542 | 253 | 16.41 |

#### Demographic Student Group Summaries

The number and the percentage of students for selected groups with test completion and valid test scores are provided for grades three through eight and high school in table 4.A.1 through table 4.A.7 of [appendix 4.A](#_Appendix_4.A:_Demographic). Grade levels reflect students’ tested grade levels during the 2020–2021 school year.

In the tables, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, length of enrollment in US schools, self-reported Spanish-language program type, and percentage of daily instruction in Spanish.

Note that data collected for program types and percentage of the school day instruction comes from the student demographic survey that was part of the CSA. Note, too, that Spanish as a foreign language programs are most often available only for students in grades six through eight and high school.

Student groups reported in [appendix 4.A](#_Appendix_4.A:_Demographic) are presented in table 4.3.

Table 4.3 Demographic Student Groups to Be Reported

|  |  |
| --- | --- |
| Student Group | Definition |
| Economic Status | * Not economically disadvantaged * Economically disadvantaged |
| English Language Fluency | * English only * Initial fluent English proficient (IFEP) * English learner (EL) * Reclassified fluent English proficient (RFEP) * Ever-ELs (EL, RFEP, or adult English learner [ADEL]) * To be determined * English proficiency unknown |
| Enrollment in US Schools | * Less than 12 months * 12 months or more |
| 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 |
| Gender | * Male * Female |

Table 4.3 *(continuation)*

|  |  |
| --- | --- |
| Student Group | Definition |
| Percentage of School Day Instruction Provided in Spanish | * 0–25% * 26–50% * 51–75% * 76–100% |
| Received Instruction in Spanish in the 2020–2021 School Year—Program Type | * One-Way Immersion * Dual-Language Immersion * Developmental Bilingual * Heritage Language or Indigenous Language * Spanish as a Foreign Language[[5]](#footnote-6) |
| Special Education Services Status | * No special education services * Special education services |

### Procedures to Maintain Standardization

The test administration procedures are designed so that the tests are administered in a standardized manner. ETS takes 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 at the beginning of the 2020–2021 school year. LEAs include public school districts, 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 assessments. In addition to the responsibilities set forth in 5*CCR* Section 857, their responsibilities included

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

#### CAASPP Test Site Coordinator

A CAASPP test site coordinator is trained by the LEA CAASPP coordinator for each test site (5 *CCR* Section 857[f]). A test site 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]). CAASPP test site coordinators’ 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 all students take the test, 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 are identified by CAASPP test site coordinators as individuals who will administer the CSA.

A test administrator signed a security affidavit (5 *CCR* Section 850[ae]). 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 assessments, including the CSA;
* reporting all test security incidents to the test site 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 test with appropriate resources and reporting potential data errors to test site 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 CSA (CDE, 2021e).

#### Instructions for Test Administrators

##### Test Administrator *Directions for Administration*

The *DFAs* of the CSAused by test administrators to administer the CSA to students are included in the *CAASPP Online Test Administration Manual* (CDE, 2021e). 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. Instructions for the CSA were written in Spanish and were to be read to students in Spanish.

Additionally, the *CAASPP Online Test Administration Manual* provided information to test administrators regarding the systems involved in testing, including sections describing the TDS, so they could become familiar with the testing application used by their students (CDE, 2021e).

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

The *CAASPP Online Test Administration Manual* (CDE, 2021e) 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:

* Roles and responsibilities of those involved with CAASPP testing
* 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 test and systems associated with the test.

##### *Spring Administration Information for Educators*

The *Spring Administration Information for Educators* (CDE, 2021f),which was developed in response to the need for remote administration of the CAASPP and English Language Proficiency Assessments for California (ELPAC), provided instructions and resources that coordinators and test administrators could use to prepare for testing and in test administration. Sections included the following topics:

* Administration options
* Requirements
* Test security
* Instructions for remote testing, including test administration directions and scripts
* Videos and quick reference guides
* Helpful links (including to the Parent/Guardian Information website)

##### *CAASPP and English Language Proficiency Assessments for California 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.

TOMS modules described in the *TOMS User Guide* included the following (CDE, 2021d):

* **Test Administration Setup—**This module allowed LEAs to determine and calculate dates for the LEA’s 2020–2021 administration of the CAASPP, including the CSA.
* **Adding and Managing Users—**This module allowed LEA CAASPP coordinators to add CAASPP test site coordinators and test administrators to TOMS so that the designated user could administer, monitor, and manage the CAASPP computer-based assessments.
* **Reports—**This module allowed LEA CAASPP coordinators and CAASPP test site coordinators access to the various reports in TOMS.
* **STAIRS/Appeals—**This module allowed LEA CAASPP coordinators and CAASPP test site coordinators access to create new STAIRS cases or search for STAIRS/Appeals cases.
* **Student Profile—**This module allowed LEA CAASPP coordinators, CAASPP test site coordinators, and test administrators and test examiners to view and manage student 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 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 assessments (CDE, 2021c).
* ***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, 2021b).
* ***CAASPP and ELPAC Accessibility Guide for Online Testing*—**This manual provided descriptions of the accessibility features for computer-based tests as well as information about supported hardware and software requirements for administering tests to students using accessibility resources, including those with a braille accommodation using Job Access With Speech (JAWS®) (software) or a braille embosser (hardware) (CDE, 2021a).

### 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 are developed with interested educator feedback and specify the audience, topics, frequency, and mode (in person, webcast, videos, modules, etc.) of the training, including such elements as format, participants, and logistics.

In 2020–2021, ETS and SCOE quickly adapted training plans to meet the needs of educators deciding how to complete testing during the COVID-19 pandemic while adhering to local health guidance. All in-person trainings were converted to a virtual format, and the longer trainings were separated into shorter segments to avoid learner fatigue.

Knowing that educators were confronted with new challenges daily that put additional demands on their time, ETS and SCOE made every effort to make the information available in a variety of ways that allowed educators access to training at a time that was responsive to their varying circumstances. This included offering training events on multiple days and times, livestreaming events, recording and archiving trainings, and converting trainings to self-paced modules that could be taken any time, at the learner’s convenience.

All training opportunities were posted in one centralized location on the CAASPP website. LEA staff were able to register for training opportunities, across both CDE contractors’ offerings, in one place, on the Upcoming Training Opportunities web page. A Past Training Opportunities web page was also created, making it easier for educators to find missed training opportunities and providing easier access to recorded trainings.

#### Workshops, Virtual Training, and Webcasts

All offered virtual trainings were recorded and made available for on-demand viewing. Most trainings were offered via Zoom, a platform that educators quickly became familiar with and comfortable using during the COVID-19 pandemic. Zoom provided an opportunity for educators to ask questions and get answers in real time. Virtual trainings were also livestreamed on YouTube so that educators still had access if a particular training reached registration capacity.

In response to an environment where educators had competing priorities to juggle, ETS and SCOE employed a variety of strategies to increase engagement during virtual trainings. Live polls were 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 was enabled to give participants an opportunity to interact with each other or to provide open-ended feedback, or it was disabled to minimize distraction and drive attendees’ focus to the information being presented. Breakout groups were used in smaller group trainings, as appropriate. Breaks and processing time were incorporated into presentations to give attendees opportunities to attend to other responsibilities that might result as part of their job or home environment. Registered participants received an email from SCOE with a link to the virtual trainings.

Working closely with the CDE, ETS and SCOE were able to increase support to educators during a particularly challenging year. ETS offered weekly Office Hours and Coffee Sessions. Office Hours included CDE and ETS leadership to provide quickly changing updates on policies related to testing. Guest speakers from LEAs were invited to offer solutions and strategies for dealing with the challenges happening at the local level. Coffee Sessions included technical staff who could answer questions about all aspects of testing, including the newly offered remote testing option. SCOE continued to offer assessment update meetings intended to provide LEA coordinators with regular updates about California’s assessment system. All trainings and meetings were recorded and archived for on-demand viewing on the Past Training Opportunities web page on the CAASPP website.

An unexpected benefit of the COVID-19 pandemic is that educators had greater access to CDE, ETS, and SCOE staff than they had in prior administration years. This challenging year provided an opportunity to provide more targeted support to educators that will have a lasting impact on the administrations to come.

#### Videos and Guides

To supplement the virtual trainings, ETS continued to produce videos on various aspects of administering the CAASPP. 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 15 additional videos to support remote test administration. The videos included videos targeted to parents/‌guardians and students to provide instruction on how to download the secure browser on a personal device, so the assessment could be taken at home, and videos on how to take an assessment at home. Videos for parents/guardians and students were produced in both English and Spanish. SCOE produced a number of quick reference guides and guides to support remote testing and those were made available in the top 10 most common languages in California according to DataQuest.

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

ETS produced short demonstration videos for every embedded accessibility resource, demonstrating how to use the resource for educators, students, and parents/guardians. The videos were available in both English and Spanish on the Accessibility Resources Demonstration Videos web page on the CAASPP website. In addition, ETS developed a video with LEA staff to help California educators learn more about the importance of implementing CAASPP accessibility resources and best practices used by educators in the field. The “Importance of Implementing CAASPP and ELPAC Accessibility Resources: Voices from Educators” video was available on the Quick Reference Guides and Videos web page on the CAASPP website.

Accessibility resource videos were also 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 2020–2021 CAASPP administration, ETS introduced a new virtual training series, “Matching Accessibility Resources to Students’ Needs.” This training focused on providing participants 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. The virtual training was originally intended as a one-time event but, because of overwhelming interest, the training was offered on four additional dates. The training was recorded and archived. LEA coordinators, test site coordinators, test administrators, and test examiners were notified via email when the recorded training was available, further extending its reach.

At the California Assessment Conference, SCOE offered three sessions on accessibility. A “Plenary Accessibility 101” session was available as a prerecorded session for all conference attendees and was intended to build a shared understanding of basic accessibility-related terms and considerations. The “Digging Deeper into Accessibility” breakout session focused on developing an equitable and systematic process for matching students with appropriate accessibility resources. “Universal Design for Learning and Accessibility Resources: A Pathway to Success for All Students” was another breakout session focused on providing an opportunity to practice appropriately matching student needs to the various accessibility resources.

### Accessibility Resources

The Every Student Succeeds Act reaffirms the importance of ensuring that assessments are accessible to special 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 CSA, 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 was reported based on available data.

The 2020–2021 CSA 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’s California Assessment Accessibility Resources Matrix (Accessibility Matrix) (CDE, 2020) 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 toolswere 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 2020–2021 CSA administration.

###### Embedded

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

* Breaks
* Digital notepad
* Expandable items
* Expandable passages
* Highlighter
* Keyboard navigation
* Line reader
* Mark for review
* Spanish glossary (for specific items)
* Strikethrough
* Zoom (in or out)

###### Non-Embedded

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

* Breaks
* Scratch paper

##### Designated Supports

Designated supports were available to all students 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 TDS (through the CAASPP secure browser).

The designated supports in the following subsections were available in the 2020–2021 CSA 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)
* Turn off any universal tool(s)

###### Non-Embedded

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

* Amplification
* Color contrast
* Color overlay
* Magnification
* Medical supports
* Noise buffers
* Read aloud (items)
* Scribe (nonwriting items)
* Separate setting (special lighting or acoustics, adaptive furniture, time of day)
* Simplified test directions

##### Accommodations

Accommodations were changes in procedures or materials that increased equitable access during the CAASPP assessments. 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 2020–2021 CSA administration.

###### Embedded

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

* Audio transcript
* Braille (embossed and refreshable)
* Closed-captioning
* Text-to-speech (reading passages)

###### Non-Embedded

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

* Alternate response options
* Print on demand
* Read aloud (reading passages)

##### 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 Accessibility Matrix included an inventory of unlisted resources that were already identified and were preapproved (CDE, 2020). During the 2020–2021 CAASPP administration, an LEA CAASPP coordinator or a CAASPP test site coordinator would use TOMS to submit a request for use of an unlisted resource. A preidentified, preapproved unlisted resource 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. The student’s score status would be changed to “Invalid” and the student’s scale score would be reported but appear on the Student Score Report (SSR) with an asterisk and a footnote that the test was administered under conditions that resulted in a score that may not be an accurate representation of the student’s achievement.

The CDE preidentified an English dictionary as a non-embedded unlisted resource that changes the construct being measured.

The LEA CAASPP coordinator or CAASPP test site 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. The lowest obtainable scale scores were reported for the affected domain when administrations included unlisted resources that changed the construct of that assessment.

#### Identification

All public school students 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, 2020) and the CDE Accessibility Matrix (CDE, 2020) 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. The CSA follows 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 document, the *Smarter Balanced Resources and Practices Comparison Crosswalk* [[6]](#footnote-7)(Smarter Balanced, 2018), connects the assessment resources described in the *Guidelines* with associated classroom practices.

Another manual, the *Smarter Balanced Usability, Accessibility, and Accommodations Implementation Guide* (Smarter Balanced, 2014),provides suggestions for implementation of these resources. Test administrators are given the opportunity to administer the CSA practice and training tests so that students have the opportunity to familiarize themselves with a designated support or accommodation prior to testing.

#### Assignment

Once a student’s IEP or Section 504 plan team decided which accessibility resource(s) the student should use, LEA CAASPP coordinators and CAASPP test site 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 [*2.4.1.1 Selection*](#_Resources_for_Selection).)
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 CAASPP test site 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.

[Appendix 4.B](#_Appendix_4.B:_Special) provides information on the number of students who were assigned accommodations and designated supports.

#### Usage of Designated Supports and Accommodations

LEA CAASPP coordinators and CAASPP test site 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 test 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 test so the student could be retested without the accommodation or designated support.

After schools and LEAs assigned eligible students to accommodations or designated supports, Cambium Assessment, Inc.’s (CAI’s) TDS provided and captured whether a certain accommodation or designated support (or multiple accommodations or supports) were used by a student as the student progresses through the test.

[Appendix 4.B](#_Appendix_4.B:_Special) reports the number of students who were assigned to a certain accommodation or designated support at school. Table 4.4 reports the number of students who, based on the availability of data, actually used this accommodation or designated support at least once. The number of students in [appendix 4.B](#_Appendix_4.B:_Special) is higher than the number of students in table 4.4 because a student who was assigned to an accessibility resource may not have used it during testing. Table 4.4 reports the available data only.

Types of accommodations and designated supports—labeled “ACC” and “DS” in the *Resource Type* column—included in table 4.4 are as follows:

1. **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 impairment who are accustomed to accessing information presented via audio in the form of braille.
2. **Text-To-Speech:** Text is read aloud to the student via embedded text-to-speech technology.
3. **Print on Demand:** Paper copies of passages and stimuli, items, or all of these are printed for students.
4. **Masking:** This resource involves blocking off content that is not of immediate need or that may be distracting to the student.

Table 4.4 Summary of Accommodations and Designated Supports Used by Students, by Grade Level or Grade Band

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Accessibility Resource | Resource Type | Students Assigned | Students Used |
| 3 | Embedded Audio Transcript | ACC | 0 | 0 |
| 3 | Embedded Text-to-Speech Passages | ACC | 6 | 5 |
| 3 | Non-Embedded Print on Demand | ACC | 0 | 0 |
| 3 | Embedded Masking | DS | 1 | 0 |
| 3 | Embedded Text-to-Speech Items | DS | 16 | 5 |
| 4 | Embedded Audio Transcript | ACC | 0 | 0 |
| 4 | Embedded Text-to-Speech Passages | ACC | 7 | 4 |
| 4 | Non-Embedded Print on Demand | ACC | 1 | 0 |
| 4 | Embedded Masking | DS | 3 | 0 |
| 4 | Embedded Text-to-Speech Items | DS | 11 | 4 |
| 5 | Embedded Audio Transcript | ACC | 0 | 0 |
| 5 | Embedded Text-to-Speech Passages | ACC | 8 | 3 |
| 5 | Non-Embedded Print on Demand | ACC | 0 | 0 |
| 5 | Embedded Masking | DS | 2 | 1 |
| 5 | Embedded Text-to-Speech Items | DS | 12 | 5 |
| 6 | Embedded Audio Transcript | ACC | 0 | 0 |
| 6 | Embedded Text-to-Speech Passages | ACC | 8 | 4 |
| 6 | Non-Embedded Print on Demand | ACC | 0 | 0 |
| 6 | Embedded Masking | DS | 0 | 0 |
| 6 | Embedded Text-to-Speech Items | DS | 11 | 3 |

Table 4.4 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Accessibility Resource | Resource Type | Students Assigned | Students Used |
| 7 | Embedded Audio Transcript | ACC | 0 | 0 |
| 7 | Embedded Text-to-Speech Passages | ACC | 11 | 4 |
| 7 | Non-Embedded Print on Demand | ACC | 0 | 0 |
| 7 | Embedded Masking | DS | 1 | 0 |
| 7 | Embedded Text-to-Speech Items | DS | 12 | 2 |
| 8 | Embedded Audio Transcript | ACC | 0 | 0 |
| 8 | Embedded Text-to-Speech Passages | ACC | 11 | 3 |
| 8 | Non-Embedded Print on Demand | ACC | 0 | 0 |
| 8 | Embedded Masking | DS | 3 | 0 |
| 8 | Embedded Text-to-Speech Items | DS | 12 | 2 |
| HS | Embedded Audio Transcript | ACC | 0 | 0 |
| HS | Embedded Text-to-Speech Passages | ACC | 0 | 0 |
| HS | Non-Embedded Print on Demand | ACC | 0 | 0 |
| HS | Embedded Masking | DS | 0 | 0 |
| HS | Embedded Text-to-Speech Items | DS | 0 | 0 |

### Practice and Training Tests

Practice and training tests are available publicly for the CSA. These tests simulate the experience of the computer-based CSA. For the 2020–2021 school year, accommodated versions of CSA 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 Online Practice and Training Tests Portal web page on the CAASPP website.

The practice tests, offered at each grade level for grades three through eight with one test for high school, were released to prepare students for the CSA. These tests more closely simulate the CSA’s length and complexity and align with the CSA blueprint.

The grade-level-specific training tests can be taken by students in all tested grades. All unique item types available on the operational test are covered in the training tests.

The scoring guides for the practice and training tests are available on the Online Practice Test Scoring Guides and *Directions for Administration (DFAs)* web page on the CAASPP website.

### Test Security and Confidentiality

For the operational CSA, every person who worked with the assessments, communicated test results, or received testing information was responsible for maintaining the security and confidentiality of the tests, including CDE staff, ETS staff, ETS subcontractors, LEA assessment coordinators, school assessment coordinators, students, parents/guardians, teachers, and cooperative educational service agency staff. ETS’ Code of Ethics required that all test information, including tangible materials (e.g., test items), confidential files (e.g., those containing personally identifiable student information), and processes related to test administration (e.g., 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 tests that ETS develops or handles, ETS maintains an Office of Testing Integrity (OTI), which is described in the next subsection.

All tests 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* (AERA, APA, & NCME, 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 *CSA 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.

#### 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 the OTI’s 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’s mission is to

* 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 balanced way that reflects the laws and professional standards governing the integrity of testing.

In an effort to enforce secure testing practices, the OTI strives to safeguard the various processes involved in a test development and administration cycle. For the CSA, those processes included the following:

* Test 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 CAASPP test site coordinators and test administrators are properly trained regarding security policies and procedures.

The CAASPP test site 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 CAASPP test site coordinator and securely destroying printed and digital media for items and passages generated by the print-on-demand feature of the TDS (CDE, 2021e).

The following measures ensured the security of the CAASPP:

* LEA CAASPP coordinators and test site coordinators must have electronically signed and submitted a “CAASPP Test Security Agreement for LEA CAASPP coordinators and CAASPP test site coordinators” form in TOMS before ETS can grant the coordinators access to TOMS (5 *CCR*, Education, Division 1, Chapter 2, Subchapter 3.75, Article 1, Section 859[a]).
* 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]).

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 CAASPP test site 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. Typically, ETS would perform site visits and testing procedure audits during the testing window; however, these visits were not made during the 2020–2021 CAASPP administration because many schools and LEAs were not open for in‑person instruction as a result of the COVID-19 pandemic. It is expected that these visits will resume in future administrations per state health and safety guidelines. However, selected LEAs were audited remotely, with LEA CAASPP coordinators or other LEA staff responding to a series of questions about test administration.

#### 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 [*1.10 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 currently 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 2020–2021 CSA, 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 it had been posted.

#### Data Management in the Secure Database

ETS currently 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, 2021d).

#### 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 (e.g., 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 CSA:

* Individual student results for computer-based assessments in the California Educator Reporting System
* Individual SSRs (electronic)
* Internet reports—available on a public web reporting site—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. ETS 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 the BS 25999-2 standard for business continuity and conducts disaster recovery exercises annually. ETS routinely backs up all data to either disks through deduplication or to tapes, all of which are stored off site.

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 events:

* Scoring
* Transfer of scores by means of secure data exchange
* Reporting
* Posting of aggregated data
* Storage

In addition to protecting the confidentiality of testing materials, 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 a test developed by ETS (e.g., a CAASPP assessment). 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 tests, which, in turn, compromise the reliability and validity of test results (CDE, 2021b). Whether intentional or unintentional, failure by staff or students to comply with security rules constitutes a test security incident. Test security incidents have impacts on scoring and affect students’ performance on the test.

LEA CAASPP coordinators and CAASPP test site coordinators ensured that all test security and summative 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 test. If an Appeal to a student’s test 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 CAASPP test site coordinator). The STAIRScase in TOMS provided the LEA CAASPP coordinator, the CDE, and CalTAC with the opportunity to interact and communicate regarding the STAIRS process (CDE, 2021b).

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 (i.e., 20-minute pause rule)

Appeals requests were reviewed by the CDE or CalTAC. When a request to submit an Appeal was approved, the coordinator received a system-generated email with the Appeal type that was approved (CDE, 2021b).

Types of Appeals available during the 2020–2021 CAASPP administration are described in table 4.5.

Table 4.5 Types of Appeals

|  |  |
| --- | --- |
| Type of Appeal | Description |
| Reset | Resetting a student’s summative assessment removed that assessment from the system and enabled the student to start a new assessment from the beginning. |
| Invalidate | Invalidated summative 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. The score was counted as “not proficient” for aggregation into the CAASPP results. |
| Re-open | Reopening a summative assessment allowed a student to access an assessment that had already been submitted or had expired. |
| Restore | Restoring a summative assessment returned an assessment from the Reset status to its prior status. This action could be performed only on tests that were reset previously. |
| Grace Period Extension | Permitting a grace period extension allowed the student to review previously answered questions 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 test, test security, or test validity. An impropriety can be corrected and contained at a local level. An impropriety should be reported to the LEA CAASPP coordinator and CAASPP test site coordinator immediately. The coordinator should 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 test or impact test security or test validity. 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 CAASPP test site 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 test. 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 telephone. Following the call, the CAASPP test site 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.

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.

#### Appeals

For test security incidents reported in STAIRS that resulted in a need to reset, reopen, invalidate, or restore individual computer-based student assessments, the request had to be approved by the CDE. Requests to reset and reopen assessments were processed by CalTAC.

In most instances, an Appeal was submitted to address a test security breach or irregularity. The LEA CAASPP coordinator or CAASPP test site coordinator submitted Appeals in TOMS. All submitted Appeals were available for retrieval and review by the appropriate credentialed users within a given organization. However, the view of Appeals was restricted according to the user role as established in TOMS. An Appeal could be requested only by the LEA CAASPP coordinator or CAASPP test site coordinator if prompted while filing a STAIRS case in TOMS (CDE, 2021b). Types of Appeals available during the 2020–2021 CAASPP administration are described in table 4.5.

Table 4.6 presents the number and types of incidents submitted in STAIRS in the 2020–‍2021 administration for the CSA as well as the number of Statewide Student Identifiers (SSIDs) submitted and approved.

Table 4.6 Number and Types of Incidents Submitted in STAIRS in the 2020–2021 Administration

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Description | Appeal Type | Number of Incidents | Total Number of SSIDs Submitted | Appeals SSIDs Approved |
| Accessibility Issue | Reset | 2 | 2 | 2 |
| Accidental Summative Access | Reset or Re-open or No Appeal | 0 | 0 | 0 |
| Administered Incorrect Assessment | Reset | 0 | 0 | 0 |
| Administration Error | No Appeal | 0 | 0 | 0 |
| Data Entry Issue | Reset or Re-open | 0 | 0 | 0 |
| Expired or Accidentally Submitted Test | Re-open | 23 | 23 | 23 |
| Exposing Secure Materials | Invalidate or No Appeal | 0 | 0 | 0 |
| Incorrect SSID Used | Reset or No Appeal | 0 | 0 | 0 |
| Restore from Reset | Restore | 0 | 0 | 0 |
| Student Cheating or Accessing Unauthorized Devices | Invalidate | 1 | 1 | 1 |
| Student Disruption | No Appeal | 0 | 0 | 0 |
| Technical Issues | Grace Period Extension or No Appeal | 4 | 1 | 0 |
| Validity Issue | Invalidate or Reset | 1 | 1 | 1 |

Table 4.7 and table 4.8 present the number of Appeals approved and rejected in STAIRS in the 2020–2021 administration, respectively.

Table 4.7 Number of Appeals Approved in STAIRS in the 2020–2021 Administration

|  |  |
| --- | --- |
| Appeal Type | Number of Appeals Approved for the CSA |
| Reset | 2 |
| Re-open | 23 |
| Invalidate | 2 |
| Grace Period Extension | 0 |
| Restore | 0 |

Table 4.8 Number of Appeals Rejected in STAIRS in the 2020–2021 Administration

|  |  |
| --- | --- |
| Appeal Type | Number of Appeals Rejected for the CSA |
| Reset | 0 |
| Re-open | 0 |
| Invalidate | 0 |
| Grace Period Extension | 1 |
| Restore | 0 |

### 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 Department of Education. (2020). *California assessment accessibility resources matrix*. California Department of Education website.

California Department of Education. (2021a). *CAASPP and ELPAC accessibility guide for online testing*. Sacramento, CA: California Department of Education.

California Department of Education. (2021b). *CAASPP and ELPAC security incidents and appeals procedure guide.* Sacramento, CA: California Department of Education.

California Department of Education. (2021c). *CAASPP and ELPAC technical specifications and configuration guide*. Sacramento, CA: California Department of Education.

California Department of Education. (2021d). *CAASPP and ELPAC Test Operations Management System user guide*. Sacramento, CA: California Department of Education.

California Department of Education. (2021e). *CAASPP online test administration manual.* Sacramento, CA: California Department of Education.

California Department of Education. (2021f). *Spring administration information for educators.* 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). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations implementation guide.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes.

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

Smarter Balanced Assessment Consortium. (2020). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Los Angeles, CA: Smarter Balanced Assessment Consortium and National Center on Educational Outcomes.

### Appendix 4.A: Demographic Summaries

**Notes:**

* Data collected for Spanish instruction program types and percentage of daily instruction in Spanish is derived from the student survey as part of the operational assessment.
* Percent of valid scores for individual student demographic groups may not sum to 100 because of rounding.
* Data for the category of “Received in Spanish in the 2020–2021 school year—Spanish as a Foreign Language Program” are available for grades six through eight and for high school.

Table 4.A.1 **Demographic Summary, Grade Three**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 1,213 | 100.00% |
| Male | 581 | 47.90% |
| Female | 632 | 52.10% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 9 | 0.74% |
| Asian | 19 | 1.57% |
| Native Hawaiian or Other Pacific Islander | 0 | 0.00% |
| Filipino | 6 | 0.49% |
| Hispanic or Latino | 1,002 | 82.61% |
| Black or African American | 12 | 0.99% |
| White | 133 | 10.96% |
| Two or more races | 32 | 2.64% |
| Unknown | 0 | 0.00% |
| English only | 388 | 31.99% |
| IFEP | 57 | 4.70% |
| EL | 622 | 51.28% |
| RFEP | 144 | 11.87% |
| Ever-ELs (EL or RFEP) | 766 | 63.15% |
| To be determined | 2 | 0.16% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 1,116 | 92.00% |
| Special education services | 97 | 8.00% |
| Not economically disadvantaged | 431 | 35.53% |
| Economically disadvantaged | 782 | 64.47% |
| In U.S. schools less than 12 months | 9 | 0.74% |
| In U.S. schools 12 months or more | 1,204 | 99.26% |

Table 4.A.1 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 1,136 | 93.65% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 71 | 5.85% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 906 | 74.69% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 88 | 7.25% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 18 | 1.48% |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 38 | 3.13% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 303 | 24.98% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 476 | 39.24% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 319 | 26.30% |

Table 4.A.2 **Demographic Summary, Grade Four**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 1,052 | 100.00% |
| Male | 511 | 48.57% |
| Female | 541 | 51.43% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 4 | 0.38% |
| Asian | 13 | 1.24% |
| Native Hawaiian or Other Pacific Islander | 1 | 0.10% |
| Filipino | 5 | 0.48% |
| Hispanic or Latino | 885 | 84.13% |
| Black or African American | 11 | 1.05% |
| White | 106 | 10.08% |
| Two or more races | 27 | 2.57% |
| Unknown | 0 | 0.00% |
| English only | 326 | 30.99% |
| IFEP | 46 | 4.37% |
| EL | 463 | 44.01% |
| RFEP | 216 | 20.53% |
| Ever-ELs (EL or RFEP) | 679 | 64.54% |
| To be determined | 0 | 0.00% |
| English proficiency unknown | 1 | 0.10% |
| No special education services | 963 | 91.54% |
| Special education services | 89 | 8.46% |
| Not economically disadvantaged | 368 | 34.98% |
| Economically disadvantaged | 684 | 65.02% |
| In U.S. schools less than 12 months | 7 | 0.67% |
| In U.S. schools 12 months or more | 1,045 | 99.33% |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 1,001 | 95.15% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 45 | 4.28% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 789 | 75.00% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 73 | 6.94% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 18 | 1.71% |

Table 4.A.2 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 29 | 2.76% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 492 | 46.77% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 342 | 32.51% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 138 | 13.12% |

Table 4.A.3 **Demographic Summary, Grade Five**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 916 | 100.00% |
| Male | 448 | 48.91% |
| Female | 468 | 51.09% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 9 | 0.98% |
| Asian | 10 | 1.09% |
| Native Hawaiian or Other Pacific Islander | 0 | 0.00% |
| Filipino | 0 | 0.00% |
| Hispanic or Latino | 769 | 83.95% |
| Black or African American | 6 | 0.66% |
| White | 98 | 10.70% |
| Two or more races | 24 | 2.62% |
| Unknown | 0 | 0.00% |
| English only | 300 | 32.75% |
| IFEP | 42 | 4.59% |
| EL | 352 | 38.43% |
| RFEP | 222 | 24.24% |
| Ever-ELs (EL or RFEP) | 574 | 62.66% |
| To be determined | 0 | 0.00% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 848 | 92.58% |
| Special education services | 68 | 7.42% |
| Not economically disadvantaged | 316 | 34.50% |
| Economically disadvantaged | 600 | 65.50% |
| In U.S. schools less than 12 months | 2 | 0.22% |
| In U.S. schools 12 months or more | 914 | 99.78% |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 887 | 96.83% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 17 | 1.86% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 705 | 76.97% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 91 | 9.93% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 14 | 1.53% |

Table 4.A.3 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 31 | 3.38% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 475 | 51.86% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 274 | 29.91% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 107 | 11.68% |

Table 4.A.4 **Demographic Summary, Grade Six**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 948 | 100.00% |
| Male | 471 | 49.68% |
| Female | 477 | 50.32% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 3 | 0.32% |
| Asian | 11 | 1.16% |
| Native Hawaiian or Other Pacific Islander | 0 | 0.00% |
| Filipino | 1 | 0.11% |
| Hispanic or Latino | 806 | 85.02% |
| Black or African American | 9 | 0.95% |
| White | 92 | 9.70% |
| Two or more races | 26 | 2.74% |
| Unknown | 0 | 0.00% |
| English only | 282 | 29.75% |
| IFEP | 38 | 4.01% |
| EL | 296 | 31.22% |
| RFEP | 331 | 34.92% |
| Ever-ELs (EL or RFEP) | 627 | 66.14% |
| To be determined | 1 | 0.11% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 889 | 93.78% |
| Special education services | 59 | 6.22% |
| Not economically disadvantaged | 364 | 38.40% |
| Economically disadvantaged | 584 | 61.60% |
| In U.S. schools less than 12 months | 6 | 0.63% |
| In U.S. schools 12 months or more | 942 | 99.37% |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 903 | 95.25% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 7 | 0.74% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 770 | 81.22% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 61 | 6.43% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 5 | 0.53% |
| Received Instruction in Spanish in the 2020–2021 School Year—Spanish as a Foreign Language Program | 15 | 1.58% |

Table 4.A.4 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 89 | 9.39% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 497 | 52.43% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 213 | 22.47% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 104 | 10.97% |

Table 4.A.5 **Demographic Summary, Grade Seven**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 729 | 100.00% |
| Male | 349 | 47.87% |
| Female | 380 | 52.13% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 0 | 0.00% |
| Asian | 10 | 1.37% |
| Native Hawaiian or Other Pacific Islander | 3 | 0.41% |
| Filipino | 3 | 0.41% |
| Hispanic or Latino | 603 | 82.72% |
| Black or African American | 6 | 0.82% |
| White | 85 | 11.66% |
| Two or more races | 19 | 2.61% |
| Unknown | 0 | 0.00% |
| English only | 234 | 32.10% |
| IFEP | 25 | 3.43% |
| EL | 161 | 22.09% |
| RFEP | 309 | 42.39% |
| Ever-ELs (EL or RFEP) | 470 | 64.47% |
| To be determined | 0 | 0.00% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 697 | 95.61% |
| Special education services | 32 | 4.39% |
| Not economically disadvantaged | 320 | 43.90% |
| Economically disadvantaged | 409 | 56.10% |
| In U.S. schools less than 12 months | 1 | 0.14% |
| In U.S. schools 12 months or more | 728 | 99.86% |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 717 | 98.35% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 18 | 2.47% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 604 | 82.85% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 54 | 7.41% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 4 | 0.55% |
| Received Instruction in Spanish in the 2020–2021 School Year—Spanish as a Foreign Language Program | 11 | 1.51% |

Table 4.A.5 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 109 | 14.95% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 328 | 44.99% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 193 | 26.47% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 87 | 11.93% |

Table 4.A.6 **Demographic Summary, Grade Eight**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 572 | 100.00% |
| Male | 286 | 50.00% |
| Female | 286 | 50.00% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 1 | 0.17% |
| Asian | 5 | 0.87% |
| Native Hawaiian or Other Pacific Islander | 1 | 0.17% |
| Filipino | 2 | 0.35% |
| Hispanic or Latino | 495 | 86.54% |
| Black or African American | 3 | 0.52% |
| White | 59 | 10.31% |
| Two or more races | 6 | 1.05% |
| Unknown | 0 | 0.00% |
| English only | 133 | 23.25% |
| IFEP | 30 | 5.24% |
| EL | 107 | 18.71% |
| RFEP | 302 | 52.80% |
| Ever-ELs (EL or RFEP) | 409 | 71.50% |
| To be determined | 0 | 0.00% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 546 | 95.45% |
| Special education services | 26 | 4.55% |
| Not economically disadvantaged | 212 | 37.06% |
| Economically disadvantaged | 360 | 62.94% |
| In U.S. schools less than 12 months | 2 | 0.35% |
| In U.S. schools 12 months or more | 570 | 99.65% |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 559 | 97.73% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 8 | 1.40% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 475 | 83.04% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 33 | 5.77% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 5 | 0.87% |
| Received Instruction in Spanish in the 2020–2021 School Year—Spanish as a Foreign Language Program | 16 | 2.80% |

Table 4.A.6 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 88 | 15.38% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 331 | 57.87% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 79 | 13.81% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 61 | 10.66% |

Table 4.A.7 **Demographic Summary, High School**

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| All students | 253 | 100.00% |
| Male | 122 | 48.22% |
| Female | 131 | 51.78% |
| Nonbinary | 0 | 0.00% |
| American Indian or Alaska Native | 0 | 0.00% |
| Asian | 0 | 0.00% |
| Native Hawaiian or Other Pacific Islander | 0 | 0.00% |
| Filipino | 0 | 0.00% |
| Hispanic or Latino | 245 | 96.84% |
| Black or African American | 0 | 0.00% |
| White | 7 | 2.77% |
| Two or more races | 1 | 0.40% |
| Unknown | 0 | 0.00% |
| English only | 20 | 7.91% |
| IFEP | 3 | 1.19% |
| EL | 91 | 35.97% |
| RFEP | 135 | 53.36% |
| Ever-ELs (EL, RFEP, or ADEL) | 226 | 89.33% |
| ADEL | 0 | 0.00% |
| To be determined | 4 | 1.58% |
| English proficiency unknown | 0 | 0.00% |
| No special education services | 237 | 93.68% |
| Special education services | 16 | 6.32% |
| Not economically disadvantaged | 35 | 13.83% |
| Economically disadvantaged | 218 | 86.17% |
| In U.S. schools less than 12 months | 35 | 13.83% |
| In U.S. schools 12 months or more | 218 | 86.17% |

Table 4.A.7 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Student Group | Number of Valid Scores | Percent of Valid Scores |
| Received Instruction in Spanish in the 2020–2021 School Year—Total | 217 | 85.77% |
| Received Instruction in Spanish in the 2020–2021 School Year—One-Way Immersion Program | 23 | 9.09% |
| Received Instruction in Spanish in the 2020–2021 School Year—Dual-Language Immersion Program | 79 | 31.23% |
| Received Instruction in Spanish in the 2020–2021 School Year—Developmental Bilingual Program | 22 | 8.70% |
| Received Instruction in Spanish in the 2020–2021 School Year—Heritage Language or Indigenous Language Program | 10 | 3.95% |
| Received Instruction in Spanish in the 2020–2021 School Year—Spanish as a Foreign Language Program | 62 | 24.51% |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 44 | 17.39% |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 79 | 31.23% |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 33 | 13.04% |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 61 | 24.11% |

### Appendix 4.B: Special Service Summaries

Table 4.B.1 Special Services Summary, Grades Three Through Five

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accessibility Resource | Grade 3 Number | Grade 3 Pct. of Total | Grade 4 Number | Grade 4 Pct. of Total | Grade 5 Number | Grade 5 Pct. of Total |
| Embedded Accommodation—Audio Transcript | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Accommodation—Braille | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Accommodation—Closed Captioning | 1 | 0.08 | 0 | 0.00 | 1 | 0.11 |
| Embedded Accommodation—Text-to-Speech | 6 | 0.49 | 7 | 0.67 | 8 | 0.87 |
| Non-Embedded Accommodation—Alternate Response Options | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Accommodation—Print on Demand | 0 | 0.00 | 1 | 0.10 | 0 | 0.00 |
| Non-Embedded Accommodation—Read Aloud (for Passages) | 5 | 0.41 | 4 | 0.38 | 7 | 0.76 |
| Non-Embedded Accommodation—Scribe (for Writing) | 0 | 0.00 | 2 | 0.19 | 2 | 0.22 |
| Embedded Designated Support—Color Contrast | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Designated Support—Masking | 1 | 0.08 | 3 | 0.29 | 2 | 0.22 |
| Embedded Designated Support—Mouse Pointer | 0 | 0.00 | 1 | 0.10 | 0 | 0.00 |
| Embedded Designated Support—Permissive Mode | 0 | 0.00 | 2 | 0.19 | 0 | 0.00 |
| Embedded Designated Support—Print Size | 0 | 0.00 | 1 | 0.10 | 0 | 0.00 |
| Embedded Designated Support—Streamlining | 3 | 0.25 | 6 | 0.57 | 6 | 0.66 |
| Embedded Designated Support—Text-to-Speech | 16 | 1.32 | 11 | 1.05 | 12 | 1.31 |
| Embedded Designated Support—Turn off Universal Tools | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |

Table 4.B.1 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accessibility Resource | Grade 3 Number | Grade 3 Pct. of Total | Grade 4 Number | Grade 4 Pct. of Total | Grade 5 Number | Grade 5 Pct. of Total |
| Non-Embedded Designated Support—Amplification | 1 | 0.08 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Color Contrast | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Color Overlay | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Magnification | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Medical Device | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Noise Buffers | 18 | 1.48 | 16 | 1.52 | 11 | 1.20 |
| Non-Embedded Designated Support—Read Aloud | 18 | 1.48 | 12 | 1.14 | 7 | 0.76 |
| Non-Embedded Designated Support—Scribe | 2 | 0.16 | 2 | 0.19 | 1 | 0.11 |
| Non-Embedded Designated Support—Separate Setting | 27 | 2.23 | 25 | 2.38 | 23 | 2.51 |
| Other—Unlisted Resources | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Other—Designated support or accommodation is in IEP | 25 | 2.06 | 28 | 2.66 | 25 | 2.73 |
| Other—Designated support or accommodation is in Section 504 plan | 0 | 0.00 | 0 | 0.00 | 1 | 0.11 |

Table 4.B.2 Special Services Summary, Grades Six Through Eight

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accessibility Resource | Grade 6 Number | Grade 6 Pct. of Total | Grade 7 Number | Grade 7 Pct. of Total | Grade 8 Number | Grade 8 Pct. of Total |
| Embedded Accommodation—Audio Transcript | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Accommodation—Braille | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Accommodation—Closed Captioning | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Accommodation—Text-to-Speech | 8 | 0.84 | 11 | 1.51 | 11 | 1.92 |
| Non-Embedded Accommodation—Alternate Response Options | 0 | 0.00 | 0 | 0.00 | 1 | 0.17 |
| Non-Embedded Accommodation—Print on Demand | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Accommodation—Read Aloud (for Passages) | 4 | 0.42 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Accommodation—Scribe (for Writing) | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Designated Support—Color Contrast | 1 | 0.11 | 0 | 0.00 | 0 | 0.00 |
| Embedded Designated Support—Masking | 0 | 0.00 | 1 | 0.14 | 3 | 0.52 |
| Embedded Designated Support—Mouse Pointer | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Embedded Designated Support—Permissive Mode | 0 | 0.00 | 1 | 0.14 | 0 | 0.00 |
| Embedded Designated Support—Print Size | 1 | 0.11 | 0 | 0.00 | 0 | 0.00 |
| Embedded Designated Support—Streamlining | 7 | 0.74 | 9 | 1.23 | 12 | 2.10 |
| Embedded Designated Support—Text-to-Speech | 11 | 1.16 | 12 | 1.65 | 12 | 2.10 |
| Embedded Designated Support—Turn off Universal Tools | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |

Table 4.B.2 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accessibility Resource | Grade 6 Number | Grade 6 Pct. of Total | Grade 7 Number | Grade 7 Pct. of Total | Grade 8 Number | Grade 8 Pct. of Total |
| Non-Embedded Designated Support—Amplification | 1 | 0.11 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Color Contrast | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Color Overlay | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Magnification | 1 | 0.11 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Medical Device | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Noise Buffers | 3 | 0.32 | 9 | 1.23 | 2 | 0.35 |
| Non-Embedded Designated Support—Read Aloud | 6 | 0.63 | 1 | 0.14 | 0 | 0.00 |
| Non-Embedded Designated Support—Scribe | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Non-Embedded Designated Support—Separate Setting | 17 | 1.79 | 12 | 1.65 | 16 | 2.80 |
| Other—Unlisted Resources | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Other—Designated support or accommodation is in IEP | 18 | 1.90 | 16 | 2.19 | 14 | 2.45 |
| Other—Designated support or accommodation is in Section 504 plan | 3 | 0.32 | 1 | 0.14 | 4 | 0.70 |

Table 4.B.3 Special Services Summary, High School

|  |  |  |
| --- | --- | --- |
| Accessibility Resource | High School Number | High School Pct. of Total |
| Embedded Accommodation—Audio Transcript | 0 | 0.00 |
| Embedded Accommodation—Braille | 0 | 0.00 |
| Embedded Accommodation—Closed Captioning | 0 | 0.00 |
| Embedded Accommodation—Text-to-Speech | 0 | 0.00 |
| Non-Embedded Accommodation—Alternate Response Options | 0 | 0.00 |
| Non-Embedded Accommodation—Print on Demand | 0 | 0.00 |
| Non-Embedded Accommodation—Read Aloud (for Passages) | 0 | 0.00 |
| Non-Embedded Accommodation—Scribe (for Writing) | 0 | 0.00 |
| Embedded Designated Support—Color Contrast | 0 | 0.00 |
| Embedded Designated Support—Masking | 0 | 0.00 |
| Embedded Designated Support—Mouse Pointer | 0 | 0.00 |
| Embedded Designated Support—Permissive Mode | 0 | 0.00 |
| Embedded Designated Support—Print Size | 1 | 0.40 |
| Embedded Designated Support—Streamlining | 0 | 0.00 |
| Embedded Designated Support—Text-to-Speech | 0 | 0.00 |
| Embedded Designated Support—Turn off Universal Tools | 0 | 0.00 |

Table 4.B.3 *(continuation)*

|  |  |  |
| --- | --- | --- |
| Accessibility Resource | High School Number | High School Pct. of Total |
| Non-Embedded Designated Support—Amplification | 0 | 0.00 |
| Non-Embedded Designated Support—Color Contrast | 0 | 0.00 |
| Non-Embedded Designated Support—Color Overlay | 0 | 0.00 |
| Non-Embedded Designated Support—Magnification | 0 | 0.00 |
| Non-Embedded Designated Support—Medical Device | 0 | 0.00 |
| Non-Embedded Designated Support—Noise Buffers | 0 | 0.00 |
| Non-Embedded Designated Support—Read Aloud | 0 | 0.00 |
| Non-Embedded Designated Support—Scribe | 0 | 0.00 |
| Non-Embedded Designated Support—Separate Setting | 0 | 0.00 |
| Other—Unlisted Resources | 0 | 0.00 |
| Other—Designated support or accommodation is in IEP | 0 | 0.00 |
| Other—Designated support or accommodation is in Section 504 plan | 0 | 0.00 |

## Standard Setting

This chapter summarizes the standard setting process, undertaken in the 2018–2019 administration, through which California Spanish Assessment (CSA) achievement levels and threshold scores were recommended. Included are an overview of the standard setting methodology, a summary of the standard setting procedure, a description of the achievement level descriptors (ALDs), and the results. The detailed standard setting information for the CSA is described in the *Standard Setting Technical Report for the California Spanish Assessment* (California Department of Education [CDE], 2019).

### Background

The content of the CSA is aligned with the California Common Core State Standards en Español (CCCSSeE) (Council of Chief State School Officers, CDE, and San Diego County Office of Education, 2012). The CDE and the administration of the CSA required a standard setting process to evaluate students’ Spanish skills in reading, writing mechanics, and listening against the new expectations.

Standard setting refers to a class of methodologies by which one or more performance threshold scores are used to determine achievement levels. The purpose of the standard setting process for the CSA was to collect recommendations from Spanish-language educators in California for the placement of the CSA threshold scores for review by the CDE, with final determination by the California State Board of Education (SBE).

ETS conducted standard setting workshops from August 6–9, 2019, following the first operational administration of the CSA. The Bookmark standard setting method was applied to all items on each test, by grade. Refer to section [*5.3 Standard Settin**g Methodology*](#_Standard_Setting_Methodology) for more information about the Bookmark method.

Through the standard setting process, input and recommendations on the threshold scores were solicited from Spanish-language educators in California. The CDE reviewed the educator input and recommendations and provided these recommendations to the SBE along with recommendations from the State Superintendent of Public Instruction (SSPI). The SSPI recommendation was to establish preliminary score reporting ranges beginning with the 2018–2019 administration, giving consideration to the sample size of test takers and the wide variety of educational programs across local educational agencies in California. The SBE established the reporting ranges based on the SSPI preliminary threshold score recommendations. There are three reporting ranges (Range 1 through Range 3); two preliminary threshold scores are needed to define the three ranges. Students with scale scores lower than the preliminary threshold score for Range 2 are assigned to the lowest score range, Range 1. Students with scale scores that are equal to or greater than the preliminary threshold score for Range 3 are assigned to the highest score range, Range 3. The rest of the students with valid scores are assigned to Range 2.

### Achievement Level Descriptors

The CSA ALDs describe expectations of what students can do at each level. The general, or policy, ALDs are short policy descriptors that convey the expectation across all grades and were approved by the California SBE in November 2017 (CDE, 2017). From July 18–‍19, 2018, 21 California educators convened in Sacramento to review and provide input on the range ALDs, which are descriptions of the Spanish reading/language arts knowledge and skills necessary for students in grades three through eight and high school to be placed into one of three achievement levels. These range ALDs were used to inform the standard setting process.

Appendix 5.A of the *California Spanish Assessment 2018–2019 Technical Report* (CDE, 2020) provides a description of the three range ALDs, with Level 3 reflecting the highest level of achievement.

### Standard Setting Methodology

For the CSA, the Bookmark method was used for standard setting. The Bookmark method is an item-mapping procedure that allows multiple performance threshold scores to be set in an efficient manner. This method represents an appropriate balance between statistical rigor and informed opinion, as explained in the following subsection. In the case of the CSA, three of four panels worked on two tests (i.e., grades three and four, grades five and six, and grades seven and eight; the high school panel worked only on that test).

#### Bookmark Method

The Bookmark method (Lewis, et al., 1998; Mitzel, et al., 2001) is a commonly used item-mapping procedure in which test items are ordered from easiest to most difficult based on actual student performance; the ordered items are presented in a booklet known as an ordered item booklet (OIB). The task of each panelist is to place a “bookmark” in the OIB that differentiates content that a student with just enough content knowledge and skills to be performing at a defined achievement level would likely know from content that the student would not likely know. A bookmark is placed in the OIB for each item defined at the border of each achievement level. For each CSA, two bookmarks were required to set the three achievement levels.

The Bookmark method has its basis in item response theory (IRT) analysis. IRT is used to estimate item difficulties. Based on the first-year operational test data, a response probability of 0.67 (RP67) estimated by the IRT model was employed to order the items from easiest to hardest and to place item difficulty estimates on the score scale. Panelists were instructed to consider the definition of “most likely” as having a two-thirds likelihood of answering a multiple-choice (MC) item correctly; thus, the instructions to the panelists and the analytical model were aligned. One benefit of this approach is that once panelists make judgments in the OIB, the difficulty values associated with each item have a built-in relationship to scale scores through theta, a fact that allows results to be provided to score users and policy makers on the familiar metric of the scale score.

### Standard Setting Procedures

This section describes what occurred prior to and during the standard setting workshop.

#### Panelists

A diverse group, representative of Spanish-language educators in California, was recruited to participate as panelists in the standard setting sessions. In recruiting panelists, the goal was to include a representative group of California educators who were familiar with the CCCSSeE and who have experience in the education of students in grades three through twelve who will take the CSA. It was important to include teachers working with these students, as those educators provided a perspective on learning goals for the students taking the CSA, as well as students’ progress toward Spanish reading/language arts proficiency.

The educators who participated in the CSA standard setting included representatives from across regions in California (north, south, and central) and across gender, race, and ethnic categories. The composition of each panel included the following as criteria for selection:

* Educators who were teaching Spanish-language learners, in the grade level(s) assigned to the panel
* Educators who were teaching students who would take the CSA
* Educators who were familiar with the CCCSSeE

The final selection of panelists invited to the workshops was made by the CDE. The total number of panelists who participated and completed the CSA standard setting process was 56.

#### Materials

Panelists were provided with a letter describing the purpose and procedures of the standard setting workshop along with a preworkshop assignment specific to the individual educator’s panel assignment, instructions, a notetaking form for the assignment, and the links to the training tests and to the general and range ALDs for the tests the panelists would be reviewing.

During the workshop, panelists received training materials and a set of operational materials. The set of operational materials included a printed version of the CSA and the answer key with scoring rules for 2-point items, the OIB, judgment recording forms, and an item map. The detailed procedure with regard to securing those materials was described in the *Standard Setting Technical Report for the California Spanish Assessment* (CDE, 2019).

#### Process (Including Articulation)

Prior to making judgements in the OIB, as part of a preworkshop assignment, panelists were provided with a link to the CSA training test on the CDE website and asked to take the training test for the grade level the panelists were scheduled to work with first. Panelists were also asked to become familiar with the general ALDs and the range ALDs and to access a link to the CCCSSeE. Panelists were asked to consider the expectations of a student in each of the achievement levels, take notes about the knowledge and skills of students at the beginning of Level 2 and Level 3, and bring those notes to the standard setting workshop.

At the workshop, each panel began with the test familiarization by reviewing one of the two tests assigned to that panel, and then they developed borderline student definitions as a group for that grade level. The process to arrive at borderline student definitions involved small-group discussions and the development of draft borderline-student definitions, followed by a whole-panel discussion of the draft definitions to reach a panel consensus of what was expected. For each CSA grade level or the grade band, two definitions were developed for two thresholds—the Level 3 borderline student definition followed by the Level 2 borderline student definition. The Level 3 definition was developed first to allow cross grade articulation early in the process.

After the “borderline Level 3 student” definition was drafted, two pairs of two panels working on adjacent CSA grade-level or grade-band tests met to discuss the drafts, provide feedback to each other, and finalize the definitions. These discussions and this work focused on cross-grade consistency of the ALDs and the description of the borderline student for Level 3. Each panel then reconvened and completed the “borderline Level 2 student” definition.

Each panel, with the exception of the high school panel, completed the standard setting process on two CSA grade-level tests. After completing the process for the first of two grade levels on which the panel completed the standard setting process (i.e., grade three, or grade five or grade seven), the panel began the entire process again with the second assessment (e.g., grade four). The grades five and six panel and the grades seven and eight panel met again to consider cross-grade consistency when creating the borderline student definitions for their second CSA (for grades six and seven, respectively). The process of developing the borderline student definitions provided vertical articulation of the expectations across grades prior to bookmark judgments.

To make judgments and place bookmarks in the OIB, panelists reviewed each item in the OIB in sequence and considered whether the student at the beginning of Level 2, known as the borderline Level 2 student, would most likely be able to answer the item correctly. A panelist placed the Level 2 bookmark on the first item encountered in the OIB that the panelist believed the borderline Level 2 student would most likely not be able to address because items beyond that point were too difficult for that borderline student. The panelist continued from that point in the OIB and then stopped at the item that the borderline Level 3 student would not likely be able to address (i.e., the item that likely exceeds the ability of the borderline Level 3 student). Note that in the Bookmark method, the definition of “most likely” is related to the IRT model. That is, panelists were instructed to think of “most likely” as having a two-thirds likelihood of answering an MC item correctly. In ordering the items in the OIB, a response probability of 0.67 (RP67) is employed in the IRT model as recommended by research; thus, the instructions to the panelists and the analytical model are aligned.[[7]](#footnote-8)

The Bookmark process was implemented in three rounds with feedback and discussion between rounds. The final recommended threshold scores were based on the median of panelists’ judgment scores. The last step in the workshop involved a subset of panelists from each panel room that were recruited to attend the cross-grade articulation meeting. The goal of this meeting was to ask panelists to consider the score recommendations by considering feedback and data across the seven sets of threshold score recommendations. Panelists were provided with the borderline student definitions across all panels. The panelists were asked to review the definitions for the assigned grade levels or the grade band, along with the two adjacent grade levels or the grade band. The panel facilitator asked the panelists to share the rationales and the discussions that occurred in each panel. Panelists next reviewed the impact data for all seven sets of threshold scores.

As part of the standard setting process, the CDE analyzed the standard setting panel’s judgments and refined the threshold scores for consistency across all the CSA grade levels tested. The CDE’s recommendations were then presented to the SBE as preliminary threshold scores for approval.

### Results of the Standard Setting Process

The results of the standard setting were educator-panel recommendations of threshold scores; these recommendations and the SSPI recommendations were presented to the SBE for review. The SBE approved the recommendation of the score reporting ranges for the CSA based on the SSPI’s recommended preliminary threshold scores. The recommendations of the SSPI are presented in table 5.1. The scale scores in this table were presented and used in the standard setting process and are not the official reporting scale. The standard setting working scale ranges from 300 to 500 score points and is more user friendly than the theta metric. The official scale score reporting scale was developed and approved after the SBE approval of score ranges.

Table 5.1 shows the percentage of students statewide that would be placed in each of the three score ranges, identified in this table as ranges 1 through 3, on the basis of the results of the 2018–2019 CSA administration. Also shown in this table is the percentage of students statewide that would be at and above Range 2 on the basis of the results of the 2018–‍2019 administration. Finally, the preliminary threshold score is the minimum standard setting scale score needed to reach this range on the 2018–‍2019 administration of tests. Note that threshold scores were generated solely for the standard setting process; reporting scales were later developed to report scores on the Student Score Report and public reporting. The definition of the score reporting range is included in section [*2.7 Standard Setting*](#_Standard_Setting_1).

Table 5.1 SSPI Recommendations for the Preliminary Threshold Scores and Three Scale Score Ranges on the CSA

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Percent of Students in Range 1 | Percent of Students in Range 2 | Standard Setting Scale Threshold Score for Range 2 | Percent at or Above Range 2 | Percent of Students in Range 3 | Standard Setting Scale Threshold Score for Range 3 |
| Grade 3 | 52.7 | 33.0 | 401 | 47.3 | 14.3 | 413 |
| Grade 4 | 53.5 | 31.5 | 401 | 46.5 | 15.0 | 414 |
| Grade 5 | 45.6 | 40.8 | 398 | 54.4 | 13.6 | 413 |
| Grade 6 | 41.4 | 40.8 | 398 | 58.6 | 17.8 | 411 |
| Grade 7 | 58.2 | 37.0 | 402 | 41.8 | 4.8 | 418 |
| Grade 8 | 57.4 | 32.9 | 402 | 42.6 | 9.7 | 415 |
| High school | 59.6 | 31.4 | 403 | 40.4 | 9.0 | 414 |

The reporting score ranges at different grades in a typical administration are presented in table 6.2. The scale score ranges are developed as part of the scaling, based on the preliminary threshold scores, and do not change from year to year. For the CSA, they will remain unchanged from administration to administration until more information about the test-taking population and other factors becomes available and new performance standards and achievement levels are adopted.

### References

California Department of Education. (2017, November 8). *California State Board of Education: Final minutes: November 8–9, 2017*. California Department of Education website.

California Department of Education. (2019). *Standard setting technical report for the California Spanish Assessment.* California Department of Education website.

California Department of Education. (2020). *California Spanish Assessment 2018–2019 technical report.* [Unpublished report]. Sacramento, CA: California Department of Education.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications.

Council of Chief State School Officers, California Department of Education, and San Diego County Office of Education. (2012). *Common Core State Standards in language arts and literacy in history/social studies, science, and technical subjects: Spanish language version*.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The Bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the 1998 annual meeting of the National Council on Measurement, San Diego, CA.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–81). Mahwah, NJ: Erlbaum.

## Scoring and Reporting

Student item responses were scored and analyzed to determine individual students’ scores for the California Spanish Assessment (CSA). On the basis of the analyses of the item responses, individual student scores (i.e., overall reporting scores) were calculated and reported. In addition, student test scores were aggregated to produce summary reports for local educational agencies (LEAs).

This chapter summarizes the scoring at the item level in the CSA and the approach implemented to produce student scores. This chapter also describes scores reported at the individual student level and various reports that were generated for 2020–2021 CSA administration.

### Student Test Scores

Overall reporting scores for the CSA were produced at the individual student level. To obtain overall reporting scores, the ability (theta) scores needed to be estimated.

ETS Assessment and Learning Technology Research & Development staff reviewed each item and determined the answer keys. The keys were provided to Cambium Assessment, Inc. (CAI) for implementation in the test delivery system (TDS). After CAI finished machine-scoring item responses, scores and responses were delivered to ETS. ETS’ Enterprise Score Key Management (eSKM) system collected and calculated individual students’ overall scores (e.g., total raw scores). ETS’ Psychometric Analysis & Research (PAR) team conducted a series of psychometric analyses such as calibration, equating, and scaling using individual item scores of the test samples and produced the raw-to-scale score conversion tables based on all psychometric analyses. When the conversion tables were produced, eSKM produced the scale score and the reporting score ranges for students who completed the test.

ETS used two parallel scoring systems to produce and verify students’ scores. The eSKM scoring system received individual students’ item scores and item responses from CAI and computed individual student scores for the ETS reporting system. ETS’ PAR team also computed individual student scores based on the same data files using statistical analysis system software. The scores from the two systems were then compared for the purpose of internal quality control. Inconsistency in the total raw scores was investigated and resolved. The parallel scoring process ensured the quality and accuracy of scoring and supported the transfer of scores into the database of the student records scoring system, the Test Operations Management System (TOMS).

#### Scoring of Incomplete Cases

Whether a test should be scored or reported depended on the “complete” status of the test and how much of the test was submitted for scoring. Depending on the nature of the missing data, different actions were taken.

As defined in the CSA scoring and reporting specifications, for a typical test administration, tests are considered “complete” and students are scored if students respond to at least 10 items. Tests are also considered “complete” with the lowest obtainable scale score (LOSS); students are assigned the LOSS if students respond to at least 1 but fewer than 10 items. Tests are considered “partially complete,” and no scores are assigned, if students log on to the test but answer no items. Tests are considered “non-complete” if students did not log on to the test.

ETS, in consultation with the California Department of Education (CDE), implemented several rules to identify an incomplete test; these rules are represented in table 6.1, which included the following four specifications:

1. Attemptedness and participation rules describing when a test is considered attempted or taken
2. When a test is scored
3. Whether incomplete tests are scored
4. When a score is reported

Table 6.1 Rules for Incomplete Tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| If the student | Classify the student as taking the test? | Score the student’s responses? | Classify the student as attempting the test? | Report a score for the student? |
| Logged on to the test and answered at least 1 item but fewer than 10 items | Yes | Yes, LOSS for the test | Yes (Completion with LOSS) | Yes (LOSS) |
| Logged on to the test and answered at least 10 items | Yes | Yes | Yes (Completion) | Yes (Actual scores) |
| Logged on to the test but answered no items | No | N/A | Partial Completion | No |
| Did not log on to the test | No | N/A | Noncompletion | No |
| Logged on and answered at least one item with a special condition code (refer to subsection [*6.3.2 Special Cases*](#_Special_Cases)*)* | No | N/A | Not Tested | No |

#### Theta Scores

Typically, a student’s raw score is the sum of scores on the individual items presented to the student. A theta score is derived from an item response theory (IRT) model and indicates a student’s ability level on the IRT scale. The test for each grade level—grades three through eight and high school—has its own theta scale. When all the items presented to a student are calibrated onto that theta scale, the student’s raw score can be transformed into an ability (theta) estimate. Refer to subsection [*7.4.2 Equating*](#_Equating_1) for details.

After all operational items are calibrated and linked onto the base scale, the raw score can be computed as a sum of dichotomous and polytomous item scores and can be transformed into an ability estimate (theta) by using the IRT inverse test characteristic curve (TCC) method (Stocking, 1996). With this method, the student’s estimated ability is the ability value at which the expected raw score is equal to the student’s raw score. Refer to section [*7.4 Item Response Theory Analyses*](#_Item_Response_Theory) for the scaling procedures and the IRT inverse TCC method. Note that the estimation of ability is implemented by using the item parameters of each form.

When a conversion table from the raw score to theta score is created for each form, the theta score of each individual student can be obtained in the conversion table. Refer to [appendix 6.A](#_Appendix_6.A:_Theta) for the score distribution of raw scores, theta scores, and scale scores.

#### Scale Scores for the Total Assessment

Raw scores obtained on each grade-level CSA are transformed to scale scores using the scaling process described in subsection [*7.4.2 Equating*](#_Equating). The following requirements were used to develop and define the CSA reporting scale ranges:

1. Each scale score has three digits (e.g., 320, 551, or 780) where the first digit is indicative of the grade being reported. The leading digit is defined by the grade for elementary and middle school, while the high school leading digit is set to “9.” The latter two digits represent the scale score as derived from the transformation from the raw scores to the scale scores. Refer to subsection [*7.4.2.4.2 Transformation from Theta Scores to Scale Scores*](#_7.4.2.3.2_Transformation_from) for details of the transformation.
2. Score ranges are grade specific. For example, the possible scale scores would be 300 to 399 for grade three with the LOSS at 300 and the highest obtainable scale score (HOSS) at 399. For grade four, this range is 400 to 499 with a LOSS of 400 and a HOSS of 499, and so on for the other grades. For high school grades, the scale ranges from 900 to 999 with a LOSS of 900 and a HOSS of 999.
3. The scores reported as Range 1, Range 2, and Range 3 on the reporting scale are the same from year to year. Across the grade levels, the last two digits corresponding to the score reporting range are the same, such as 360 for grade three, 460 for grade four, 560 for grade five, 660 for grade six, 760 for grade seven, 860 for grade eight, and 960 for high school.
4. Students who logged on to the test and answered at least 1 item but fewer than 10 items, as shown in table 6.1, are assigned the LOSS.

For students who complete a CSA, their scale scores cannot be lower than the LOSS or higher than the HOSS as a result of truncation in the scale score transformation listed in table 6.2. For example, the scale scores for grade three are truncated at a minimum of 300 and a maximum of 399. As a result, the range of student ability estimates [-6, +6] is transformed to the scale score range [300, 399] for grade three and [400, 499] for grade four. The scale score ranges for other grade levels and the grade band follow the same rules.

The complete raw-to-scale score conversion tables for each CSA test are presented in table 6.A.1 through table 6.A.7 in [appendix 6.A](#_Appendix_6.A:_Theta). The raw scores, theta scores, transformed scale scores, and the number and percentage of students at each raw score are listed in those tables. Because of the small test sample in high school, the score distribution table for high school is not broken down by high school grade levels. In addition, because IRT equating was not possible with high school data, theta scores were neither produced nor reported in table 6.A.7.

#### Score Reporting Ranges

CSA reporting scales currently classify each student’s performance into one of the three score reporting ranges. Achievement levels will be determined and implemented once the CSA has been administered for a few years, and a more stable test population, different types of programs, and curricula have been identified. Detailed information regarding the determination of the score reporting ranges can be found in the *Standard Setting Technical Report for the California Spanish Assessment* (CDE, 2019). The score reporting ranges for each grade level are presented in table 6.2.

Table 6.2 CSA Score Reporting Ranges by Grade Level and Grade Band

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level or Grade Band | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 |
| Grade 3 | 300–348 | 349–359 | 360–399 |
| Grade 4 | 400–448 | 449–459 | 460–499 |
| Grade 5 | 500–545 | 546–559 | 560–599 |
| Grade 6 | 600–647 | 648–659 | 660–699 |
| Grade 7 | 700–743 | 744–759 | 760–799 |
| Grade 8 | 800–847 | 848–859 | 860–899 |
| High school | 900–949 | 950–959 | 960–999 |

### Overview of Score Aggregation Procedures

For the CSA, the aggregated scores were generated for the selected student groups of interest (gender, ethnicity, English language fluency, etc.) and for the total population. This subsection contains a description of the types of aggregation that were performed on the CSA summary test scores.

#### Individual Student Score Distributions and Summary Statistics

Summary statistics that describe student performance are presented in table 6.3. Included in the table are the number of students taking each test and the means and standard deviations (SDs) of scale scores and theta scores. In general, the number of students who tested with valid scores decreases as the grade level increases. Caution should be taken when interpreting these results, because the sample sizes are small; the number of students who took the CSA during the 2020–2021 administration is about 13 percent of the number of students who took the CSA in 2018–2019 in grades three through five, 20 percent in grades six through eight, and 6 percent in high school.

Table 6.3 Mean and SD of Theta Scores and Scale Scores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade Level | Number of Students Tested with Valid Scores | Scale Score Mean | Scale Score SD | Theta Score Mean | Theta Score SD |
| Grade 3 | 1,213 | 344 | 10.8 | -0.2934 | 0.7671 |
| Grade 4 | 1,052 | 445 | 10.3 | -0.2553 | 0.7926 |
| Grade 5 | 916 | 545 | 10.2 | -0.2060 | 0.7307 |
| Grade 6 | 948 | 649 | 11.6 | -0.0372 | 0.8213 |
| Grade 7 | 729 | 742 | 11.3 | 0.0203 | 0.7580 |
| Grade 8 | 572 | 845 | 11.3 | -0.0367 | 0.7538 |
| High school—Grade 9 | 114 | 940 | 14.9 | NA | NA |
| High school—Grade 10 | 72 | 942 | 11.8 | NA | NA |
| High school—Grade 11 | 43 | 941 | 10.5 | NA | NA |
| High school—Grade 12 | 24 | 949 | 9.4 | NA | NA |
| High school—All grades | 253 | 942 | 13.1 | NA | NA |

**Notes:**

1. The incomplete cases (a student answered no test questions) are not included in the analysis.
2. There are no theta values for high school grades because the classical equating was used for high school.

The number and percentage of students at each score reporting range for each test is presented in table 6.4. More students are at score reporting range 1 than range 2 or range 3 for all grade levels, and score reporting range 3 has the fewest students. For grade six, a higher percentage of students scored in range 3, and a lower percentage of students scored in range 1, compared to the other grade levels.

Table 6.4 Numbers and Percentages of Students in Score Reporting Ranges

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Range 1 N | Range 1 % | Range 2 N | Range 2 % | Range 3 N | Range 3 % |
| Grade 3 | 867 | 71.48 | 231 | 19.04 | 115 | 9.48 |
| Grade 4 | 713 | 67.78 | 234 | 22.24 | 105 | 9.98 |
| Grade 5 | 527 | 57.53 | 299 | 32.64 | 90 | 9.83 |
| Grade 6 | 458 | 48.31 | 286 | 30.17 | 204 | 21.52 |
| Grade 7 | 406 | 55.69 | 285 | 39.09 | 38 | 5.21 |
| Grade 8 | 332 | 58.04 | 176 | 30.77 | 64 | 11.19 |
| High school—Grade 9 | 79 | 69.30 | 28 | 24.56 | 7 | 6.14 |
| High school—Grade 10 | 51 | 70.83 | 16 | 22.22 | 5 | 6.94 |
| High school—Grade 11 | 35 | 81.40 | 7 | 16.28 | 1 | 2.33 |
| High school—Grade 12 | 13 | 54.17 | 9 | 37.50 | 2 | 8.33 |
| High school—All grades | 178 | 70.36 | 60 | 23.72 | 15 | 5.93 |

#### Demographic Student Group Summaries

Statistics summarizing student performance by grade level for selected demographic student groups are provided in [appendix 6.B](#_Appendix_6.B:_Raw). In table 6.B.1 through table 6.B.7, students are grouped by demographic characteristics, including gender, ethnicity, English language fluency, economic status (disadvantaged or not), special education services status, length of enrollment in US schools, Spanish-language program type, and percentage of daily instruction in Spanish. For each demographic student group, the number of students who completed testing with a valid reporting scale score, reporting score means and SDs, and the percentage of students in each score reporting range is reported.

Table 4.1 provides definitions of the demographic student groups. To protect student privacy, when the number of students in a student group is 10 or fewer, the summary statistics are not reported and are presented as “N/A.”

### Reports Produced and Scores for Each Report

The CSA provides 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 the California Assessment of Student Performance and Progress (CAASPP) reporting for students. Scores for the CSA, as one of the components in CAASPP, are reported through the CAASPP reporting system.

#### Online Reporting

TOMS is a secure website hosted by ETS that permits LEA users to manage the CAASPP computer-based summative 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 (i.e., 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 2020–‍2021 administration: the California Educator Reporting System (CERS).

TOMS communicated with CERS, which provided authorized users with interactive and cumulative online reports for the CSA at the student, school, and LEA levels. CERS provided preliminary score data for each administered test available in the reporting system.

Based on the CSA reporting requirements, CERS provided summative reports containing information outlining student knowledge and skills, as well as reporting ranges aligned with the assessment-specific performance areas. CERS also permitted access to individual score reports, which provided preliminary score data for each administered test available in the reporting system. The online aggregated reports were available to be downloaded in PDF, Excel, and comma-separated value formats.

CERS was the primary source for LEA staff to analyze CSA results at the LEA, school, grade, classroom, or customized group level. CERS provided these reports, which could be downloaded to plan instruction. LEA staff with TOMS logon credentials could enter CERS through the CAASPP 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.
* The student’s parent/guardian requested exemption from testing.
* The student did not log on to test systems.
* The student was administered out-of-grade level tests.
* The student score was invalidated in the system (not reported in aggregated reporting).

#### Types of Score Reports

CAASPP reports fall into two categories. The specific reports within each category are presented in this subsection.

* **Student Score Report (SSR)—**The SSR was the official score report for parents/‌guardians. An SSR described the student’s results.
* **LEA student data files and aggregations—**LEA student data files were available for download on demand by the LEA in TOMS to coincide with availability of the SSRs.

##### Student Score Report

The CSA SSR is the official score report for parents/guardians and includes the following metrics:

* Reported scale scores (The ranges of scale scores are provided in table 6.2.)
* Reported reporting ranges (CSA reporting ranges are “Score Reporting Range 1,” “Score Reporting Range 2,” and “Score Reporting Range 3.”)
* A description of score reporting ranges

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 [*4.4.1 Accessibility Resource Categories*](#_Universal_Tools,_Designated).

LEAs had three options for accessing and distributing SSRs to parents/guardians:

1. Accessing electronic SSR PDFs using a locally provided parent/guardian or student portal
2. Downloading SSR PDFs 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

The LEA CAASPP coordinator could forward the appropriate reports to test sites. In the case of a locally printed CSA SSR, the LEA sent the printed report(s) to the child’s parent/‌guardian. CSA SSRs that included individual student results were not distributed beyond the student’s school.

Further information about the SSR and its interpretation is provided on the Smarter Balanced 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 forward the appropriate reports to test sites. Optionally, the LEA could download and then print the SSR PDF and then send 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 CAASPP test site coordinator to download from TOMS.

Preliminary 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 tests taken.

Current and historical aggregated results are accessible to the public on the CDE Test Results for California’s Assessments website.

#### Score Report Applications

CSA results provide parents/guardians with information about their child’s progress. The results are one tool for increasing communication and collaboration between parents/‌guardians and teachers about how to identify priorities to help the student progress in Spanish reading/language arts competency. They provided limited information about one measure of a student’s academic performance. Like any important measure of student performance, the test results should be viewed with other available information such as progress on individualized education program goals, assignments, and teacher conferences.

There may be a low, moderate, or high degree of alignment between the CSA results and the LEA’s instructional programs. Factors that determine this alignment are as follows:

* Does the LEA’s Spanish language program provide Spanish reading/language arts instruction?
* Is the LEA’s Spanish language program aligned with the California Common Core State Standards en Español?
* Is there a percentage of the LEA’s instructional day that is conducted in Spanish?

If all three statements are true, then an LEA may have a high degree of alignment between its CSA results and its instructional program. The less true the statements are, the lower the degree of the alignment.

With this in mind, schools may use the CSA results to help make decisions about how to support student achievement. CSA results, however, should never be used as the only source of information to make important decisions about a child’s education. CSA results help schools and LEAs identify strengths and weaknesses in their instructional programs.

#### Criteria for Interpreting Individual Test Scores

In a typical year, LEAs may use the CSA results to help inform decisions around instructional needs or other considerations related to student achievement, but the CSA results should not be used in isolation to make inferences about instructional needs. It is important to remember that results from a single test 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 student’s CSA results. It is also important to note that a student’s score in a content area contains measurement error and could vary to some extent 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 reporting scale scores, the user is limited to the comparison within a grade level. The user may compare reporting scale scores for the same grade within a school, between schools, or between a school and its LEA, its county, or the state.

For more details on the criteria for interpreting information provided on the score reports, refer to the Starting Smarter website. Refer also to *CAASPP Post-Test Guide* (CDE, 2021).

### References

California Department of Education. (2019). *Standard setting technical report for the California Spanish Assessment.* California Department of Education website.

California Department of Education. (2021). *CAASPP post-test guide: Technical information for student score reports for CAASPP LEA and test site coordinators and research specialists.* Sacramento, CA: California Department of Education.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21,* 365–89.

### Appendix 6.A: Raw Score, Theta Score, and Scale Score Distributions of Students Taking Each Test

**Notes:**

* A student with an incomplete test did not receive a score.
* When a student logged on to the TDS but did not answer any item, the student did not receive a score.
* When a student logged on and answered at least 1 item but fewer than 10 items, the student received the lowest obtainable scale score (such as 300 for grade three, 400 for grade four, …, and 900 for high school grades).
* Table 6.A.1 through table 6.A.6 include students who have completed testing with a valid theta score.
* Table 6.A.7 does not include theta scores because IRT equating was not applied to high school results in the 2020–2021 administration.
* Table 6.A.1 through table 6.A.7 only include students taking the regular forms.

Table 6.A.1 Overall Raw, Theta, and Scale Score Distribution Grade Three

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 3 | -3.224 | 303 | 1 | 0.1 | 1 | 0.1 |
| 8 | -2.046 | 319 | 1 | 0.1 | 2 | 0.2 |
| 9 | -1.893 | 321 | 2 | 0.2 | 4 | 0.3 |
| 10 | -1.754 | 323 | 1 | 0.1 | 5 | 0.4 |
| 11 | -1.625 | 325 | 5 | 0.4 | 10 | 0.8 |
| 12 | -1.505 | 327 | 6 | 0.5 | 16 | 1.3 |
| 13 | -1.392 | 329 | 21 | 1.8 | 37 | 3.1 |
| 14 | -1.286 | 330 | 29 | 2.4 | 66 | 5.5 |
| 15 | -1.184 | 331 | 33 | 2.8 | 99 | 8.3 |
| 16 | -1.087 | 333 | 51 | 4.3 | 150 | 12.5 |
| 17 | -0.994 | 334 | 61 | 5.1 | 211 | 17.6 |
| 18 | -0.904 | 335 | 77 | 6.4 | 288 | 24.1 |
| 19 | -0.816 | 337 | 75 | 6.3 | 363 | 30.4 |
| 20 | -0.732 | 338 | 66 | 5.5 | 429 | 35.9 |
| 21 | -0.649 | 339 | 72 | 6.0 | 501 | 41.9 |
| 22 | -0.569 | 340 | 59 | 4.9 | 560 | 46.8 |
| 23 | -0.490 | 341 | 63 | 5.3 | 623 | 52.1 |
| 24 | -0.412 | 342 | 47 | 3.9 | 670 | 56.0 |
| 25 | -0.336 | 343 | 45 | 3.8 | 715 | 59.8 |
| 26 | -0.260 | 344 | 38 | 3.2 | 753 | 63.0 |

Table 6.A.1 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 27 | -0.186 | 345 | 35 | 2.9 | 788 | 65.9 |
| 28 | -0.112 | 346 | 26 | 2.2 | 814 | 68.1 |
| 29 | -0.039 | 347 | 26 | 2.2 | 840 | 70.2 |
| 30 | 0.034 | 348 | 11 | 0.9 | 851 | 71.2 |
| 31 | 0.107 | 349 | 23 | 1.9 | 874 | 73.1 |
| 32 | 0.180 | 351 | 28 | 2.3 | 902 | 75.4 |
| 33 | 0.252 | 352 | 29 | 2.4 | 931 | 77.8 |
| 34 | 0.325 | 353 | 17 | 1.4 | 948 | 79.3 |
| 35 | 0.398 | 354 | 21 | 1.8 | 969 | 81.0 |
| 36 | 0.472 | 355 | 24 | 2.0 | 993 | 83.0 |
| 37 | 0.546 | 356 | 19 | 1.6 | 1,012 | 84.6 |
| 38 | 0.620 | 357 | 26 | 2.2 | 1,038 | 86.8 |
| 39 | 0.696 | 358 | 23 | 1.9 | 1,061 | 88.7 |
| 40 | 0.773 | 359 | 20 | 1.7 | 1,081 | 90.4 |
| 41 | 0.851 | 360 | 12 | 1.0 | 1,093 | 91.4 |
| 42 | 0.931 | 361 | 11 | 0.9 | 1,104 | 92.3 |
| 43 | 1.013 | 362 | 16 | 1.3 | 1,120 | 93.6 |
| 44 | 1.097 | 363 | 12 | 1.0 | 1,132 | 94.6 |
| 45 | 1.183 | 365 | 9 | 0.8 | 1,141 | 95.4 |
| 46 | 1.272 | 366 | 8 | 0.7 | 1,149 | 96.1 |
| 47 | 1.365 | 367 | 6 | 0.5 | 1,155 | 96.6 |
| 48 | 1.462 | 368 | 9 | 0.8 | 1,164 | 97.3 |
| 49 | 1.563 | 370 | 6 | 0.5 | 1,170 | 97.8 |
| 50 | 1.670 | 371 | 4 | 0.3 | 1,174 | 98.2 |
| 51 | 1.783 | 373 | 7 | 0.6 | 1,181 | 98.7 |
| 52 | 1.905 | 375 | 4 | 0.3 | 1,185 | 99.1 |
| 53 | 2.036 | 377 | 7 | 0.6 | 1,192 | 99.7 |
| 54 | 2.179 | 379 | 1 | 0.1 | 1,193 | 99.7 |
| 55 | 2.338 | 381 | 3 | 0.3 | 1,196 | 100.0 |

Table 6.A.2 Overall Raw, Theta, and Scale Score Distribution for Grade Four

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 2 | -4.108 | 400 | 1 | 0.1 | 1 | 0.1 |
| 6 | -2.785 | 412 | 1 | 0.1 | 2 | 0.2 |
| 7 | -2.580 | 414 | 2 | 0.2 | 4 | 0.4 |
| 10 | -2.077 | 421 | 3 | 0.3 | 7 | 0.7 |
| 11 | -1.935 | 423 | 3 | 0.3 | 10 | 1.0 |
| 12 | -1.802 | 425 | 3 | 0.3 | 13 | 1.3 |
| 13 | -1.677 | 426 | 1 | 0.1 | 14 | 1.3 |
| 14 | -1.558 | 428 | 13 | 1.3 | 27 | 2.6 |
| 15 | -1.445 | 429 | 17 | 1.6 | 44 | 4.2 |
| 16 | -1.337 | 431 | 10 | 1.0 | 54 | 5.2 |
| 17 | -1.233 | 432 | 26 | 2.5 | 80 | 7.7 |
| 18 | -1.133 | 433 | 29 | 2.8 | 109 | 10.5 |
| 19 | -1.036 | 435 | 41 | 3.9 | 150 | 14.4 |
| 20 | -0.942 | 436 | 54 | 5.2 | 204 | 19.6 |
| 21 | -0.851 | 437 | 52 | 5.0 | 256 | 24.6 |
| 22 | -0.762 | 438 | 60 | 5.8 | 316 | 30.4 |
| 23 | -0.675 | 439 | 63 | 6.1 | 379 | 36.4 |
| 24 | -0.589 | 440 | 54 | 5.2 | 433 | 41.6 |
| 25 | -0.506 | 441 | 49 | 4.7 | 482 | 46.3 |
| 26 | -0.423 | 443 | 60 | 5.8 | 542 | 52.1 |
| 27 | -0.342 | 444 | 42 | 4.0 | 584 | 56.2 |
| 28 | -0.262 | 445 | 35 | 3.4 | 619 | 59.5 |
| 29 | -0.183 | 446 | 30 | 2.9 | 649 | 62.4 |
| 30 | -0.104 | 447 | 26 | 2.5 | 675 | 64.9 |
| 31 | -0.026 | 448 | 27 | 2.6 | 702 | 67.5 |
| 32 | 0.052 | 449 | 18 | 1.7 | 720 | 69.2 |
| 33 | 0.129 | 450 | 19 | 1.8 | 739 | 71.1 |
| 34 | 0.207 | 451 | 32 | 3.1 | 771 | 74.1 |
| 35 | 0.284 | 452 | 27 | 2.6 | 798 | 76.7 |
| 36 | 0.362 | 453 | 22 | 2.1 | 820 | 78.8 |
| 37 | 0.440 | 454 | 21 | 2.0 | 841 | 80.9 |
| 38 | 0.518 | 455 | 27 | 2.6 | 868 | 83.5 |
| 39 | 0.598 | 456 | 24 | 2.3 | 892 | 85.8 |
| 40 | 0.678 | 457 | 14 | 1.3 | 906 | 87.1 |
| 41 | 0.759 | 458 | 15 | 1.4 | 921 | 88.6 |
| 42 | 0.842 | 459 | 14 | 1.3 | 935 | 89.9 |
| 43 | 0.926 | 460 | 8 | 0.8 | 943 | 90.7 |
| 44 | 1.012 | 461 | 21 | 2.0 | 964 | 92.7 |
| 45 | 1.100 | 462 | 11 | 1.1 | 975 | 93.8 |

Table 6.A.2 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 46 | 1.191 | 463 | 12 | 1.2 | 987 | 94.9 |
| 47 | 1.285 | 465 | 13 | 1.3 | 1,000 | 96.2 |
| 48 | 1.382 | 466 | 9 | 0.9 | 1,009 | 97.0 |
| 49 | 1.483 | 467 | 3 | 0.3 | 1,012 | 97.3 |
| 50 | 1.588 | 469 | 12 | 1.2 | 1,024 | 98.5 |
| 51 | 1.699 | 470 | 7 | 0.7 | 1,031 | 99.1 |
| 52 | 1.817 | 472 | 3 | 0.3 | 1,034 | 99.4 |
| 53 | 1.942 | 473 | 4 | 0.4 | 1,038 | 99.8 |
| 54 | 2.078 | 475 | 1 | 0.1 | 1,039 | 99.9 |
| 56 | 2.387 | 479 | 1 | 0.1 | 1,040 | 100.0 |

Table 6.A.3 Overall Raw, Theta, and Scale Score Distribution for Grade Five

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 4 | -3.008 | 506 | 1 | 0.1 | 1 | 0.1 |
| 8 | -2.135 | 518 | 2 | 0.2 | 3 | 0.3 |
| 10 | -1.834 | 522 | 1 | 0.1 | 4 | 0.4 |
| 11 | -1.701 | 524 | 4 | 0.4 | 8 | 0.9 |
| 12 | -1.576 | 526 | 5 | 0.6 | 13 | 1.4 |
| 13 | -1.459 | 528 | 11 | 1.2 | 24 | 2.7 |
| 14 | -1.349 | 529 | 11 | 1.2 | 35 | 3.9 |
| 15 | -1.243 | 531 | 19 | 2.1 | 54 | 6.0 |
| 16 | -1.142 | 532 | 20 | 2.2 | 74 | 8.2 |
| 17 | -1.046 | 533 | 32 | 3.5 | 106 | 11.8 |
| 18 | -0.952 | 535 | 29 | 3.2 | 135 | 15.0 |
| 19 | -0.861 | 536 | 47 | 5.2 | 182 | 20.2 |
| 20 | -0.773 | 537 | 55 | 6.1 | 237 | 26.3 |
| 21 | -0.688 | 538 | 45 | 5.0 | 282 | 31.3 |
| 22 | -0.604 | 540 | 37 | 4.1 | 319 | 35.4 |
| 23 | -0.521 | 541 | 39 | 4.3 | 358 | 39.7 |
| 24 | -0.441 | 542 | 40 | 4.4 | 398 | 44.1 |
| 25 | -0.361 | 543 | 40 | 4.4 | 438 | 48.6 |
| 26 | -0.283 | 544 | 41 | 4.5 | 479 | 53.1 |
| 27 | -0.205 | 545 | 37 | 4.1 | 516 | 57.2 |
| 28 | -0.128 | 546 | 25 | 2.8 | 541 | 60.0 |
| 29 | -0.052 | 547 | 29 | 3.2 | 570 | 63.2 |
| 30 | 0.024 | 548 | 36 | 4.0 | 606 | 67.2 |
| 31 | 0.100 | 549 | 27 | 3.0 | 633 | 70.2 |
| 32 | 0.176 | 550 | 23 | 2.5 | 656 | 72.7 |
| 33 | 0.252 | 552 | 26 | 2.9 | 682 | 75.6 |
| 34 | 0.328 | 553 | 21 | 2.3 | 703 | 77.9 |
| 35 | 0.404 | 554 | 26 | 2.9 | 729 | 80.8 |
| 36 | 0.481 | 555 | 22 | 2.4 | 751 | 83.3 |
| 37 | 0.558 | 556 | 11 | 1.2 | 762 | 84.5 |
| 38 | 0.637 | 557 | 16 | 1.8 | 778 | 86.3 |
| 39 | 0.716 | 558 | 18 | 2.0 | 796 | 88.2 |
| 40 | 0.797 | 559 | 16 | 1.8 | 812 | 90.0 |
| 41 | 0.879 | 560 | 16 | 1.8 | 828 | 91.8 |
| 42 | 0.962 | 561 | 12 | 1.3 | 840 | 93.1 |
| 43 | 1.048 | 563 | 16 | 1.8 | 856 | 94.9 |
| 44 | 1.136 | 564 | 8 | 0.9 | 864 | 95.8 |
| 45 | 1.226 | 565 | 8 | 0.9 | 872 | 96.7 |
| 46 | 1.319 | 566 | 5 | 0.6 | 877 | 97.2 |

Table 6.A.3 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 47 | 1.416 | 568 | 7 | 0.8 | 884 | 98.0 |
| 48 | 1.517 | 569 | 3 | 0.3 | 887 | 98.3 |
| 49 | 1.623 | 571 | 5 | 0.6 | 892 | 98.9 |
| 50 | 1.734 | 572 | 4 | 0.4 | 896 | 99.3 |
| 51 | 1.851 | 574 | 3 | 0.3 | 899 | 99.7 |
| 52 | 1.977 | 576 | 3 | 0.3 | 902 | 100.0 |

Table 6.A.4 Overall Raw, Theta, and Scale Score Distribution for Grade Six

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 6 | -2.769 | 611 | 1 | 0.1 | 1 | 0.1 |
| 7 | -2.579 | 614 | 2 | 0.2 | 3 | 0.3 |
| 9 | -2.258 | 618 | 1 | 0.1 | 4 | 0.4 |
| 11 | -1.990 | 622 | 1 | 0.1 | 5 | 0.5 |
| 12 | -1.869 | 624 | 2 | 0.2 | 7 | 0.7 |
| 13 | -1.756 | 625 | 3 | 0.3 | 10 | 1.1 |
| 15 | -1.547 | 628 | 3 | 0.3 | 13 | 1.4 |
| 16 | -1.450 | 630 | 5 | 0.5 | 18 | 1.9 |
| 17 | -1.356 | 631 | 12 | 1.3 | 30 | 3.2 |
| 18 | -1.265 | 632 | 10 | 1.1 | 40 | 4.3 |
| 19 | -1.178 | 634 | 14 | 1.5 | 54 | 5.8 |
| 20 | -1.093 | 635 | 35 | 3.7 | 89 | 9.5 |
| 21 | -1.010 | 636 | 27 | 2.9 | 116 | 12.4 |
| 22 | -0.929 | 637 | 24 | 2.6 | 140 | 14.9 |
| 23 | -0.849 | 638 | 28 | 3.0 | 168 | 17.9 |
| 24 | -0.772 | 639 | 26 | 2.8 | 194 | 20.7 |
| 25 | -0.695 | 640 | 38 | 4.1 | 232 | 24.8 |
| 26 | -0.620 | 641 | 35 | 3.7 | 267 | 28.5 |
| 27 | -0.545 | 642 | 34 | 3.6 | 301 | 32.1 |
| 28 | -0.472 | 643 | 31 | 3.3 | 332 | 35.4 |
| 29 | -0.399 | 644 | 33 | 3.5 | 365 | 39.0 |
| 30 | -0.326 | 645 | 27 | 2.9 | 392 | 41.8 |
| 31 | -0.254 | 646 | 24 | 2.6 | 416 | 44.4 |
| 32 | -0.182 | 647 | 32 | 3.4 | 448 | 47.8 |
| 33 | -0.111 | 648 | 26 | 2.8 | 474 | 50.6 |
| 34 | -0.039 | 649 | 29 | 3.1 | 503 | 53.7 |
| 35 | 0.032 | 650 | 19 | 2.0 | 522 | 55.7 |
| 36 | 0.104 | 651 | 32 | 3.4 | 554 | 59.1 |
| 37 | 0.177 | 652 | 37 | 3.9 | 591 | 63.1 |
| 38 | 0.249 | 653 | 24 | 2.6 | 615 | 65.6 |
| 39 | 0.323 | 655 | 24 | 2.6 | 639 | 68.2 |
| 40 | 0.397 | 656 | 23 | 2.5 | 662 | 70.7 |
| 41 | 0.472 | 657 | 24 | 2.6 | 686 | 73.2 |
| 42 | 0.548 | 658 | 22 | 2.3 | 708 | 75.6 |
| 43 | 0.626 | 659 | 25 | 2.7 | 733 | 78.2 |
| 44 | 0.705 | 660 | 25 | 2.7 | 758 | 80.9 |
| 45 | 0.785 | 661 | 30 | 3.2 | 788 | 84.1 |
| 46 | 0.868 | 662 | 21 | 2.2 | 809 | 86.3 |
| 47 | 0.953 | 663 | 21 | 2.2 | 830 | 88.6 |

Table 6.A.4 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 48 | 1.041 | 665 | 18 | 1.9 | 848 | 90.5 |
| 49 | 1.132 | 666 | 17 | 1.8 | 865 | 92.3 |
| 50 | 1.226 | 667 | 11 | 1.2 | 876 | 93.5 |
| 51 | 1.324 | 669 | 11 | 1.2 | 887 | 94.7 |
| 52 | 1.427 | 670 | 15 | 1.6 | 902 | 96.3 |
| 53 | 1.536 | 672 | 7 | 0.7 | 909 | 97.0 |
| 54 | 1.651 | 673 | 7 | 0.7 | 916 | 97.8 |
| 55 | 1.774 | 675 | 4 | 0.4 | 920 | 98.2 |
| 56 | 1.908 | 677 | 8 | 0.9 | 928 | 99.0 |
| 57 | 2.053 | 679 | 5 | 0.5 | 933 | 99.6 |
| 58 | 2.214 | 681 | 4 | 0.4 | 937 | 100.0 |

Table 6.A.5 Overall Raw, Theta, and Scale Score Distribution for Grade Seven

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 4 | -3.022 | 700 | 1 | 0.1 | 1 | 0.1 |
| 5 | -2.754 | 701 | 4 | 0.6 | 5 | 0.7 |
| 6 | -2.528 | 704 | 3 | 0.4 | 8 | 1.1 |
| 7 | -2.332 | 707 | 2 | 0.3 | 10 | 1.4 |
| 8 | -2.159 | 710 | 2 | 0.3 | 12 | 1.7 |
| 9 | -2.002 | 712 | 1 | 0.1 | 13 | 1.8 |
| 11 | -1.727 | 716 | 4 | 0.6 | 17 | 2.4 |
| 13 | -1.488 | 720 | 4 | 0.6 | 21 | 2.9 |
| 14 | -1.379 | 721 | 2 | 0.3 | 23 | 3.2 |
| 15 | -1.275 | 723 | 6 | 0.8 | 29 | 4.0 |
| 16 | -1.177 | 724 | 8 | 1.1 | 37 | 5.2 |
| 17 | -1.082 | 726 | 6 | 0.8 | 43 | 6.0 |
| 18 | -0.991 | 727 | 7 | 1.0 | 50 | 7.0 |
| 19 | -0.903 | 728 | 13 | 1.8 | 63 | 8.8 |
| 20 | -0.818 | 730 | 24 | 3.3 | 87 | 12.1 |
| 21 | -0.735 | 731 | 19 | 2.6 | 106 | 14.8 |
| 22 | -0.654 | 732 | 22 | 3.1 | 128 | 17.9 |
| 23 | -0.575 | 733 | 22 | 3.1 | 150 | 20.9 |
| 24 | -0.498 | 735 | 25 | 3.5 | 175 | 24.4 |
| 25 | -0.422 | 736 | 24 | 3.3 | 199 | 27.8 |
| 26 | -0.347 | 737 | 23 | 3.2 | 222 | 31.0 |
| 27 | -0.274 | 738 | 29 | 4.0 | 251 | 35.0 |
| 28 | -0.201 | 739 | 24 | 3.3 | 275 | 38.4 |
| 29 | -0.129 | 740 | 28 | 3.9 | 303 | 42.3 |
| 30 | -0.057 | 741 | 31 | 4.3 | 334 | 46.6 |
| 31 | 0.014 | 742 | 35 | 4.9 | 369 | 51.5 |
| 32 | 0.085 | 743 | 25 | 3.5 | 394 | 55.0 |
| 33 | 0.156 | 744 | 31 | 4.3 | 425 | 59.3 |
| 34 | 0.227 | 745 | 26 | 3.6 | 451 | 62.9 |
| 35 | 0.298 | 746 | 25 | 3.5 | 476 | 66.4 |
| 36 | 0.369 | 748 | 16 | 2.2 | 492 | 68.6 |
| 37 | 0.440 | 749 | 28 | 3.9 | 520 | 72.5 |
| 38 | 0.512 | 750 | 17 | 2.4 | 537 | 74.9 |
| 39 | 0.585 | 751 | 24 | 3.3 | 561 | 78.2 |
| 40 | 0.659 | 752 | 21 | 2.9 | 582 | 81.2 |
| 41 | 0.733 | 753 | 19 | 2.6 | 601 | 83.8 |
| 42 | 0.809 | 754 | 16 | 2.2 | 617 | 86.1 |
| 43 | 0.887 | 755 | 14 | 2.0 | 631 | 88.0 |
| 44 | 0.965 | 756 | 21 | 2.9 | 652 | 90.9 |

Table 6.A.5 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 45 | 1.046 | 758 | 11 | 1.5 | 663 | 92.5 |
| 46 | 1.129 | 759 | 16 | 2.2 | 679 | 94.7 |
| 47 | 1.214 | 760 | 8 | 1.1 | 687 | 95.8 |
| 48 | 1.303 | 762 | 8 | 1.1 | 695 | 96.9 |
| 49 | 1.394 | 763 | 6 | 0.8 | 701 | 97.8 |
| 50 | 1.489 | 764 | 5 | 0.7 | 706 | 98.5 |
| 51 | 1.589 | 766 | 2 | 0.3 | 708 | 98.7 |
| 52 | 1.694 | 767 | 4 | 0.6 | 712 | 99.3 |
| 53 | 1.805 | 769 | 2 | 0.3 | 714 | 99.6 |
| 54 | 1.924 | 771 | 1 | 0.1 | 715 | 99.7 |
| 55 | 2.051 | 773 | 1 | 0.1 | 716 | 99.9 |
| 56 | 2.189 | 775 | 1 | 0.1 | 717 | 100.0 |

Table 6.A.6 Overall Raw, Theta, and Scale Score Distribution for Grade Eight

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 8 | -2.511 | 808 | 2 | 0.4 | 2 | 0.4 |
| 11 | -2.053 | 815 | 1 | 0.2 | 3 | 0.5 |
| 12 | -1.922 | 817 | 3 | 0.5 | 6 | 1.1 |
| 13 | -1.800 | 819 | 1 | 0.2 | 7 | 1.3 |
| 14 | -1.684 | 821 | 2 | 0.4 | 9 | 1.6 |
| 15 | -1.574 | 822 | 4 | 0.7 | 13 | 2.3 |
| 16 | -1.469 | 824 | 1 | 0.2 | 14 | 2.5 |
| 17 | -1.369 | 825 | 10 | 1.8 | 24 | 4.3 |
| 18 | -1.272 | 827 | 7 | 1.3 | 31 | 5.5 |
| 19 | -1.179 | 828 | 8 | 1.4 | 39 | 7.0 |
| 20 | -1.089 | 830 | 4 | 0.7 | 43 | 7.7 |
| 21 | -1.001 | 831 | 9 | 1.6 | 52 | 9.3 |
| 22 | -0.916 | 832 | 11 | 2.0 | 63 | 11.3 |
| 23 | -0.833 | 834 | 18 | 3.2 | 81 | 14.5 |
| 24 | -0.751 | 835 | 19 | 3.4 | 100 | 17.9 |
| 25 | -0.671 | 836 | 23 | 4.1 | 123 | 22.0 |
| 26 | -0.593 | 837 | 11 | 2.0 | 134 | 23.9 |
| 27 | -0.516 | 838 | 17 | 3.0 | 151 | 27.0 |
| 28 | -0.439 | 839 | 23 | 4.1 | 174 | 31.1 |
| 29 | -0.364 | 841 | 17 | 3.0 | 191 | 34.1 |
| 30 | -0.289 | 842 | 28 | 5.0 | 219 | 39.1 |
| 31 | -0.215 | 843 | 21 | 3.8 | 240 | 42.9 |
| 32 | -0.141 | 844 | 27 | 4.8 | 267 | 47.7 |
| 33 | -0.068 | 845 | 21 | 3.8 | 288 | 51.4 |
| 34 | 0.006 | 846 | 13 | 2.3 | 301 | 53.8 |
| 35 | 0.079 | 847 | 20 | 3.6 | 321 | 57.3 |
| 36 | 0.152 | 848 | 24 | 4.3 | 345 | 61.6 |
| 37 | 0.226 | 849 | 21 | 3.8 | 366 | 65.4 |
| 38 | 0.300 | 851 | 19 | 3.4 | 385 | 68.8 |
| 39 | 0.375 | 852 | 24 | 4.3 | 409 | 73.0 |
| 40 | 0.450 | 853 | 19 | 3.4 | 428 | 76.4 |
| 41 | 0.527 | 854 | 23 | 4.1 | 451 | 80.5 |
| 42 | 0.604 | 855 | 17 | 3.0 | 468 | 83.6 |
| 43 | 0.683 | 856 | 9 | 1.6 | 477 | 85.2 |
| 44 | 0.763 | 857 | 6 | 1.1 | 483 | 86.3 |
| 45 | 0.844 | 859 | 13 | 2.3 | 496 | 88.6 |
| 46 | 0.928 | 860 | 10 | 1.8 | 506 | 90.4 |
| 47 | 1.014 | 861 | 8 | 1.4 | 514 | 91.8 |
| 48 | 1.102 | 863 | 8 | 1.4 | 522 | 93.2 |

Table 6.A.6 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 49 | 1.194 | 864 | 7 | 1.3 | 529 | 94.5 |
| 50 | 1.288 | 865 | 8 | 1.4 | 537 | 95.9 |
| 51 | 1.387 | 867 | 7 | 1.3 | 544 | 97.1 |
| 52 | 1.491 | 868 | 4 | 0.7 | 548 | 97.9 |
| 53 | 1.600 | 870 | 4 | 0.7 | 552 | 98.6 |
| 54 | 1.716 | 872 | 1 | 0.2 | 553 | 98.8 |
| 55 | 1.840 | 874 | 3 | 0.5 | 556 | 99.3 |
| 56 | 1.973 | 876 | 2 | 0.4 | 558 | 99.6 |
| 57 | 2.119 | 878 | 1 | 0.2 | 559 | 99.8 |
| 58 | 2.280 | 880 | 1 | 0.2 | 560 | 100.0 |

Table 6.A.7 Overall Raw, Theta, and Scale Score Distribution for High School

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 4 | N/A | 900 | 1 | 0.4 | 1 | 0.4 |
| 5 | N/A | 900 | 2 | 0.8 | 3 | 1.2 |
| 8 | N/A | 905 | 2 | 0.8 | 5 | 2.0 |
| 9 | N/A | 909 | 1 | 0.4 | 6 | 2.4 |
| 10 | N/A | 912 | 2 | 0.8 | 8 | 3.2 |
| 11 | N/A | 915 | 1 | 0.4 | 9 | 3.6 |
| 12 | N/A | 917 | 2 | 0.8 | 11 | 4.3 |
| 13 | N/A | 920 | 4 | 1.6 | 15 | 5.9 |
| 14 | N/A | 922 | 2 | 0.8 | 17 | 6.7 |
| 15 | N/A | 924 | 2 | 0.8 | 19 | 7.5 |
| 16 | N/A | 925 | 12 | 4.7 | 31 | 12.3 |
| 17 | N/A | 927 | 3 | 1.2 | 34 | 13.4 |
| 18 | N/A | 929 | 8 | 3.2 | 42 | 16.6 |
| 19 | N/A | 930 | 5 | 2.0 | 47 | 18.6 |
| 20 | N/A | 931 | 14 | 5.5 | 61 | 24.1 |
| 21 | N/A | 933 | 4 | 1.6 | 65 | 25.7 |
| 22 | N/A | 934 | 7 | 2.8 | 72 | 28.5 |
| 23 | N/A | 935 | 7 | 2.8 | 79 | 31.2 |
| 24 | N/A | 937 | 6 | 2.4 | 85 | 33.6 |
| 25 | N/A | 938 | 8 | 3.2 | 93 | 36.8 |
| 26 | N/A | 939 | 10 | 4.0 | 103 | 40.7 |
| 27 | N/A | 940 | 5 | 2.0 | 108 | 42.7 |
| 28 | N/A | 941 | 7 | 2.8 | 115 | 45.5 |
| 29 | N/A | 942 | 8 | 3.2 | 123 | 48.6 |
| 30 | N/A | 943 | 13 | 5.1 | 136 | 53.8 |
| 31 | N/A | 944 | 9 | 3.6 | 145 | 57.3 |
| 32 | N/A | 945 | 6 | 2.4 | 151 | 59.7 |
| 33 | N/A | 946 | 10 | 4.0 | 161 | 63.6 |
| 34 | N/A | 947 | 7 | 2.8 | 168 | 66.4 |
| 35 | N/A | 948 | 4 | 1.6 | 172 | 68.0 |
| 36 | N/A | 949 | 6 | 2.4 | 178 | 70.4 |
| 37 | N/A | 951 | 7 | 2.8 | 185 | 73.1 |
| 38 | N/A | 952 | 6 | 2.4 | 191 | 75.5 |
| 39 | N/A | 953 | 9 | 3.6 | 200 | 79.1 |
| 40 | N/A | 954 | 8 | 3.2 | 208 | 82.2 |
| 41 | N/A | 955 | 7 | 2.8 | 215 | 85.0 |
| 42 | N/A | 956 | 5 | 2.0 | 220 | 87.0 |
| 43 | N/A | 957 | 7 | 2.8 | 227 | 89.7 |
| 44 | N/A | 958 | 3 | 1.2 | 230 | 90.9 |

Table 6.A.7 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Raw Score | Theta Score | Scale Score | N | Percent | Cumulative Frequency | Cumulative Percent |
| 45 | N/A | 959 | 8 | 3.2 | 238 | 94.1 |
| 46 | N/A | 960 | 3 | 1.2 | 241 | 95.3 |
| 47 | N/A | 961 | 5 | 2.0 | 246 | 97.2 |
| 48 | N/A | 962 | 1 | 0.4 | 247 | 97.6 |
| 49 | N/A | 964 | 3 | 1.2 | 250 | 98.8 |
| 50 | N/A | 965 | 1 | 0.4 | 251 | 99.2 |
| 52 | N/A | 967 | 1 | 0.4 | 252 | 99.6 |
| 54 | N/A | 970 | 1 | 0.4 | 253 | 100.0 |

### Appendix 6.B: Demographic Summary of Students in Each Reporting Score Range

**Notes:**

* To protect privacy when the number of students in a student group is 10 or fewer, the summary statistics at the test and reporting levels are not reported and are presented as “N/A” in the tables in [appendix 6.B](#_Appendix_6.B:_Raw).
* Percentages in these tables may not sum up to 100 because of rounding.
* Table 6.B.1 through table 6.B.7, include students who have completed testing with a valid reporting scale score.
* Data for the category of “Received in Spanish in the 2020–2021 school year—Spanish as a Foreign Language Program” are available for grades six through eight and for high school.
* SD = standard deviation.

Table 6.B.1 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Three

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 1,213 | 344 | 10.8 | 71.5 | 19.0 | 9.5 |
| Male | 581 | 343 | 10.4 | 74.2 | 18.4 | 7.4 |
| Female | 632 | 345 | 11.0 | 69.0 | 19.6 | 11.4 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 9 | N/A | N/A | N/A | N/A | N/A |
| Asian | 19 | 349 | 10.5 | 52.6 | 31.6 | 15.8 |
| Native Hawaiian or Other Pacific Islander | 0 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 6 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 1,002 | 343 | 10.4 | 73.1 | 18.6 | 8.4 |
| Black or African American | 12 | 345 | 14.9 | 75.0 | 8.3 | 16.7 |
| White | 133 | 347 | 11.5 | 65.4 | 21.1 | 13.5 |
| Two or more races | 32 | 349 | 13.7 | 53.1 | 25.0 | 21.9 |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |

Table 6.B.1 *(continuation one)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| English only | 388 | 344 | 11.1 | 70.6 | 18.6 | 10.8 |
| Initial fluent English proficient (IFEP) | 57 | 349 | 12.3 | 49.1 | 40.4 | 10.5 |
| EL | 622 | 342 | 9.7 | 78.3 | 15.0 | 6.8 |
| Reclassified fluent English proficient (RFEP) | 144 | 348 | 11.7 | 52.8 | 29.9 | 17.4 |
| Ever-ELs (EL or RFEP) | 766 | 343 | 10.4 | 73.5 | 17.8 | 8.7 |
| To be determined | 2 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 1,116 | 344 | 10.9 | 70.3 | 19.7 | 10.0 |
| Special education services | 97 | 341 | 8.9 | 85.6 | 11.3 | 3.1 |
| Economically disadvantaged | 782 | 342 | 9.3 | 80.2 | 14.7 | 5.1 |
| Not economically disadvantaged | 431 | 348 | 12.0 | 55.7 | 26.9 | 17.4 |
| In U.S. schools less than 12 months | 9 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 1,204 | 344 | 10.8 | 71.4 | 19.1 | 9.5 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 1,136 | 344 | 10.9 | 70.7 | 19.4 | 9.9 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 71 | 338 | 6.3 | 93.0 | 7.0 | 0.0 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 906 | 345 | 11.0 | 67.4 | 21.3 | 11.3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 88 | 342 | 11.2 | 77.3 | 13.6 | 9.1 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 18 | 337 | 10.3 | 88.9 | 11.1 | 0.0 |

Table 6.B.1 *(continuation two)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 38 | 337 | 5.6 | 94.7 | 5.3 | 0.0 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 303 | 343 | 11.3 | 72.9 | 17.8 | 9.2 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 476 | 345 | 11.2 | 67.2 | 21.0 | 11.8 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 319 | 344 | 10.2 | 70.8 | 20.1 | 9.1 |

Table 6.B.2 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Four

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 1,052 | 445 | 10.3 | 67.8 | 22.2 | 10.0 |
| Male | 511 | 444 | 10.1 | 70.8 | 20.2 | 9.0 |
| Female | 541 | 445 | 10.4 | 64.9 | 24.2 | 10.9 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 4 | N/A | N/A | N/A | N/A | N/A |
| Asian | 13 | 447 | 10.5 | 46.2 | 46.2 | 7.7 |
| Native Hawaiian or Other Pacific Islander | 1 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 5 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 885 | 445 | 10.3 | 68.0 | 22.0 | 9.9 |
| Black or African American | 11 | 440 | 7.7 | 90.9 | 9.1 | 0.0 |
| White | 106 | 445 | 10.6 | 67.9 | 21.7 | 10.4 |
| Two or more races | 27 | 447 | 11.2 | 55.6 | 25.9 | 18.5 |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 326 | 443 | 9.8 | 74.8 | 18.7 | 6.4 |
| IFEP | 46 | 450 | 11.4 | 52.2 | 21.7 | 26.1 |
| EL | 463 | 443 | 9.6 | 72.1 | 21.2 | 6.7 |
| RFEP | 216 | 450 | 10.6 | 50.9 | 30.1 | 19.0 |
| Ever-ELs (EL or RFEP) | 679 | 445 | 10.3 | 65.4 | 24.0 | 10.6 |
| To be determined | 0 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 1 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 963 | 445 | 10.3 | 66.0 | 23.5 | 10.5 |
| Special education services | 89 | 441 | 9.1 | 86.5 | 9.0 | 4.5 |
| Economically disadvantaged | 684 | 444 | 10.0 | 70.8 | 20.3 | 8.9 |
| Not economically disadvantaged | 368 | 446 | 10.8 | 62.2 | 25.8 | 12.0 |
| In U.S. schools less than 12 months | 7 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 1,045 | 445 | 10.3 | 68.0 | 22.1 | 9.9 |

Table 6.B.2 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 1,001 | 445 | 10.3 | 66.5 | 23.1 | 10.4 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 45 | 440 | 8.8 | 84.4 | 13.3 | 2.2 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 789 | 446 | 10.5 | 63.6 | 24.5 | 11.9 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 73 | 442 | 8.9 | 79.5 | 16.4 | 4.1 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 18 | 441 | 7.6 | 83.3 | 11.1 | 5.6 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 29 | 443 | 9.2 | 79.3 | 10.3 | 10.3 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 492 | 445 | 10.8 | 66.3 | 21.3 | 12.4 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 342 | 445 | 10.5 | 64.6 | 24.9 | 10.5 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 138 | 444 | 8.2 | 69.6 | 27.5 | 2.9 |

Table 6.B.3 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Five

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 916 | 545 | 10.2 | 57.5 | 32.6 | 9.8 |
| Male | 448 | 544 | 9.7 | 62.5 | 29.9 | 7.6 |
| Female | 468 | 546 | 10.6 | 52.8 | 35.3 | 12.0 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 9 | N/A | N/A | N/A | N/A | N/A |
| Asian | 10 | N/A | N/A | N/A | N/A | N/A |
| Native Hawaiian or Other Pacific Islander | 0 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 0 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 769 | 545 | 10.1 | 57.2 | 33.3 | 9.5 |
| Black or African American | 6 | N/A | N/A | N/A | N/A | N/A |
| White | 98 | 545 | 10.7 | 59.2 | 29.6 | 11.2 |
| Two or more races | 24 | 546 | 11.5 | 62.5 | 29.2 | 8.3 |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 300 | 544 | 10.3 | 62.7 | 27.7 | 9.7 |
| IFEP | 42 | 551 | 12.2 | 33.3 | 40.5 | 26.2 |
| EL | 352 | 542 | 8.5 | 71.6 | 25.9 | 2.6 |
| RFEP | 222 | 551 | 9.6 | 32.9 | 48.6 | 18.5 |
| Ever-ELs (EL or RFEP) | 574 | 545 | 9.9 | 56.6 | 34.7 | 8.7 |
| To be determined | 0 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 848 | 546 | 10.2 | 55.8 | 33.7 | 10.5 |
| Special education services | 68 | 539 | 8.0 | 79.4 | 19.1 | 1.5 |
| Economically disadvantaged | 600 | 544 | 10.1 | 59.8 | 31.3 | 8.8 |
| Not economically disadvantaged | 316 | 546 | 10.4 | 53.2 | 35.1 | 11.7 |
| In U.S. schools less than 12 months | 2 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 914 | 545 | 10.2 | 57.7 | 32.6 | 9.7 |

Table 6.B.3 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 887 | 545 | 10.2 | 57.0 | 32.8 | 10.1 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 17 | 540 | 12.8 | 76.5 | 17.6 | 5.9 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 705 | 546 | 10.3 | 54.0 | 34.6 | 11.3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 91 | 544 | 9.3 | 63.7 | 29.7 | 6.6 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 14 | 539 | 7.7 | 85.7 | 14.3 | 0.0 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 31 | 538 | 9.7 | 80.6 | 16.1 | 3.2 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 475 | 546 | 10.2 | 53.5 | 36.0 | 10.5 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 274 | 545 | 10.3 | 58.8 | 30.3 | 10.9 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 107 | 544 | 9.7 | 61.7 | 29.9 | 8.4 |

Table 6.B.4 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Six

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 948 | 649 | 11.6 | 48.3 | 30.2 | 21.5 |
| Male | 471 | 647 | 10.9 | 56.1 | 27.8 | 16.1 |
| Female | 477 | 651 | 12.0 | 40.7 | 32.5 | 26.8 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 3 | N/A | N/A | N/A | N/A | N/A |
| Asian | 11 | 660 | 11.9 | 27.3 | 9.1 | 63.6 |
| Native Hawaiian or Other Pacific Islander | 0 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 1 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 806 | 649 | 11.2 | 49.4 | 31.9 | 18.7 |
| Black or African American | 9 | N/A | N/A | N/A | N/A | N/A |
| White | 92 | 652 | 12.7 | 45.7 | 21.7 | 32.6 |
| Two or more races | 26 | 655 | 15.7 | 34.6 | 19.2 | 46.2 |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 282 | 650 | 12.2 | 49.3 | 25.5 | 25.2 |
| IFEP | 38 | 655 | 12.9 | 26.3 | 31.6 | 42.1 |
| EL | 296 | 644 | 9.2 | 68.6 | 25.0 | 6.4 |
| RFEP | 331 | 653 | 11.0 | 32.0 | 38.7 | 29.3 |
| Ever-ELs (EL or RFEP) | 627 | 649 | 11.1 | 49.3 | 32.2 | 18.5 |
| To be determined | 1 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 889 | 650 | 11.6 | 46.2 | 31.0 | 22.7 |
| Special education services | 59 | 642 | 7.6 | 79.7 | 16.9 | 3.4 |
| Economically disadvantaged | 584 | 648 | 11.1 | 51.5 | 32.0 | 16.4 |
| Not economically disadvantaged | 364 | 652 | 11.9 | 43.1 | 27.2 | 29.7 |
| In U.S. schools less than 12 months | 6 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 942 | 649 | 11.6 | 48.4 | 30.3 | 21.3 |

Table 6.B.4 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 903 | 650 | 11.5 | 47.5 | 30.8 | 21.7 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 7 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 770 | 650 | 11.6 | 45.2 | 31.4 | 23.4 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 61 | 648 | 10.6 | 47.5 | 37.7 | 14.8 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 5 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Spanish as a Foreign Language Program | 15 | 640 | 3.9 | 100.0 | 0.0 | 0.0 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 89 | 650 | 12.0 | 49.4 | 21.3 | 29.2 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 497 | 651 | 11.6 | 41.4 | 33.8 | 24.7 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 213 | 648 | 11.5 | 54.9 | 27.2 | 17.8 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 104 | 646 | 9.2 | 59.6 | 31.7 | 8.7 |

Table 6.B.5 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Seven

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 729 | 742 | 11.3 | 55.7 | 39.1 | 5.2 |
| Male | 349 | 741 | 11.5 | 61.9 | 34.1 | 4.0 |
| Female | 380 | 744 | 11.0 | 50.0 | 43.7 | 6.3 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 0 | N/A | N/A | N/A | N/A | N/A |
| Asian | 10 | N/A | N/A | N/A | N/A | N/A |
| Native Hawaiian or Other Pacific Islander | 3 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 3 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 603 | 742 | 11.5 | 58.5 | 36.2 | 5.3 |
| Black or African American | 6 | N/A | N/A | N/A | N/A | N/A |
| White | 85 | 745 | 10.4 | 41.2 | 55.3 | 3.5 |
| Two or more races | 19 | 745 | 10.3 | 36.8 | 52.6 | 10.5 |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 234 | 743 | 11.4 | 51.3 | 43.6 | 5.1 |
| IFEP | 25 | 749 | 10.4 | 40.0 | 44.0 | 16.0 |
| EL | 161 | 736 | 9.0 | 83.9 | 15.5 | 0.6 |
| RFEP | 309 | 744 | 11.3 | 45.6 | 47.6 | 6.8 |
| Ever-ELs (EL or RFEP) | 470 | 742 | 11.2 | 58.7 | 36.6 | 4.7 |
| To be determined | 0 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 697 | 743 | 11.2 | 54.1 | 40.5 | 5.5 |
| Special education services | 32 | 733 | 9.8 | 90.6 | 9.4 | 0.0 |
| Economically disadvantaged | 409 | 741 | 10.9 | 61.6 | 34.0 | 4.4 |
| Not economically disadvantaged | 320 | 744 | 11.7 | 48.1 | 45.6 | 6.3 |
| In U.S. schools less than 12 months | 1 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 728 | 742 | 11.3 | 55.6 | 39.1 | 5.2 |

Table 6.B.5 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 717 | 742 | 11.3 | 55.6 | 39.1 | 5.3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 18 | 734 | 9.8 | 83.3 | 16.7 | 0.0 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 604 | 743 | 11.3 | 51.5 | 42.2 | 6.3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 54 | 739 | 9.3 | 75.9 | 24.1 | 0.0 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 4 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Spanish as a Foreign Language Program | 11 | 734 | 12.9 | 81.8 | 18.2 | 0.0 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 109 | 744 | 13.3 | 48.6 | 44.0 | 7.3 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 328 | 743 | 10.7 | 50.9 | 44.2 | 4.9 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 193 | 743 | 10.1 | 54.9 | 39.4 | 5.7 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 87 | 735 | 11.7 | 83.9 | 12.6 | 3.4 |

Table 6.B.6 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for Grade Eight

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 572 | 845 | 11.3 | 58.0 | 30.8 | 11.2 |
| Male | 286 | 844 | 10.9 | 66.1 | 27.3 | 6.6 |
| Female | 286 | 847 | 11.5 | 50.0 | 34.3 | 15.7 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 1 | N/A | N/A | N/A | N/A | N/A |
| Asian | 5 | N/A | N/A | N/A | N/A | N/A |
| Native Hawaiian or Other Pacific Islander | 1 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 2 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 495 | 845 | 11.3 | 59.4 | 31.1 | 9.5 |
| Black or African American | 3 | N/A | N/A | N/A | N/A | N/A |
| White | 59 | 850 | 11.5 | 45.8 | 28.8 | 25.4 |
| Two or more races | 6 | N/A | N/A | N/A | N/A | N/A |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 133 | 846 | 11.1 | 58.6 | 26.3 | 15.0 |
| IFEP | 30 | 852 | 9.1 | 33.3 | 46.7 | 20.0 |
| EL | 107 | 837 | 9.8 | 84.1 | 15.0 | 0.9 |
| RFEP | 302 | 847 | 10.7 | 51.0 | 36.8 | 12.3 |
| Ever-ELs (EL or RFEP) | 409 | 845 | 11.4 | 59.7 | 31.1 | 9.3 |
| To be determined | 0 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 546 | 846 | 11.3 | 56.8 | 31.5 | 11.7 |
| Special education services | 26 | 838 | 8.4 | 84.6 | 15.4 | 0.0 |
| Economically disadvantaged | 360 | 845 | 11.3 | 61.1 | 30.3 | 8.6 |
| Not economically disadvantaged | 212 | 847 | 11.3 | 52.8 | 31.6 | 15.6 |
| In U.S. schools less than 12 months | 2 | N/A | N/A | N/A | N/A | N/A |
| In U.S. schools 12 months or more | 570 | 846 | 11.3 | 57.9 | 30.9 | 11.2 |

Table 6.B.6 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 559 | 846 | 11.3 | 58.0 | 30.6 | 11.4 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 8 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 475 | 846 | 11.4 | 56.0 | 30.9 | 13.1 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 33 | 843 | 12.5 | 69.7 | 24.2 | 6.1 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 5 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Spanish as a Foreign Language Program | 16 | 842 | 11.1 | 75.0 | 25.0 | 0.0 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 88 | 845 | 11.4 | 62.5 | 27.3 | 10.2 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 331 | 847 | 10.8 | 51.7 | 34.4 | 13.9 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 79 | 841 | 13.4 | 70.9 | 21.5 | 7.6 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 61 | 843 | 9.3 | 68.9 | 26.2 | 4.9 |

Table 6.B.7 Percent of Students in Each Score Reporting Range for Total Scores by Demographic Variables for High School, All Grade Levels

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| All students | 253 | 942 | 13.1 | 70.4 | 23.7 | 5.9 |
| Male | 122 | 939 | 13.0 | 79.5 | 16.4 | 4.1 |
| Female | 131 | 945 | 12.6 | 61.8 | 30.5 | 7.6 |
| Nonbinary | 0 | N/A | N/A | N/A | N/A | N/A |
| American Indian or Alaska Native | 0 | N/A | N/A | N/A | N/A | N/A |
| Asian | 0 | N/A | N/A | N/A | N/A | N/A |
| Native Hawaiian or Other Pacific Islander | 0 | N/A | N/A | N/A | N/A | N/A |
| Filipino | 0 | N/A | N/A | N/A | N/A | N/A |
| Hispanic or Latino | 245 | 942 | 13.2 | 70.2 | 23.7 | 6.1 |
| Black or African American | 0 | N/A | N/A | N/A | N/A | N/A |
| White | 7 | N/A | N/A | N/A | N/A | N/A |
| Two or more races | 1 | N/A | N/A | N/A | N/A | N/A |
| Unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| English only | 20 | 940 | 12.5 | 75.0 | 25.0 | 0.0 |
| IFEP | 3 | N/A | N/A | N/A | N/A | N/A |
| EL | 91 | 939 | 13.4 | 74.7 | 19.8 | 5.5 |
| RFEP | 135 | 943 | 13.0 | 67.4 | 25.9 | 6.7 |
| Adult English learner (ADEL) | 0 | N/A | N/A | N/A | N/A | N/A |
| Ever-ELs (EL, RFEP, or ADEL) | 226 | 942 | 13.3 | 70.4 | 23.5 | 6.2 |
| To be determined | 4 | N/A | N/A | N/A | N/A | N/A |
| English proficiency unknown | 0 | N/A | N/A | N/A | N/A | N/A |
| No special education services | 237 | 943 | 12.1 | 68.8 | 25.3 | 5.9 |
| Special education services | 16 | 924 | 15.5 | 93.8 | 0.0 | 6.3 |
| Economically disadvantaged | 218 | 942 | 13.4 | 69.3 | 24.3 | 6.4 |
| Not economically disadvantaged | 35 | 942 | 11.4 | 77.1 | 20.0 | 2.9 |

Table 6.B.7 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Number Valid Score | Scale Score Mean | Scale Score SD | Reporting Score Range 1 | Reporting Score Range 2 | Reporting Score Range 3 |
| In U.S. schools less than 12 months | 35 | 943 | 12.0 | 68.6 | 22.9 | 8.6 |
| In U.S. schools 12 months or more | 218 | 941 | 13.3 | 70.6 | 23.9 | 5.5 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Total | 217 | 941 | 13.3 | 71.9 | 22.6 | 5.5 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—One-Way Immersion Program | 23 | 944 | 12.9 | 65.2 | 21.7 | 13.0 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Dual-Language Immersion Program | 79 | 941 | 13.6 | 70.9 | 22.8 | 6.3 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Developmental Bilingual Program | 22 | 936 | 13.5 | 86.4 | 9.1 | 4.5 |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Heritage Language or Indigenous Language Program | 10 | N/A | N/A | N/A | N/A | N/A |
| Received Instruction in Spanish in the 2020–‍2021 School Year—Spanish as a Foreign Language Program | 62 | 941 | 13.1 | 71.0 | 25.8 | 3.2 |
| Percentage of School-Day Instruction Provided in Spanish—0–25% | 44 | 942 | 16.6 | 63.6 | 29.5 | 6.8 |
| Percentage of School-Day Instruction Provided in Spanish—26–50% | 79 | 943 | 11.7 | 68.4 | 26.6 | 5.1 |
| Percentage of School-Day Instruction Provided in Spanish—51–75% | 33 | 936 | 12.9 | 84.8 | 12.1 | 3.0 |
| Percentage of School-Day Instruction Provided in Spanish—76–100% | 61 | 940 | 12.2 | 75.4 | 18.0 | 6.6 |

## Psychometric Analyses

This chapter describes in detail the psychometric analyses conducted for the 2020–2021 California Spanish Assessment (CSA) and the results of the analyses. These include classical item analyses and differential item functioning (DIF) analyses; item response theory (IRT) calibration, postequating for grades three to eight, and classical equating for high school; field test scaling; omission, expiration, and completion analyses; overall testing summaries as the means and standard deviations (SDs) of scale scores and theta values; the percentages of students at each score reporting range for each test; testing time analysis; reliability analysis; consistency and accuracy of the score reporting range classifications; and student group reliability.

### Overview

This section describes the data samples used for the statistical analyses, presents the results of the item and test analyses, and explains all statistical procedures implemented in the psychometric analyses, as well as the procedures designed to ensure the validity of score uses and interpretations.

#### Summary of the Analyses

The following list identifies the analyses that are conducted for a typical CSA administration. Each analysis is described in the subsequent narrative.

1. **Classical Item Analyses—**Classical item analyses for the CSA are discussed in section [*7.2 Classical Item Analyses*](#_Classical_Item_Analyses).
2. **DIF Analyses—**DIF analyses for the CSA are described in section [*7.3 Differential Item Functioning Analyses*](#_Differential_Item_Functioning).
3. **IRT Analyses—**IRT analyses, including calibration and scaling for the CSA, are described in section [*7.4 Item Response Theory Analyses*](#_Item_Response_Theory).
4. **Classical Equating for High School—**The classical equating analysis for the CSA high school assessment—undertaken because of the small sample size—is described in section [*7.5 Classical Equating Analysis for High School*](#_Toc92714105).
5. **Testing Time Analyses—**Response time analyses are described in section [*7.6 Testing Time Analyses*](#_Response_Time_Analyses).
6. **Reliability Analyses—**Reliability estimation for the CSA is illustrated in section [*7.7 Reliability Analyses*](#_Reliability_Analyses).
7. **Validity Evidence—**Validity evidence related to the CSA is discussed in section [*7.8 Validity Evidence*](#_Validity_Evidence).

Caution should be taken when comparing the test results from the 2020–2021 administration with results from 2018–2019 because the testing samples in the 2020–2021 administration are small—approximately 13 percent of the 2018–2019 test population in grades three through five, 20 percent in grades six through eight, and 6 percent in high school.

#### Sample Used for the Analyses

In general, analyses included in this technical report are based on all valid scores in the tested population. The actual data sample used depends on the time that data source became available as well as the information contained in the data to meet the analysis timeline.

Both classical item analysis and IRT calibration include students who logged on to the test and answered at least one item. The IRT analyses in [appendix 7.B](#_Appendix_7.B:_Item) were run using the version 1 of the production data file (“P1”) that was updated on June 4, 2021. The classical item analyses in [appendix 7.A](#_Appendix_7.A:_Classical) and item-level DIF analyses in table 7.7 present a summary of gender DIF classifications.

Table 7.7 is based on the complete data file available in July 2021, after the administration testing window was closed on July 30, 2021. All other analyses, such as the testing time analyses and reliability analyses, used the final version of the production data file for student reports, which became available in August 2021. All data sources include all valid student scores.

Note that the number of students in the item analysis and DIF sample is higher than the number of the final production data because the 2019–2020 data was included in the item analysis and DIF samples to meet the sample threshold for item analysis and DIF analysis (200 students per item). The reuse of the 2019–2020 form provides such an opportunity to combine the data from the 2019–2020 and 2020–2021 administrations; otherwise, the sample size of the embedded field test items in some grade levels—fewer than 200 students—may not be included in the analysis.

Table 7.1 shows small differences in student counts among the data sources used for IRT calibration analysis, classical item analysis and DIF analysis, and the final production data file. A small number of student scores was excluded from the final production data as a result of the data validation process. The numbers in the *Item Analysis and DIF Sample N* column for grades three through eight combine data from the 2019‒2020 and 2020‒2021 CSA administrations.

Table 7.1 Sample Size by Test Form

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level or Grade Band | Calibration Sample N | Item Analysis and DIF Sample N | Final Production Data N |
| Grade 3 | 1,186 | 1,371 | 1,213 |
| Grade 4 | 1,031 | 1,288 | 1,052 |
| Grade 5 | 894 | 1,112 | 916 |
| Grade 6 | 939 | 1,185 | 948 |
| Grade 7 | 718 | 922 | 729 |
| Grade 8 | 559 | 670 | 572 |
| High school | 226 | 231 | 253 |

### Classical Item Analyses

Classical item analyses are conducted to evaluate the performance of all operational test items with respect to item difficulty, item discrimination, and distractor analysis. In addition, the distributions of score categories on key-based, selected-response items and rule-based, machine-scored items are also included in the classical item analyses results. Lastly, the associated flagging rules of these statistics are used to identify items that are not performing as expected.

Items scored as one (correct) or zero (incorrect) are referred to as dichotomous items. Items with a maximum score greater than one are called polytomous items. Table 7.2 and table 7.3 present the summary results of item difficulty and item-total correlation, respectively, by grade level. Table 7.4 presents the summary results of flagged items in each form by grade level. In addition, [appendix 7.A](#_Appendix_7.A:_Classical) presents results of the classical item analyses, including item difficulty indices and item-total correlation coefficient for grades three through eight. The item type and associated item flags are also provided.

#### Classical Item Difficulty Indices

For dichotomous items, item difficulty is indicated by its *p*-value, which is the proportion of students who answer the item correctly. The range of *p*-values is from 0.00 to 1.00. Items with high *p*-values are easier items; those with low *p*-values are more difficult items.

The formula for the *p*-value for a dichotomous item is presented in equation 7.1. *Refer to the* [*Alternative Text for Equation 7.1*](#_Alternative_Text_for) *for a description of this equation.*

 (7.1)

where,

*Xij* is the number of students who answered item *i* correctly, and

*Ni* is the total number of students who were presented with item *i*.

For polytomous items, the difficulty is indicated by the average item score (AIS). The AIS can range from 0.00 to the maximum total possible points for an item. To facilitate the interpretation, the AIS values for polytomous items are often expressed as the proportion of the maximum possible score, which are equivalent to the *p-*values of dichotomous items. Desired AIS values for polytomous items generally fall within the range of 20 percent to 95 percent of the maximum obtainable item score; items with values outside this range are flagged for review.

The formula for the *p-*value for a polytomous item 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,

*Xij* is the score assigned for a given polytomous item *i* and student *j*,

*Ni* is the total number of students who were presented with item *i*, and

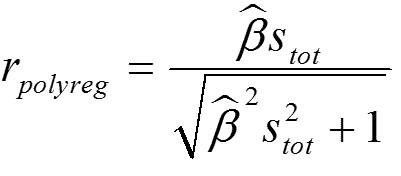
*Max (Xi)* is the maximum possible score for item *i*.

#### Item-Total Correlation

The item-total correlation statistic describes the relationship between students’ performance on a specific item and their performance on the total assessment. It is calculated as the correlation coefficient between the item score and total score.

In general, item-total correlation ranges from -1.0 (for a perfect negative relationship) to 1.0 (for a perfect positive relationship). A relatively high positive item-total correlation coefficient value is desired, as it indicates that students with higher scores on the overall test tend to perform better on the item. A negative item-total correlation typically signifies a problem with the item, as the students with higher scores on the overall test are more likely to get the item wrong or receive a low score, and the students with lower scores on the overall test are more likely to get the item correct or a high score.

For the CSA, the polyserial correlation is used for both polytomous and dichotomous items. Statistically, polyserial correlations are based on a polyserial regression model (Olsson, 1979; Drasgow, 1988), which assumes that performance on an item is determined by the examinee’s position on an underlying latent variable that is normally distributed at a given criterion score level. Polyserial correlation is an estimate of the correlation between the test score and the latent variable that determines the examinee’s performance on the item. Based on this approach, the polyserial correlation can be estimated using equation 7.3. *Refer to the* [*Alternative Text for Equation 7.3*](#_Alternative_Text_for_2) *for a description of this equation.*

 (7.3)

where,

*Stot* is the SD of the students’ total test scores as a criterion score, and

*β* is the item parameter to be estimated from the data, with the estimate denoted as , using maximum likelihood estimation (MLE). It is a regression coefficient (slope) for predicting the continuous version of an item score onto the continuous version of the total score.

There are as many regressions as there are boundaries between scores with all regressions sharing a common slope, *β*. For a polytomous item, there are *m-*1 regressions, where *m* is the number of score points on the item. Beta (*β*) is the common slope for all *m*-1 regressions. Desired polyserial correlation values of items are positive and larger than 0.20. Items with negative polyserial correlation values or values below 0.2 are flagged for review.

#### Distractor Analyses

The quality of distractors is an important component of an item’s overall quality. Distractors should be clearly incorrect, but at the same time be plausible and attractive to students who do not understand the content or skills being assessed. For the CSA, the following distractor analyses were conducted to evaluate the quality of distractors.

##### The Proportion of Students Choosing Each Distractor for High Performing Students

The percentage of students at each response option was calculated for the highest-performing 20 percent of students. If the percentage of students who selected a distractor was greater than the percentage of students who selected the correct answer for the high-performing group, the item would be flagged and examined to determine whether it has multiple correct answers or the wrong key (i.e., the item is miskeyed).

##### Polyserial Correlation

The polyserial correlation was calculated for each response option. While the key should have a positive polyserial correlation with the criterion score, the distractors should exhibit negative polyserial correlations (i.e., lower-ability students would likely choose the distractors, while higher-ability students would not). An item with a positive distractor-total correlation would be flagged for review, as this item may have multiple correct answers, be miskeyed, or have other content issues.

#### Omission and Completion Rates

An item is considered “omitted” if it was seen but not answered (i.e., it was left blank). Because students were not allowed to skip questions once they have started taking the CSA, and the only exception was when students skipped questions belonging to a reading passage or a paginated item group and exited out of the testing system, the possibility of an omission would be very low.

#### Distribution of Item Scores

For polytomous items, examination of the distribution of scores assists in showing how well the item performed. If no students achieved the highest possible score, the item may not be functioning as expected because the item may be confusing, poorly worded, unexpectedly difficult, or students may not have had an opportunity to learn the content.

Items with a low percentage (i.e., less than 3 percent) of students obtaining any possible item score were flagged for further review. Such items may pose problems during the IRT calibrations. They need to be carefully reviewed and may need to be excluded from the item calibration analyses.

#### Summary of Classical Item Analyses Flagging Criteria

In summary, an item was flagged for review if the item analysis yielded any of the following results. One item could have multiple flags if the statistics met the flagging criteria:

* **Difficulty flags** indicated extreme values of the proportion-correct (for dichotomous items) or the proportion of the possible maximum points earned (for polytomous items):
* A-flag: A *p*-value less than 0.2 for dichotomous items and polytomous items suggests that the item might be too difficult.
* H-flag: A *p*-value greater than 0.95 for dichotomous items and polytomous items suggests that the item might be too easy.
* A **discrimination flag** (R-flag) indicated that the item did not discriminate effectively between high- and low-ability students. Items with a polyserial correlation less than 0.20 were flagged.
* An **omit flag** (O-flag) was set for dichotomous items and polytomous items with nonresponse rates greater than 5 percent.
* A **distractor flag** (P-flag) was used for an item with any distractors having positive correlation with the criterion score.
* A **miskey flag** (D-flag) was used for multiple-choice items when more of the high-ability examinee group—the top 20 percent of examinees on the total assessment—choose any distractor rather than the response keyed as correct.
* An **underrepresented score point flag** (L-flag) was used for any item that had less than 3 percent of the students at any score level.

ETS’ Psychometric Analysis & Research staff and Assessment and Learning Technology Research & Development staff carefully reviewed each of the flagged items during and at the end of the item analyses. All flagged items were also reviewed by California educators at the data review meeting and then summarized for the California Department of Education (CDE) with recommendations for subsequent analyses.

#### Classical Item Analyses Results Summary

The summary statistics of the classical item analyses, which include the means and ranges of overall item difficulty and item-total correlation for all operational items, are presented in table 7.2 and table 7.3 for grades three through eight. There is a range of item difficulties with the *p-*values ranging from 0.04 to 0.86 and the average *p*-values ranging from 0.41 to 0.52. The difficulty distribution of the items shows that the largest group of items is in the range of 0.4 to 0.6, the second largest group of items is in the range of 0.2 to 0.4, and the third largest group of items is in the range of 0.6 to 0.8. Few items are in the highest difficulty group (ranging between 0 and 0.2) and the lowest difficulty group (ranging between 0.8 and 1.0).

The average item-total correlation ranged from 0.39 to 0.45. These values of the item-total correlations indicate that the items have acceptable levels of discrimination. For table 7.2 and table 7.3, the high school sample size was too small to conduct valid item analysis.

Table 7.2 Item Difficulty Distributions by Grade Level

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level | 0 ≤ *p* < 0.2 | 0.2 ≤ *p* < 0.4 | 0.4 ≤ *p* < 0.6 | 0.6 ≤ *p* < 0.8 | 0.8 ≤ *p* ≤ 1.0 | Total Number of Items | Mean *p*-value | Minimum *p*-‍value | Maximum *p*-‍value |
| Grade 3 | 2 | 23 | 23 | 4 | 0 | 52 | 0.41 | 0.17 | 0.80 |
| Grade 4 | 2 | 18 | 26 | 6 | 0 | 52 | 0.45 | 0.17 | 0.76 |
| Grade 5 | 4 | 17 | 22 | 8 | 1 | 52 | 0.45 | 0.13 | 0.83 |
| Grade 6 | 1 | 8 | 29 | 13 | 1 | 52 | 0.52 | 0.17 | 0.80 |
| Grade 7 | 1 | 14 | 27 | 9 | 1 | 52 | 0.49 | 0.04 | 0.82 |
| Grade 8 | 0 | 12 | 27 | 11 | 2 | 52 | 0.51 | 0.21 | 0.86 |

Table 7.3 Item-Total Correlation Distributions by Grade Level

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level | r < 0 | 0 ≤ r < 0.2 | 0.2 ≤ r < 0.3 | 0.3 ≤ r < 0.4 | 0.4 ≤ r < 0.5 | r ≥ 0.5 | Total Number of Items | Mean r | Minimum r | Maximum r |
| Grade 3 | 0 | 4 | 3 | 10 | 17 | 18 | 52 | 0.44 | 0.2 | 0.8 |
| Grade 4 | 0 | 1 | 11 | 9 | 15 | 16 | 52 | 0.43 | 0.1 | 0.7 |
| Grade 5 | 0 | 4 | 6 | 18 | 10 | 14 | 52 | 0.41 | 0.1 | 0.8 |
| Grade 6 | 0 | 4 | 1 | 10 | 18 | 19 | 52 | 0.45 | 0.1 | 0.8 |
| Grade 7 | 0 | 4 | 13 | 12 | 12 | 11 | 52 | 0.39 | 0.1 | 0.7 |
| Grade 8 | 0 | 3 | 9 | 13 | 15 | 12 | 52 | 0.41 | 0.1 | 0.7 |

The summary of flagged items in each test form by grade level is presented in table 7.4. Only elementary and middle school grade levels are presented; the high school sample size was too small to conduct valid item analysis. Note that there are 52 items on each form. All flagged items were reviewed by content experts and then summarized for, and reviewed by, the CDE. None of the flagged items were identified as having any content flaws during the thorough review by content experts and the CDE. The small sample size in the 2020–2021 administration could be a reason for some items being flagged.

Table 7.4 Flagged Items Summary in Each Form by Grade Level

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level | Form | No. of Items | No. of Flag A Items | Percent of Flag A Items | No. of Flag H Items | Percent of Flag H Items | No. of Flag R Items | Percent of Flag R Items | No. of Flag D Items | Percent of Flag D Items | No. of Flag P Items | Percent of Flag P Items | No. of Flag O Items | Percent of Flag O Items | No. of Flag L Items | Percent of Flag L Items |
| Grade 3 | 1 | 52 | 2 | 3.9 | 0 | 0 | 4 | 7.7 | 1 | 1.92 | 12 | 23.1 | 0 | 0 | 0 | 0 |
| Grade 4 | 1 | 52 | 3 | 5.8 | 0 | 0 | 1 | 1.9 | 6 | 11.54 | 15 | 28.9 | 0 | 0 | 1 | 1.9 |
| Grade 5 | 1 | 52 | 4 | 7.7 | 0 | 0 | 4 | 7.7 | 5 | 9.62 | 16 | 30.8 | 0 | 0 | 0 | 0 |
| Grade 6 | 1 | 52 | 1 | 1.9 | 0 | 0 | 4 | 7.7 | 1 | 1.92 | 11 | 21.2 | 0 | 0 | 1 | 1.9 |
| Grade 7 | 1 | 52 | 1 | 1.9 | 0 | 0 | 4 | 7.7 | 3 | 5.77 | 21 | 40.4 | 0 | 0 | 0 | 0 |
| Grade 8 | 1 | 52 | 1 | 1.9 | 0 | 0 | 3 | 5.8 | 3 | 5.77 | 17 | 32.7 | 0 | 0 | 3 | 5.8 |

Detailed results of the classical item analyses for each item by grade level are presented in [appendix 7.A](#_Appendix_7.A:_Classical). The summary statistics of item difficulty and item-total correlation coefficient by domain in each grade level are presented in Table 7.A.1 and table 7.A.2. The summary of item difficulty and item-total correlation by the test form in each grade level is listed in Table 7.A.3 through table 7.A.8. The maximum score points, item type, and associated item flag information for each item are also presented.

### Differential Item Functioning Analyses

Analyses of DIF can provide evidence of the degree to which an item score interpretation or use is valid for individuals who differ in their demographic characteristics. An item may be biased if it contains content or language that is differentially familiar to student groups. It is important, however, to recognize that item performance differences flagged for DIF might be related to actual difference in relevant knowledge or skills (group impact) or statistical Type I error, which might falsely assert DIF exists for an item. As a result, DIF statistics are used to identify *potential* item bias. Subsequent reviews by content experts and bias and sensitivity experts are required to determine the source and meaning of item performance differences.

DIF analyses were performed on all operational items. In examining the DIF between groups, the reference group is often designated as the group assumed to have an advantage, while the focal group refers to the group anticipated to be disadvantaged by the test. The sample size requirements for the DIF analyses were 100 in the smaller of either group and 400 in the combined focal and reference groups. These sample size requirements are based on standard operating procedures with respect to DIF analyses at ETS to ensure reliable DIF results can be obtained.

#### Differential Item Functioning Procedure for Dichotomous Items

The Mantel-Haenszel (MH) DIF (MH-DIF) statistic was calculated for dichotomous items (Mantel & Haenszel, 1959; Holland & Thayer, 1988). For this method, students are classified to relevant student groups of interest (e.g., gender or ethnicity). Students at each total-score level in the focal group (e.g., females) are compared with students at each total-score level in the reference group (e.g., males).

The common odds ratio is estimated across all levels of matched student ability using the formula in equation 7.4 (Dorans & Holland, 1993). The resulting estimate is interpreted as the relative probability of success on a particular item for members of two groups when matched on ability. *Refer to the* [*Alternative Text for Equation 7.4*](#_Alternative_Text_for_3) *for a description of this equation.*

Equation 7.4; a link to the long description for this equation is found in the preceding paragraph. (7.4)

where,

*m* is the number of score categories of the total test,

*Rrm* is the number of students in the reference group who answer the item correctly at score level *m*,

*Wfm* is the number of students in the focal group who answer the item incorrectly at score level *m*,

*Ntm* is the total number of students at score level *m*,

*Rfm* is the number of students in the focal group who answer the item correctly at score level *m*, and

*Wrm* is the number of students in the reference group who answer the item incorrectly at score level *m*.

To facilitate the interpretation of MH results, the common odds ratio is frequently transformed to the delta scale using equation 7.5 (Holland & Thayer, 1988). *Refer to the* [*Alternative Text for Equation 7.5*](#_Alternative_Text_for_4) *for a description of this equation.*

 (7.5)

Positive values indicate DIF in favor of the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values indicate DIF in favor of the reference group (i.e., negative DIF items are differentially harder for the focal group).

#### Differential Item Functioning Procedure for Polytomous Items

The standardization DIF (Dorans & Schmitt, 1993; Zwick, Thayer, & Mazzeo, 1997; Dorans, 2013) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959) is calculated for polytomous items. The standardized mean difference (SMD) compares the item means of the two groups after adjusting for differences in the distribution of students across all items and is calculated using equation 7.6. *Refer to the* [*Alternative Text for Equation 7.6*](#_Alternative_Text_for_5) *for a description of this equation.*

 (7.6)

where,

*X is* the criterion score (total raw score),

*Y is* the item score,

*M* is the number of score levels on *X*,

*Nfm* is the number of students in the focal group at score level *m*,

*Er* is the expected item score for the reference group, and

*Ef* is the expected item score for the focal group.

A positive SMDvalue means that, conditional on the criterion score, the focal group has a higher mean item score than the reference group (i.e., the item is differentially easier for the focal group). In contrast, a negative SMD value means that, conditional upon the criterion score, the focal group has a lower mean item score than the reference group (i.e., the item is differentially harder for the focal group).

#### Classifications

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C. Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values.

The flagging criteria for dichotomous items are defined in table 7.5; the flagging criteria for polytomous items are defined in table 7.6.

Table 7.5 DIF Categories for Dichotomous Items

|  |  |
| --- | --- |
| DIF Category | Criteria |
| A (negligible) | * Absolute value of MH D-DIF is not significantly different from zero at the 0.05 level or is less than one. * Positive values are classified as “A+” and negative values as “A-.” |
| B (moderate) | * Absolute value of MH D-DIF is significantly different from zero but not from one at the 0.05 level and is at least one; *or* absolute value of MH D-‍DIF is significantly different from one but is less than 1.5. * Positive values are classified as “B+” and negative values as “B-.” |
| C (large) | * Absolute value of MH D-DIF is significantly greater than one at the 0.05 level and is at least 1.5. * Positive values are classified as “C+” and negative values as “C-.” |

Table 7.6 DIF Categories for Polytomous Items

|  |  |
| --- | --- |
| DIF Category | Criteria |
| A (negligible) | * Mantel Chi-square *p*-value ≥ 0.05 level or |SMD/SD| ≤ 0.17 |
| B (moderate) | * Mantel Chi-square *p*-value < 0.05 level and 0.17< |SMD/SD| ≤ 0.25 |
| C (large) | * Mantel Chi-square *p*-value < 0.05 level and |SMD/SD| > 0.25 |

**Note:** SMD= standardized mean DIF; SD = total group standard deviation of item score

Table 7.7 presents a summary of gender DIF classifications.

Table 7.7 Gender DIF Classifications Summary by Grade Level and Grade Band

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DIF Category | Grade 3 Number | Grade 3 Percent | Grade 4 Number | Grade 4 Percent | Grade 5 Number | Grade 5 Percent | Grade 6 Number | Grade 6 Percent | Grade 7 Number | Grade 7 Percent | Grade 8 Number | Grade 8 Percent | High School Number | High School Percent |
| C- | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | N/A | N/A |
| B- | 0 | 0.00 | 0 | 0.00 | 3 | 5.77 | 0 | 0.00 | 0 | 0.00 | 4 | 7.69 | N/A | N/A |
| A- | 25 | 48.08 | 25 | 48.08 | 26 | 50.00 | 29 | 55.77 | 29 | 55.77 | 19 | 36.54 | N/A | N/A |
| A+ | 26 | 50.00 | 27 | 51.92 | 23 | 44.23 | 23 | 44.23 | 22 | 42.31 | 27 | 51.92 | N/A | N/A |
| B+ | 1 | 1.92 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 1.92 | 2 | 3.85 | N/A | N/A |
| C+ | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | N/A | N/A |
| Small N | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 52 | 100.00 |
| **Total** | **52** | **100.0** | **52** | **100.0** | **52** | **100.0** | **52** | **100.0** | **52** | **100.0** | **52** | **100.0** | **52** | **100.0** |

#### Items Exhibiting Significant Differential Item Functioning

DIF analyses for the gender group were conducted for the CSA for each grade level.

Table 7.7 provides detailed DIF results and shows the distributions of items across the DIF category classifications for each grade level and the grade band. In addition, “Small N” indicates that the DIF analysis was not performed because of insufficient sample size.

There were no C-DIF items for any grade level for the gender group comparison.

### Item Response Theory Analyses

IRT is built upon the item response function, which describes the probability of a given response as a function of a person’s true ability. IRT can be used to implement item calibrations, link item parameters, scale test scores across different forms or test administrations, evaluate item performance, build an item bank, and assemble test forms.

In the 2020–2021 administration, IRT calibration and postequating were applied in grades three through eight for item calibration, postequating, and scaling. Two steps were applied in item calibration and postequating for operational items and embedded field test items. The first step was for calibrating and equating the operational items. After operational items were calibrated in a full data matrix, robust-z procedures were used to evaluate the stability of anchor items through a comparison between the 2020–2021 calibration results and the bank item estimates. When a stable anchor set was obtained, the linking constant that linked the 2020–2021 items to the base scale was calculated and applied to the 2020–2021 operational items. Using the linked item estimates, students’ theta scores and scale scores were generated using the inverse test characteristic curve (TCC) method and linear transformation.

The second step was for the field test items. The field test items were calibrated in a sparse matrix containing both operational item scores and field test item scores. Then, the field test items were scaled back to the base scale using the equating constants calculated based on the anchor items. Refer to subsection [*7.4.2 Equating*](#_Equating_2) for detailed information about calibration, equating, and scaling.

The extremely small test sample in high school (226 students) does not support the IRT calibration and equating procedures. Instead, a classical linear equating method was used to equate the high school form to the base scale. Refer to section [*7.5 Classical Equating Analysis for High School*](#_Classical_Equating_Analysis) for detailed information. Accommodated forms were reused forms from the 2018–2019 administration. The conversion tables derived from the 2018–2019 administration were used directly for the accommodated forms.

#### Models

The one-parameter logistic item response theory (1PL-IRT) model was used for the CSA item calibration and was selected after consultation with the CDE. In particular, the generalized partial credit model (GPCM) (Muraki, 1992) restricted for 1PL-IRT, which is essentially the partial credit model (PCM) (Masters, 1982), was applied to both dichotomous and polytomous items.

The mathematical form of the GPCM is presented in equation 7.7. *Refer to the* [*Alternative Text for Equation 7.7*](#_Alternative_Text_for_6) *for a description of this equation.*

 (7.7)

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 and is fixed to 0.588 for every item,

*bi* is the location parameter for item *i*,

*div* is the category parameter for item *i* on item 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 7.7 becomes an expression of the one-parameter logistic (1PL) model for dichotomous items. Essentially, the 1PL model (Hambleton, Swaminathan, & Rogers, 1991) and the PCM (Masters, 1982) were used for dichotomous items and polytomous items, respectively.

#### Equating

Postequating was conducted using the 2020–2021 data. In this section, calibration, postequating for grades three through eight, and scaling are included. Classical linear equating for high school is discussed in section [*7.5 Classical Equating Analysis for High School*](#_Classical_Equating_Analysis).

##### Item Calibration

Operational items in the general forms at each grade level were calibrated when the 2020–‍2021 calibration data was available.

As stated in subsection [*7.4.1 Models*](#_Item_Response_Theory_1), the 1PL model (Hambleton, Swaminathan, & Rogers, 1991) and the corresponding PCM were jointly used to calibrate dichotomously and polytomously scored items*.* The software flexMIRT® (Cai, 2017) version 3.51 was used for calibration.

###### Data Preparation

Prior to IRT calibration analyses, ETS psychometricians reviewed the results of the classical item analyses to decide whether any items were of poor quality and needed to be removed from calibration. The results also were reviewed by ETS content experts and the CDE. The decision whether to remove items from calibration was made in consultation with the CDE.

For IRT calibration, scored item response data was used to create the IRT analysis input data files for each grade. The 2020–2021 IRT analysis input data file was a full matrix containing item-level scores for students who answered at least one item.

Similar to the classical item analyses, “omit” items were treated as incorrect and “not presented” items were treated as blank.

###### Description of the Calibration Procedure

FlexMIRT (Cai, 2017), a multilevel and multiple-group IRT software package for item analysis and test scoring, was used for CSA item calibration analysis. This software can fit a variety of IRT models to both single-level and multilevel data that are dichotomous, polytomous, or both, and was chosen for its superior flexibility among IRT software programs.

The procedure described next was followed to calibrate the student response data using flexMIRT for each grade:

1. Prepare and format the input data files as required by flexMIRT.
2. Prepare flexMIRT control files and specify the IRT models and analyses. (The 1PL‑IRT model and the corresponding PCM were used.)
3. Evaluate the flexMIRT output to examine whether every execution of flexMIRT analysis reached satisfactory convergence.
4. Review the item parameter estimates:
   1. At the form level, the summary statistics for the *b*-parameter estimates (location difficulty) and *d*-parameter estimates (step difficulty) are examined, including the mean, SD, median, minimum, maximum, and model-fit. The model-fit is evaluated using the root mean square error of approximation (RMSEA). RMSEA values less than 0.05 indicate good fit while RMSEA values greater than 0.10 indicate poor fit (Browne & Cudeck, 1993). The *b*-parameters are correlated with the *p*-values.
   2. At the item level, statistics of individual items are examined, including item difficulty estimates, model-fit statistics, and the IRT-based item parameters. The *b*-parameters and the *d*-parameters should be in the range of -4.0 to +4.0 with a standard error of 0.4 or less.
5. Flag items that did not perform as expected. (All flagged items were discussed thoroughly with the CDE to decide whether those items should be removed from calibration or whether the scoring categories needed to be collapsed.)

The calibration process was conducted independently by two ETS psychometricians to ensure quality and accuracy of results. Specifically, two psychometricians independently created flexMIRT control files and ran the same input data files and then compared the calibration results. Any differences in the output were investigated. Refer to section [*8.6 Quality Control of Psychometric Processes*](#_Quality_Control_of)for more details about this procedure.

Table 7.8 presents the item difficulty parameter distribution by grade level. Note that the most difficult item with a maximum value of 3.63 in grade seven is still within the interval [-‍4, 4].

Table 7.8 Item Difficulty Parameter Distribution by Grade Level

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IRT-*b* Range | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 |
| b < −3.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| −3.5 ≤ b < −3.0 | 0 | 0 | 0 | 0 | 0 | 0 |
| −3.0 ≤ b < −2.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| −2.5 ≤ b < −2.0 | 0 | 0 | 0 | 0 | 0 | 0 |
| −2.0 ≤ b < −1.5 | 1 | 2 | 1 | 1 | 0 | 2 |
| −1.5 ≤ b < −1.0 | 0 | 3 | 1 | 6 | 3 | 3 |
| −1.0 ≤ b < −0.5 | 6 | 7 | 10 | 6 | 3 | 6 |
| −0.5 ≤ b < 0 | 18 | 12 | 11 | 12 | 17 | 14 |
| 0 ≤ b < 0.5 | 7 | 14 | 15 | 18 | 11 | 16 |
| 0.5 ≤ b < 1.0 | 16 | 8 | 10 | 6 | 16 | 7 |
| 1.0 ≤ b < 1.5 | 3 | 5 | 1 | 2 | 1 | 4 |
| 1.5 ≤ b < 2.0 | 1 | 1 | 3 | 1 | 0 | 0 |
| 2.0 ≤ b < 2.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.5 ≤ b < 3.0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.0 ≤ b < 3.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| b ≥ 3.5 | 0 | 0 | 0 | 0 | 1 | 0 |
| Minimum | -1.98 | -1.72 | -1.84 | -1.86 | -1.44 | -1.91 |
| Maximum | 1.56 | 1.59 | 1.85 | 1.77 | 3.63 | 1.49 |
| Mean | 0.16 | 0.05 | 0.11 | -0.09 | 0.18 | -0.05 |
| SD | 0.65 | 0.76 | 0.74 | 0.73 | 0.75 | 0.72 |
| **Number of Items** | **52** | **52** | **52** | **52** | **52** | **52** |

##### Linking the Item Parameters

The new items in the tests for each grade level and the grade band were linked to a calibrated item pool using a common-item nonequivalent groups design (Kolen & Brennan, 2004b). The base scales for the CSA were established based on data from the 2018–‍2019 administration. The items of the new operational forms used were placed on the 2018–2019 base scale by using a set of linking items (i.e., anchor set) selected from the calibrated item pool and readministered in a typical year for each grade.

For the 2020–2021 administration, after IRT calibration was performed with the 2020–2021 items, the complete set of anchor items was used to calculate the linking constants to place the item parameters onto the 2018–2019 scale by using the mean-to-mean method described in the next subsection. The linking process was carried out iteratively by inspecting differences between the transformed 2020–2021 item estimates and base estimates for the anchor items and by removing items for which the item difficulty estimates changed significantly; this is called the robust-z procedure. Robust-z is also described in more detail in subsection [*7.4.2.2.2* *Robust-Z Procedure*](#_Robust-Z_Procedure_1).

###### Mean-to-Mean Transformation

The item difficulty estimates from a new form of calibration may not be comparable to those from the 2018–2019 calibration. The difficulty estimates based on a typical year’s data need to be transformed onto the base scale to make them comparable to the item bank parameters.

The mean-to-mean transformation assumes that the item bank and the new form difficulty values differ by a constant; that is, the item bank and the new form difficulty values can be made comparable by adding the same constant for all items. If this assumption is correct, then that constant is the difference between the means of the anchor items from the item bank and the new form difficulty values for the anchor items.

An iterative procedure was implemented to calculate the linking constants using common items in the item bank and the typical year’s administration. For each iteration of linking constants computation, the procedure described in subsection *[7.4.2.2.2](#_Robust-Z_Procedure_1)**[Robust-Z Procedure](#_Robust-Z_Procedure_1)* is intended to inspect the differences between the transformed new (a typical year) and base (2018–2019) estimates for the anchor items and remove anchor items for which the item difficulty estimates changed significantly.

There were eight steps involved in making mean-to-mean transformation:

1. Identify the anchor items in both the item bank (2018–2019 administration) and the 2020–2021 administration.
2. Obtain the item difficulty parameters (*b*-values) of these anchor items that are on the base scale from the item bank.
3. Obtain the item difficulty parameters (*b*-values) of these anchor items from the calibration of the new form.
4. Calculate the average item difficulty for the anchor set on the base scale.
5. Calculate the average item difficulty for the anchor set from the calibration of the new form.
6. Obtain the transformation constant by taking the difference between the two average item difficulties (*b*-values)—using the average item difficulty for the anchor set on the base scale—and subtracting the average item difficulty for the anchor set from the calibration of the new form.
7. Obtain a set of adjusted item difficulty parameters (*b*-values) by applying the linking constant to the item difficulty parameters of the anchor items from the new form of a typical-year administration.
8. Remove anchor items by following the procedure as described in subsection [*7.4.2.2.2 Robust-Z Procedure*](#_Robust-Z_Procedure_1). The iteration process continues by removing one unstable anchor in each round until no additional items are identified with significant differences between the item difficulty estimates for adjusted new and base items.

###### Robust-Z Procedure

To identify any unstable anchor items, ETS utilized an outlier detection procedure based on the robust-z statistic (Huynh, 2000; Huynh & Rawls, 2009). In this application, robust-z, as described in equation 7.8, was calculated based on the distribution of the difficulty difference for the anchor items between the item bank and the new form in a typical-year administration. *Refer to the* [*Alternative Text for Equation 7.8*](#_Alternative_Text_for_7) *for a description of this equation.*

 (7.8)

where,

*D* is the difference between the base and transformed new item difficulty of an anchor item,

*MdD* is the median of a distribution of *D* for all anchor items, and

*IQR* is the interquartile range of a distribution of *D* for all anchor items, which is defined as the difference between the third quartile (Q3) and the first quartile (Q1) when all the *D* values are rank ordered.

A large value of this statistic for any anchor item indicates that the reference item difficulty parameter and the new form item difficulty parameter for that item differed substantially.

The criterion for removing anchor items is that the robust-z value is greater than 1.645. One anchor item was removed at each iteration. The following criteria are evaluated at each iteration:

* The correlation between the reference item difficulty estimates and new form item difficulty estimates for the anchor sets should be no less than .95.
* The ratio of standard deviations (RSD) of the reference item difficulty estimates and the new form item difficulty estimates for the anchor items should be between .90 and 1.1.

After each iteration, the mean difference of the anchor sets between the base item-difficulty estimates and the new form item difficulty estimates was recomputed based on the remaining anchor items. Once the final anchor item set was obtained and the linking constant was calculated, ETS discussed its psychometric characteristics with the CDE and received approval from the CDE. Removed anchor items were not used in the computation of the linking constants but were still included in calibration and for deriving raw-to-theta conversions.

###### Evaluation of Linking

Three indices are used to evaluate the quality of the linking procedure: the RSD of the two sets of item difficulty estimates for the anchor items (i.e., the item bank estimates and the new form item calibration estimates), and the correlation between the two sets of item difficulty estimates for the anchor items (Huynh, 2009). If the correlation is at least 0.95 and the RSD is between 0.90 and 1.1, the linking results are considered acceptable, and all anchor items are regarded as stable in the linking process. The last index to evaluate the quality of the linking procedure is the percentage of anchor items in the form. The CDE requests a 40 percent overlap between anchor items and all items from the form after linking is complete.

Table 7.9 shows a summary of the procedure of evaluation of anchor items described previously for grades three through eight, which includes the number of all anchor items at the beginning, the number of anchor items removed as a result of mean-to-mean transformation and robust-z procedure, the number of remaining anchor items, and the linking constants of the final iteration of the test for each grade level. Note that classical equating procedures were used in the high school form. Therefore, no equating constant is reported for high school in this table.

Table 7.9 Final Linking Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade Level | Number of Items in Initial Anchor Set | Number of Items Removed from the Anchor Set | Number of Items in Final Linking Set | Linking Constant |
| Grade 3 | 35 | 5 | 30 | -0.278 |
| Grade 4 | 38 | 3 | 35 | -0.224 |
| Grade 5 | 35 | 1 | 34 | -0.180 |
| Grade 6 | 34 | 10 | 24 | -0.008 |
| Grade 7 | 38 | 3 | 35 | 0.079 |
| Grade 8 | 38 | 3 | 35 | 0.008 |

Table 7.10 presents the total number of operational items on the general form for grades three through eight, the number of remaining anchor items after robust-z evaluation, the percentage of remaining anchor items out of all the operational items on the general form, the correlation between the final anchor set of the 2020–2021 item estimates and the reference (2018–2019 administration) difficulty estimates for the anchor items, and the RSD between the final set of the 2020–2021 item estimates and the reference (2018–2019 administration) difficulty estimates for the anchor items.

As table 7.10 shows, the correlation is 0.95 or higher and the RSD is between 0.95 and 1.05. The percentage of anchor items in a form is 46 or higher across grades. All three indexes to evaluate the quality of equating are met for grades three through eight.

Table 7.10 Evaluation of Anchor Set (Common Items) Between 2018–2019 and 2020–‍2021

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade Level | Number of Operational Items | Anchor Items Remaining After Deletions | Remaining Anchor Items as a Percentage of All Operational Items | Correlation Between Item Bank Item Parameters and 2020–‍2021 Item Parameters | RSD of Item Parameters Between Item Bank and 2020–‍2021 Administration |
| Grade 3 | 52 | 30 | 58% | 0.95 | 0.96 |
| Grade 4 | 52 | 35 | 67% | 0.98 | 1.01 |
| Grade 5 | 52 | 34 | 65% | 0.98 | 1.00 |
| Grade 6 | 52 | 24 | 46% | 0.98 | 1.04 |
| Grade 7 | 52 | 35 | 67% | 0.97 | 0.96 |
| Grade 8 | 52 | 35 | 67% | 0.98 | 0.97 |

Table 7.11 presents the summary statistics of the final linking results after items with unstable parameters were detected and removed from the anchor set in grades three through eight. The statistics include the number of remaining items in the final anchor set, average item difficulties of the anchor set both in the 2018–2019 base scale and from the 2020–2021 administration, along with their differences, as well as the criteria for evaluating the absolute differences.

Table 7.11 Linked Item Parameter Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade Level | Number of Anchor Items | Operational Baseline Scale Average *b*-‍parameter | Linked 2021 Test Average *b*-‍parameter | Difference of Average *b*-‍parameters | Criteria for the Acceptable Absolute Difference |
| Grade 3 | 30 | 0.0948 | 0.0948 | 0.0001 | <0.1 |
| Grade 4 | 35 | -0.0571 | -0.0571 | 0.0000 | <0.1 |
| Grade 5 | 34 | 0.1558 | 0.1614 | -0.0056 | <0.1 |
| Grade 6 | 24 | -0.1750 | -0.1750 | -0.0001 | <0.1 |
| Grade 7 | 35 | 0.1330 | 0.1330 | 0.0000 | <0.1 |
| Grade 8 | 35 | -0.0904 | -0.0904 | 0.0000 | <0.1 |

##### Summary of Linked Item Response Theory *b*-parameters

The overall summary of the linked IRT *b*-value estimates for the 2020–2021 CSA administration calibration is shown in table 7.12. The mean, SD, minimum, and maximum values are presented, in addition to the number of items for each grade level. The RMSEA values are also provided in table 7.12, which are below 0.05 for all grade levels, indicating good model fit (Browne & Cudeck, 1993).

Table 7.12 IRT Summary *b*-value Estimates for All CSA Operational Items

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level | Number of Items | Average of *b-*value | SD *b*‑value | Minimum *b*-value | Maximum *b*-value | RMSEA |
| Grade 3 | 52 | 0.16 | 0.65 | -1.98 | 1.56 | 0.04 |
| Grade 4 | 52 | 0.05 | 0.76 | -1.72 | 1.59 | 0.04 |
| Grade 5 | 52 | 0.11 | 0.74 | -1.84 | 1.85 | 0.04 |
| Grade 6 | 52 | -0.09 | 0.73 | -1.86 | 1.77 | 0.04 |
| Grade 7 | 52 | 0.18 | 0.75 | -1.44 | 3.63 | 0.03 |
| Grade 8 | 52 | -0.05 | 0.72 | -1.91 | 1.49 | 0.04 |

All *b*-values were between -4.0 and +4.0. The average *b-*parameters for CSA tests in grades three through five and grade seven were above zero, indicating that, in general, the items were relatively difficult for the students in those grade levels. The mean *b*-parameters for grades six and eight are below zero, indicating the items, on average, were relatively easier for those students in those grade levels.

Table 7.B.1 through table 7.B.6 in [appendix 7.B](#_Appendix_7.B:_Item) provide the linked IRT difficulty and step parameter estimates at the item level in each grade level.

##### Scaling the Scores

The number-correct scores (raw scores) of each form were transformed to scale scores by a three-step process for grades three through eight. First, the item-difficulty estimates for each grade level were calibrated and linked to the base scale for the item bank. Then, the number-correct scores (raw scores) of each form were transformed to ability (theta) scores that were used to establish the reporting scale by the inverse TCC procedure as described in subsection [7.4.2.4.1](#_Inverse_Test_Characteristic). Finally, these ability (theta) scores were transformed to scale scores through a linear transformation.

The requirements that are particularly applied to the CSA reporting scale are also listed in subsection [*7.4.2.4.2 Transformation from Theta Scores to Scale Scores*](#_Transformation_from_Theta).

###### Inverse Test Characteristic Curve Procedure

After all the item difficulty estimates were calibrated and linked to the reference scale derived from administration data, students’ overall ability estimates were derived from the input data file*,* through the IRT inverse TCC method (Stocking, 1996). This method transformed the sum of the student’s item scores into an ability estimate. That estimate is the ability value that makes the sum of the expected scores on the items administered to the student equal to the sum of the scores that the student actually received on those items.

The TCC expresses the expected total score on a set of items as a function of the student’s ability, which is shown 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)

where,

*ndich* is the number of dichotomous items in the test,

*Pi(θ)* is the probability of a correct response to item *i* at ability *θ* on the dichotomous item in equation 7.7,

*npoly* is the number of polytomous items in the test,

*m* is the number of score categories for each polytomous item,

*sxj* is the value for score category *x* for the polytomous item *j*,

*Pxj(θ)* is the probability that an examinee with ability *θ* obtains score *sx* on the polytomous item *j* in equation 7.7, and

*ξ(θ)* is the corresponding expected total score.

###### Transformation from Theta Scores to Scale Scores

Students’ ability estimates (theta scores) were transformed to the scale score metric by applying a linear transformation based on threshold theta values. Those threshold values were determined after standard setting and approved by the California State Board of Education (SBE). Table 7.13 shows the theta and scale score thresholds established from standard setting. The scaling transformation from theta to scale score is shown 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,

*SS* represents the reporting scale score,

*θ* represents student ability,

*a* represents the intercept of the linear transformation, and

*b* represents the slope of the linear transformation.

The slope is calculated in equation 7.11 with range 3 thresholds. *Refer to the* [*Alternative Text for Equation 7.11*](#_Alternative_Text_for_19) *for a description of this equation.*

 (7.11)

The intercept is calculated in equation 7.12 with range 3 thresholds. *Refer to the* [*Alternative Text for Equation 7.12*](#_Alternative_Text_for_20) *for a description of this equation.*

 (7.12)

where,

*SSthreshold3* represents the reporting scale score range 3 threshold,

 represents the SD of the reporting scale score,

 is the SD of the theta scores, and

*θ threshold3* is the reporting range 3 threshold theta score.

The slope and intercepts derived from equations 7.11 and 7.12 defined the linear relationship between the theta scale and the scale-score scale based on the 2018–2019 data. *Refer to the* [*Alternative Text for Equation 7.13*](#_Alternative_Text_for_25) *for a description of this equation.*

 (7.13)

The slopes and intercepts for each grade and content area are shown in table 7.13**Error! No bookmark name given.**.

Table 7.13 Conversion of Theta Score to Reporting Scores by Grade Level or Grade Band Based on the 2018–2019 Administration (the Baseline)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | *σθ* | *σSS* | Reporting Score Range 2 Threshold Theta Score | Reporting Score Range 3 Threshold Theta Score | Reporting Score Range 2 Threshold | Reporting Score Range 3 Threshold | Slope *b* | Intercept *a* |
| Grade 3 | 0.732 | 10 | 0.037 | 0.860 | 349 | 360 | 14 | 348 |
| Grade 4 | 0.798 | 10 | 0.045 | 0.922 | 449 | 460 | 13 | 448 |
| Grade 5 | 0.727 | 10 | -0.114 | 0.886 | 546 | 560 | 14 | 548 |
| Grade 6 | 0.692 | 10 | -0.128 | 0.703 | 648 | 660 | 14 | 650 |
| Grade 7 | 0.656 | 10 | 0.128 | 1.179 | 744 | 760 | 15 | 742 |
| Grade 8 | 0.681 | 10 | 0.137 | 0.981 | 848 | 860 | 15 | 846 |
| High school | 0.727 | 10 | 0.171 | 0.916 | 950 | 960 | 14 | 947 |

##### Field Test Item Calibration and Linking

The CSA uses the embedded field test design to update item pools for future form assembly. As table 3.7 indicates, 30 embedded field test items were tested per grade in grades three through eight in the 2020–2021 administration. Among the 30 field test items, each 10-item set was embedded in one of the three versions of the test form per grade. Under the same design, 46 embedded field test items were included for high school. Because of the small test sample in high school, item analysis and IRT analysis were not conducted for high school.

In the 2020–2021 administration, the testing samples across grade levels were much smaller than the 2018–2019 administration, as shown in table 1.1. Field test item calibration and scaling were conducted in August, after the testing window was closed, to allow all possible data to be included in the analyses. The procedures for field test calibration and scaling were as follows:

1. Generate a sparse matrix with all operational items and field test items for each grade, including all item scores
2. Calibrate the embedded field test items and the operational items using flexMIRT software
3. Scale the calibrated field test items to the base scale by updating the field test item estimates using the calculated linking constant based on all operational items (In this process, the robust-z and mean-to-mean procedures mentioned previously were applied.)
4. Update the item pools

### Classical Equating Analysis for High School

Results for small samples cannot be equated effectively. The smaller the sample size, the more restricted the class of stable equating methods (Dorans, Moses, & Eignor, 2010). In the 2020–2021 administration, only 226 valid scores from the CSA for high school were available for calibration and equating analysis. Because only four students in high school took the 2019–2020 CSA, combining results for the two administrations still did not produce an effective sample size.

The IRT calibration and linking analyses based on the 226 scores produced undesirable results, including the following:

1. There were larger standard errors of IRT *b*-parameter estimates. The standard errors of IRT *b*-parameters for grades three through eight were in the range of 0.05 to 0.08, while standard errors of IRT *b*-parameters for high school range from 0.13 to 0.27.
2. The correlation between IRT *b*-parameter (difficulty) and *p*-value was 0.87, while the correlation coefficients for grades three through eight ranged from 0.97 to 0.99.
3. The equating function could not adequately adjust for form difficulty.
   1. The spring 2021 sample appears to be a less able sample based on anchor item performance (refer to table 7.C.1).
   2. The spring 2021 form appears to be easier because, despite being a less able sample, the performance on the total test was better than 2019. However, the high school IRT equating results showed that the percentage of students who were in the highest score reporting range in 2020–2021 nearly doubled the percentage of students in the highest score reporting range in 2018–2019 (16.7 percent to 8.9 percent).
   3. There was a lower percentage of students in the low score reporting ranges in 2020–2021 than those in 2018–2019. In contrast, the results in grades three through eight were more consistent and reasonable between 2018–2019 and 2020–2021.

Those undesirable practical effects of IRT calibration and equating were caused mostly by the small sample size. With a small sample size, large sampling errors may lead to an equating function that differs substantially from that of the population. ETS psychometricians conducted empirical investigation of the alternative—classical linear equating.

#### Methodology of the Classical Linear Equating

Literature suggests the potential and feasibility of classical linear equating methods with small sample sizes. Kolen (1985) examined small samples of 100 and 250 and found standard errors of equating in classical linear equating, when derived without the normality assumption, to be sufficiently accurate with sample sizes of 250. Parshall, Houghton, and Kromrey (1995) examined the effects of sample size on the stability and bias of linear equating with small samples, such as 15, 20, 50, and 100. Their results suggested that if the cutoff score is near the mean score for the test, where standard errors are smallest, equating can probably be supported for small samples. A study from Kim, von Davier, and Haberman (2006) suggested that the identity function was preferable with very small samples (e.g., 10 or 25), but the use of the synthetic function was more promising than other functions with moderately small samples (e.g., 50 or 100). When the sample size increases to more than 100, some nontrivial linking functions, such as chain linear equating, are preferable.

ETS psychometricians explored several classic linear equating methods under the nonequivalent groups with anchor test (NEAT) design, as has been used with the CSA. Under the NEAT design, observed score equating methods and test linking methods are generally based on the following linear equating function. *Refer to the* [*Alternative Text for Equation 7.14*](#_Alternative_Text_for_26) *for a description of this equation.*

 (7.14)

where,

*x* is the score from the form to be equated,

*y* is the score from the reference form,

*µ(X)* and *µ(Y)* are means of scores from Form X and Form Y, respectively,

*σ(X)* and *σ(Y)* are the SDs of Form X and Form Y scores, respectively, and

*x* and *y* are individual scores from Form X and Form Y.

Solving for *y* results in the formula for the linear equating function. *Refer to the* [*Alternative Text for Equation 7.15*](#_Alternative_Text_for_23) *for a description of this equation.*

 (7.15)

where,

*ly(x)* is the linear conversion equation for converting observed scores on Form X to the scale of Form Y (Kolen, M.J. & Brennan, R.L. 2004b, p. 106; von Davier, 2003, p.2).

When the SDs of the two forms are equal to one, or equal to each other, the linear equating becomes mean equating.

Usually, *X* and *Y* are two tests given to two samples from the two test administrations *P* and *Q*, respectively, and *V* is the anchor test given to both *P* and *Q*. *X*, *V*, and *Y* refer to both tests and test scores. The target population, *T*, for the NEAT design is assumed to be a mixture of *P* and *Q* and denoted by equation 7.16. *Refer to the* [*Alternative Text for Equation 7.16*](#_Alternative_Text_for_24) *for a description of this equation.*

 (7.16)

where,

*w* is the ratio of the sample size of the group from *P* and the sum of the sample sizes of the two groups.

When *w* = 1, then *T* = *P*, and when *w* = 0, then *T* = *Q*. Other choices of *w* may be used as well (von Davier, 2003, p.2).

A number of classical linear equating methods, such as Tucker mean equating, Tucker linear equating, Levine mean equating (Kolen & Brennan, 2004b), Levine observed score linear equating (Kolen & Brennan, 2004b), and chain linear equating (Kolen & Brennan, 2004b) are based on the linear equating functions of equation 7.13 through equation 7.15, although the underlying assumptions of those equating procedures are different. Mainly, Tucker equating methods are based on the assumptions that the linear regression of *X* on *V*, or the linear regression of *Y* on *V*, is the same in the two populations, *P* and *Q*. The main assumption for Levine equatings is that the true scores of the tests, *Tx* and *Ty* and the tests of the anchor (common items that link the two forms), *Tv* , in the two populations are perfectly correlated. The regression of *Tx* on *Tv*, or the regression of *Ty* on *Tv*, is linear and the same in the two populations, *P* and *Q*. The assumption for the chain linear equating is that the (linear) linking function from *X* on *V*, or the linear linking function from *V* on *Y*, is the same in the two populations, *P* and *Q* (von Davier, 2003).

#### Empirical Analysis with the Classical Linear Equatings

Table 7.C.1 in [appendix 7.C](#_Appendix_7.C:_Classical) summarizes the number of students, mean overall raw scores, mean anchor scores, and the SDs for the means over the 2018–2019 and 2020–2021 administrations. Also, the correlations of overall and anchor scores for the two administrations, respectively, are presented. The high correlation between overall and anchor scores are 0.95 and 0.96 for the 2018–2019 and 2020–2021 forms. Note that 58 percent of the anchor items are in the 2020–2021 form. Also, the content representativeness of the anchor set was verified.

Several classical linear equating methods, such as Tucker mean equating, Tucker linear equating, Levine mean equating, Levine observed score linear equating, Levine true score linear equating, and chain linear equating, were applied and evaluated for the high school CSA equating study. Several equating analysis software, such as the CIPE equating and linking program (Kolen & Brennan, 2004a) and an R package, “Equating” (Albano, 2016), were used for linear equating. All equating results were verified by at least two different equating software for identical results.

Table 7.C.2 and figure 7.C.1 summarize the standard errors of equating for the selected equating functions. The standard errors of equating were derived using Bootstrap methods with 100 replications. The standard errors of equating are the SDs of the obtained equated scores over the 100 bootstrap samples.

#### Evaluation of the Classical Equating Results

Several procedures were used to evaluate the classical linear equating results and select the most robust procedure. First, consistency checks between current results and the 2018–‍2019 results were conducted. In particular, the raw score thresholds for the reporting score ranges as well as the percentages of students in each reporting score range over the two years were compared. Results showed that the Tucker equating methods adequately adjusted the form difficulty and produced reasonable and consistent equating results over the two years.

Second, the standard errors of equating were evaluated. The mean equating functions showed lower standard errors across all equating functions, as presented in figure 7.C.1. Although the Tucker mean standard error and Levine mean standard error were similar, the Tucker mean equating process showed better results in the consistency checks. Therefore, the Tucker mean equating method was selected for use to report the scores for high school test takers because of its robustness with small sample sizes.

### Testing Time Analyses

The length of time it takes students to complete an assessment is recorded and analyzed to build a profile describing what a typical testing event looks like for each grade-level assessment. In addition, variability in testing time is investigated to determine whether a student’s testing time should be viewed as unusual or irregular for further investigation. It should be noted that the CSA tests are untimed.

In these analyses, all students who completed testing with a valid reporting scale score are included. The testing population is partitioned into performance quartiles based on all operational items. The descriptive statistics—for example, the number of students, mean, SD, minimum and maximum, percentiles—of the time required to complete the total test are computed for each of the four performance quartile groups for each grade level (i.e., grades three through eight and the high school grade levels).

[Appendix 7.D](#_Appendix_7.D:_Response) summarizes the results of the testing time analysis. Table 7.D.1 through table 7.D.7 provide descriptive statistics of total testing time for the full student population at each ability level for each grade level. The four quartiles, Q1 through Q4, are represented by four raw score intervals. The unit of testing time is in minutes; for example, in Table 7.D.1, the median (i.e., 50th percentile) of the testing time is 44.24 minutes for the grade three, Q1 (raw score interval 0–18) group of students who took the general form.

In general, students who took the general forms at the lowest quartile level (Q1), the lowest raw score interval, have shorter testing times than students in the other quartile groups. The median total testing time generally increases as the quartile level increases from the first quartile to the last quartile (Q4), the highest raw score interval, meaning that the students who performed better on the CSA tended to spend more time on the test.

### Reliability Analyses

Reliability is the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations due to measurement error. Thus, reliability is the consistency of scores across conditions that do not differ systematically and only contain random measurement errors. In statistical terms, the variance in the distributions of test scores—essentially, the differences among individuals—is due partly to real differences in the knowledge, skill, or ability being tested (true variance) and due partly to measurement error inherent in the measurement process (error variance). The reliability coefficient is an estimate of the proportion of the total variance that is true variance.

Reliability coefficients range from 0 to 1. 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 test.

The standard error of measurement (SEM) quantifies the amount of inconsistency in the test scores. SEM is the extent to which students’ scores tend to differ from the scores they would receive if the assessment were perfectly reliable. The larger the SEM, the more the students’ scores would tend to vary over repeated testing. Observed scores with large SEM pose a challenge to the valid interpretation of a single test score. For the CSA, reliability and SEM estimates were calculated at the test-form level.

For a typical administration, the CSA also reports the reliability of classification, which is an estimate of the proportion of students who are accurately and consistently classified into score reporting ranges. There are two kinds of classification reliability statistics: decision accuracy and decision consistency. Decision accuracy is the agreement between the classifications actually made and the classifications that would be made if the test scores were perfectly reliable. Decision consistency is the agreement between the classifications that would be made on two test forms.

#### Marginal Reliability

Coefficient alpha (Cronbach, 1951), which measures internal consistency reliability, is the most commonly used estimate of alternate-forms reliability. Coefficient alpha is estimated by substituting sample estimates for the parameters and is defined using equation 7.17. *Refer to the* [*Alternative Text for Equation 7.17*](#_Alternative_Text_for_10) *for a description of this equation.*

 (7.17)

where,

*K* is the number of items in the test,

 is the observed variance of item *i* in the test, and

 is the observed variance of the total test score.

Since CSA forms have mixed item types (dichotomous and polytomous items), it is more appropriate to report stratified alpha (Feldt & Brennan, 1989). Stratified alpha is a weighted average of coefficient alphas for item sets with different maximum score points, or “strata.” It is a reliability estimate computed by dividing the test into parts (strata), computing coefficient alpha separately for each part, and using the results to estimate a reliability coefficient for the total score. The formula for the stratified alpha is presented in equation 7.18. *Refer to the* [*Alternative Text for Equation 7.18*](#_Alternative_Text_for_11) *for a description of this equation.*

 (7.18)

where,

 is the variance for strata *j* of the test,

 is the total variance of the test, and

 is the Cronbach’s alpha for strata *j* of the test.

Estimates of stratified alpha are computed by substituting sample estimates for the parameters in the formula.

#### Standard Error of Measurement for Raw Scores

The SEM provides a measure of score instability in a different metric.

The formula for the SEM is presented in equation 7.19. *Refer to the* [*Alternative Text for Equation 7.19*](#_Alternative_Text_for_12) *for a description of this equation.*

 (7.19)

where,

 is the reliability estimated in equation 7.19, and

 is the SD of the total score.

Table 7.14 gives the reliability and SEM for each CSA form, along with the number of items and students upon which those analyses were performed. In the table, form 1 represents the general form while form A represents the accommodated form.

Table 7.14 Test Reliability of the Total Scores

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Form | No. of Items | N Points | N Students | Mean | SD | Reliability | SEM |
| Grade 3 | 1 | 52 | 62 | 1,183 | 26.06 | 9.60 | 0.87 | 3.52 |
| Grade 3 | A | 52 | 60 | 17 | 18.82 | 4.99 | 0.57 | 3.26 |
| Grade 4 | 1 | 52 | 63 | 1,020 | 28.73 | 9.16 | 0.86 | 3.45 |
| Grade 4 | A | 52 | 61 | 12 | 22.67 | 6.13 | 0.72 | 3.26 |
| Grade 5 | 1 | 52 | 62 | 890 | 27.50 | 8.78 | 0.84 | 3.50 |
| Grade 5 | A | 52 | 63 | 14 | 23.43 | 5.59 | 0.62 | 3.43 |
| Grade 6 | 1 | 52 | 65 | 918 | 34.34 | 10.14 | 0.87 | 3.61 |
| Grade 6 | A | 52 | 64 | 11 | 22.91 | 6.83 | 0.76 | 3.34 |
| Grade 7 | 1 | 52 | 65 | 691 | 32.24 | 8.74 | 0.83 | 3.66 |
| Grade 7 | A | 52 | 63 | 12 | 22.92 | 4.21 | 0.35 | 3.39 |
| Grade 8 | 1 | 52 | 65 | 541 | 34.14 | 8.97 | 0.84 | 3.59 |
| Grade 8 | A | 52 | 58 | 12 | 20.83 | 6.26 | 0.75 | 3.14 |
| High school | 1 | 52 | 62 | 224 | 31.21 | 9.61 | 0.87 | 3.53 |
| High school | A | 52 | 62 | N/A | N/A | N/A | N/A | N/A |

These results indicate that the reliability estimates for all tests are moderately high. Reliability coefficients for the 2020–2021 CSA range from 0.83 to 0.87 for the general forms and from 0.35 to 0.76 for the accommodated forms. The reliability coefficients for the general forms are above 0.80, which is acceptable for standardized assessments. Caution should be taken when reviewing reliability estimates for accommodated forms, because the extremely small test samples for accommodated forms—11 to 17 students in grades three through eight—may lead to very limited score variation, which could result in low reliability estimates.

Results based on samples that contain 50 or fewer examinees should be interpreted with caution because of small sample sizes.

#### Student Group Reliabilities and Standard Errors of Measurement

For a typical administration, CSA reliabilities are estimated for various student groups that tested. The student groups included in these analyses are defined by their gender, economic status, provision of special services, length of attendance in US schools, whether they received instruction in Spanish, and English language acquisition status. Reliabilities and SEM information for the total test scores by test form are reported for each student group analysis.

Reliability values are estimates that approach the true reliability as the number of students whose scores contribute to the estimates increases. Reliabilities are not reported for samples that comprise 10 or fewer students. Results based on samples that contain 50 or fewer students should be interpreted with caution, because these estimates may meaningfully deviate from the true reliability. In some cases, score reliabilities were not estimable and are presented in the tables as “N/A.”

Table 7.E.1 through table 7.E.7 in [appendix 7.E](#_Appendix_7.E:_Reliability) present the overall test reliabilities for the various student groups. Most student groups have reliability greater than 0.80 for the general forms across all seven grade levels, with the exception of

* students who received special education services;
* students who were not receiving Spanish instruction for grades three through five and grade seven; and
* English learner (EL) students in grades five, seven, and eight.

Among those groups, the lowest reliability value, 0.66, is for students who received special education services in grade five. It should be noted that in this case, the low reliability was likely due to the lack of variation in student performance because of the small number of students in this student group.

#### Theta Scores Standard Error

The SEM is the SD of the distribution of theta scores that the student would earn under different testing conditions. The test information function (TIF) is the sum of information from each item on the test. In the framework of IRT, when theta is estimated through an MLE, the reciprocal of the square root of the TIF provides an approximate value for the SEM. For the CSA, theta scores are obtained through an IRT inverse TCC approach of the 1PL-IRT model. For the 1PL-IRT model, the inverse TCC method produces the same estimate of theta as MLE. Therefore, the SEM for a student with proficiency 𝜃𝑗 is calculated using equation 7.20. *Refer to the* [*Alternative Text for Equation 7.20*](#_Alternative_Text_for_13) *for a description of this equation.*

 (7.20)

where,

*I(θj)* is the test information for student *j*, and is calculated as presented in equation 7.21. *Refer to the* [*Alternative Text for Equation 7.21*](#_Alternative_Text_for_14) *for a description of this equation.*

 (7.21)

where,

*Ii(θj)* is the item information of item *i* for student *j*.

Item information is calculated as presented in equation 7.22. *Refer to the* [*Alternative Text for Equation 7.22*](#_Alternative_Text_for_15) *for a description of this equation.*

 (7.22)

where,

*Si(θj)* is the expected item score for item *i* on a theta score *θj* calculated using equation 7.23 (*Refer to the* [*Alternative Text for Equation 7.23*](#_Alternative_Text_for_16) *for a description of this equation.)*

, (7.23)

and equation 7.24 (*Refer to the* [*Alternative Text for Equation 7.24*](#_Alternative_Text_for_17) *for a description of this equation.)*

 (7.24)

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 7.7; and

*ni* is the maximum number of score points for item *i*.

#### Scale Scores Conditional Standard Errors

Conditional standard errors of measurement (CSEMs) for scale scores are computed by transforming SEMs of theta scores onto the reporting scale. Refer to subsection [*7.4.2.4 Scaling the Scores*](#_Scaling_the_Scores) for scaling factors of transformation. A student’s CSEM under the IRT framework is equal to the reciprocal of the square root of the TIF multiplied by the scaling factor *a*, as presented in equation 7.25. *Refer to the* [*Alternative Text for Equation 7.25*](#_Alternative_Text_for_18) *for a description of this equation.*

 (7.25)

where,

;

*CSEM*(*SS*) is the CSEM on the reporting score scale;

 is the TIF at ability level  as shown in equations 7.19, 7.20, and 7.21;

*b* is the scaling factor (the slope) needed to transform theta to the scale score metric; and

*a* is the intercept of the linear transformation.

Refer to equations 7.10, 7.11, and 7.12 for calculation.

CSEMs vary across the scale and are typically smaller in scale score units toward the center of the scale where more items are located, whereas they are larger at the extreme ends of the scale. When a test has threshold scores, it is important to provide CSEMs at the threshold scores.

Table 7.15 presents the scale score CSEMs at the lowest score required for a student to be classified in the score reporting range 2 and score reporting range 3 for each CSA. The CSEMs for high school were calculated using method IV, based on Lord’s study (Lord, 1984).

Table 7.15 Scale Score CSEMs at the Reporting Score Range Thresholds

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Level 2―Reporting Score Range Threshold | Level 2 CSEM | Level 3―Reporting Score Range Threshold | Level 3 CSEM |
| Grade 3 | 349 | 4 | 360 | 4 |
| Grade 4 | 449 | 3 | 460 | 4 |
| Grade 5 | 546 | 4 | 560 | 4 |
| Grade 6 | 648 | 4 | 660 | 4 |
| Grade 7 | 744 | 4 | 760 | 4 |
| Grade 8 | 848 | 4 | 860 | 4 |
| High school | 950 | 4 | 960 | 4 |

#### 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 about the quality of the assessment.

The methodology used for estimating the reliability of classification decisions is described in Livingston and Lewis (1995). The method was implemented by the ETS proprietary computer program RELCLASS-COMP (Version 4.14).

Decision accuracy describes the extent to which students are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the question of how closely 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 describes 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 question of what is the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test.

Decision consistency values are always lower than the corresponding decision accuracy values because in decision consistency, both of the classifications of the student are based on scores that depend on which form of the test the student took. In decision accuracy, only one of the classifications is based on a score that can vary in this way.

In each case, the proportion of classifications with exact agreement is the sum of the entries in the diagonal of the contingency table representing the bivariate distribution.

Reliability of classification at a threshold score is estimated by combining the score reporting ranges above a particular threshold score and combining the score reporting ranges below that threshold. The result is a two-by-two table indicating whether the students reach the threshold score or not. The sum of the entries in the main diagonal is the number of students accurately (or consistently) classified as not reaching versus reaching the threshold score. Table 7.16 and table 7.17 illustrate these two-by-two contingency tables.

Table 7.16 Decision Accuracy for Reaching a Score Reporting Range Threshold

|  |  |  |
| --- | --- | --- |
| Status on the Form Taken | True Status on All Forms Average: Does Not Reach a Reporting Range Threshold | True Status on All Forms Average: Reaches a Reporting Range Threshold |
| Does not reach a reporting range threshold | Correct classification | Incorrect classification |
| Reaches a reporting range threshold | Incorrect classification | Correct classification |

Table 7.17 Decision Consistency for Reaching a Score Reporting Range Threshold

|  |  |  |
| --- | --- | --- |
| Status on the Form Taken | Decision Made on a Single Form: Does Not Reach a Reporting Range Threshold | Decision Made on a Single Form: Reaches a Reporting Range Threshold |
| Does not reach a reporting range threshold | Correct classification | Incorrect classification |
| Reaches a reporting range threshold | Incorrect classification | Correct classification |

Each table includes the contingency tables for both accuracy and consistency of the various reporting range classifications. The proportion of students being accurately classified is determined by summing across the diagonals of the upper tables. The proportion of consistently classified students is determined by summing the diagonals of the lower tables.

The results of these analyses are presented in table 7.E.8 through table 7.E.21 in [appendix 7.E](#_Appendix_7.E:_Reliability). Each table pair includes the contingency tables for both accuracy and consistency of the various reporting range classifications. The proportion of students being accurately classified is determined by summing across the diagonals of the first table of the table pairs for decision accuracy (i.e., table 7.E.8, table 7.E.10, etc.). The proportion of consistently classified students is determined by summing the diagonals of the second table of the table pairs for decision consistency (i.e., table 7.E.9, table 7.E.11, etc.).

The overall decision accuracy is greater than 0.75 for all tests—grades three through eight and the high school grade band—with the highest accuracy of 0.88 occurring for grade three and the lowest level of accuracy of 0.79 occurring in grade seven. The overall decision consistency is relatively lower, with the lowest consistency of 0.72 occurring for grade seven and the highest consistency of 0.83 occurring for grade three.

### 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). It constitutes the central notion underlying the development, administration, and scoring of tests, and the uses and interpretations of test scores.

The 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, 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 CSA. This section is organized primarily around the principles prescribed by AERA, APA, and NCME’s *Standards for Educational and Psychological Testing* (2014). These *Standards* require a clear definition of the purpose of the test, 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 CSA, followed by a description and discussion of different kinds of validity evidence that have been gathered.

#### Evidence in the Design of the CSA

##### Purpose

The CSA is designed to measure a student’s Spanish skills in reading, writing mechanics, and listening for the purposes of

* providing student-level data in Spanish competency,
* providing aggregated data that may be used for evaluating the implementation of Spanish language arts programs at the local level, and
* providing a high school measure suitable to be used, in part, for the California State Seal of Biliteracy.

The assessment provides students an annual opportunity to measure their reading/‌language arts competency in Spanish.

##### The Constructs to Be Measured

As a voluntary assessment to measure a student’s Spanish skills in reading, writing mechanics, and listening, the CSA is designed to show how well students perform relative to the Spanish version of Common Core English language arts/literacy standards (i.e., California Common Core State Standards en Español [CCCSSeE]), which was developed as a joint effort between the San Diego County Office of Education, Council of Chief State School Officers, and the CDE.

The CCCSSeE are organized into the following domains:

* Reading standards
* Writing standards
* Speaking/Listening standards
* Language standards

It should also be noted that while the focus of the CCCSSeE is acquired language arts competency, the domains in the previous list are also harmonious with a four-skill, language-learning framework (e.g., listeningand *reading*, known as “receptive” skills, and speakingandwriting,known as “productive” skills).[[8]](#footnote-9)

Test blueprints are used to measure students’ mastery of the standards included in theCCCSSeE. They also provide an operational definition of the construct to which each set of standards refers and define the following:

* Subject to be assessed
* Tasks to be presented
* Administration instructions to be given
* 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).

ETS developed all CSA items to conform to the SBE-approved test blueprints (CDE, 2017).

##### The Interpretations and Uses of the Scores

Overall student performance expressed as scale scores are generated for the CSA. The scale score is also used to classify students in terms of their score reporting range by grade.

The grade-specific score reporting range descriptors help to communicate the different skills in Spanish reading/language arts knowledge that correspond with one of the three score reporting ranges. The importance of these range descriptors is that they define the knowledge or skill expectations at each range on a functional basis, based on the standards and general achievement level descriptors, and give standardized meaning to scores or score reporting ranges.

A description of the uses and applications of the CSA results is presented in [*Chapter 6*: *Scoring and Reporting*](#_Scoring_and_Reporting).

The CSA results have four primary purposes:

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 local educational agencies identify strengths and areas that need improvement in their educational programs
4. Provide the public and policymakers with information about student achievement

More detailed descriptions regarding score use can be found in the *Education Code* Section 60602 web page on the California Legislative Information website.

##### Intended Test Population

The intended test population for the CSA consists of students enrolled in grades three through twelve receiving instruction in Spanish in California and students seeking a measure that recognizes their Spanish-specific academic reading, writing mechanics, and listening skills. It is critical to recognize the diverse characteristics of the test population for the CSA and the contexts in which the test purpose and use are situated.

#### Evidence Based on the Test 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 tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman, Slattery, Vranek, & Resnick, 2002; Bhola, Impara & Buckendahl, 2003; Martone & Sireci, 2009).

##### Description of the State Standards

The CSA is aligned with the CCCSSeE. The purpose of the CCCSSeE is to guide instruction in a multitude of contexts, including in-class, collaborative activities. The focuses of the CCCSSeE are acquired language arts competency and the necessary knowledge and skills needed to reach the standards in each grade level.

##### Item Specifications

Item specifications describe the characteristics of items that are written to measure each content standard. ETS maintains item specifications for each grade-level CSA. The specifications for the CSA are described in [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1).

##### Assessment Blueprints

The CSA blueprints describe each of the Spanish language arts domains, including reading, writing mechanics, and listening for all grades tested and how that content domain is assessed through the testable standards (CDE, 2017). Each test is described by a single blueprint. The degree to which test forms administered in 2020–2021 meet the blueprint is provided in [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1) and in [appendix 3.A](#_Appendix_3.A:_CSA).

##### Form Assembly Process

Once items are developed and field-tested, ETS selects all CSA items to conform to the SBE-approved CSA content standards and test blueprints. The content standards, blueprints, and test specifications were used as the basis for choosing items for the CSA. Refer to [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1) for information on the test assembly process.

#### Evidence Based on 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 students” (AERA et al., 2014, p. 12). This type of evidence generally includes documentation of activities such as

* systematic observations of test response behavior,
* analysis of student item-response-time data, and
* evaluation of the reasoning processes students employ when solving test items (Embretson, 1983; Messick, 1989).

This type of evidence is used to confirm that the CSA is measuring the cognitive skills that are intended as the objects of measurement, and that students are using these targeted skills to respond to the items.

##### Analysis of Testing Time

Testing time for each administration can be evaluated for consistency by examining the expected response processes for the items presented to students. The length of time it takes students to complete an assessment is collected and analyzed to build a profile describing what a typical testing event looks like for each 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 CSA grade-level assessments are untimed.

Students with no item response and those who did not answer at least 10 items were removed from these analyses. The remaining testing population was partitioned into quartiles based on scale scores. These quartile groupings are not the same as the reporting levels.

Descriptive statistics of the time required to complete the total test are computed for each of the four quartile groups by grade level. Some cases of extremely long testing time may be attributed to a test’s not being closed down properly. Note that the overall raw scores are used to get the raw score intervals, but the total testing time will be based on the whole test, including the field test items taken by a student.

Table 7.D.1 through table 7.D.7 in [appendix 7.D](#_Appendix_7.D:_Response) present total testing time and raw score interval information at each student performance quartile level by grade level. Overall, students who took the general forms at the lowest raw score interval have shorter testing times than students in the other quartile groups. The median total testing time generally increases as the quartile level increases from the lowest raw score interval to the highest raw score interval—that is, students who performed better on the CSA tended to spend more time on the test.

Note that the mean testing time should be interpreted with caution. Some cases of extremely long testing time may be attributed to students with special needs taking longer to complete a test or the test not being closed down properly. Therefore, mean testing times may be misleading.

The medians (50th 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 times is a modest exclusion that leaves some students with very short durations in the results for each of the tests. 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.

#### Evidence Based on Internal Structure

Internal structure evidence evaluates the strength or salience of the major dimensions underlying an assessment. For the CSA, it is assumed that a single construct underlies the total scores obtained on each assessment. Evidence to support this assumption in a typical testing year can be gathered from the results of item analyses, DIF analyses, evaluations of internal consistency, and studies of reliability.

##### Classical Statistics

Polyserial correlations calculated for the items in an assessment show the degree to which the items discriminate between students with low and high scores on an assessment. To the degree that the correlations are high, evidence that the items assess the same construct is provided. As shown in table 7.3, the mean polyserial correlation was between 0.39 and 0.45. The polyserial correlations for the individual items in the CSA are presented in table 7.A.3 through table 7.A.8 in [appendix 7.A](#_Appendix_7.A:_Classical).

Also relevant to the validity of a score interpretation are the ranges of item difficulty for the items on which a test score will be based. The finding—that items have difficulties that span the range of student ability—provides evidence that students at all levels of ability are adequately measured by the items. Information on average item *p*-values is given in table 7.2; the data in table 7.2 indicates that these assessments had average *p*-values ranging from 0.49 to 0.52. Individual item *p*-values are also presented in table 7.A.3 through table 7.A.8.

##### Differential Item Functioning

DIF analyses are used to assess differences in the item performance of groups of students who differ in their demographic characteristics. Refer to section [*7.3 Differential Item Functioning Analyses*](#_Differential_Item_Functioning) for a description of the DIF analyses and table 7.7, where the results of the DIF analyses are reported.

##### Overall Reliability Estimates

The reliability analysis, including reliability at the student level and group level, is described in [*7.7 Reliability Analyses*](#_Reliability_Analyses). The results of reliability analyses on the overall raw score for each CSA form are presented in table 7.14. The results indicate that the reliability estimates for all tests are satisfactory, ranging from 0.83 to 0.87 for the general forms.

##### Student Groups Reliability Estimates

The reliabilities were also examined for various student groups that differ in their demographic characteristics within the student population. The student groups considered were gender, economic status, provision of special services, length of attendance in US schools, whether students received instruction in Spanish, and English language fluency levels. Student groups for accommodated forms were not analyzed because of the extremely small sample size for these forms.

Across student groups, reliability coefficients were generally higher than 0.80, except for some groups of students who received in special education services, groups not receiving instruction in Spanish, EL student groups, or some reclassified fluent English proficient groups. The reliability was also lower because of a lack of variation in performance caused by small group size and homogenous group members. Refer to subsection [*7.7.3 Student Group Reliabilities and Standard Errors of Measurement*](#_Student_Group_Reliabilities) for details.

Reliability estimates and SEM information for the total test scores are reported for each student group in table 7.E.1 through table 7.E.7 in [appendix 7.E](#_Appendix_7.E:_Reliability).

##### Reliability of Performance Classifications

The methodology used for estimating the reliability of classification decisions is described with the decision classification analyses in subsection [*7.7.6 Decision Classification Analyses*](#_Decision_Classification_Analyses). The overall decision accuracy is greater than 0.79 for all seven assessments. The overall decision consistency is relatively lower, with the lowest being 0.72 for grade seven**.** The results of these analyses are presented in table 7.E.8 through table 7.E.21 in [appendix 7.E](#_Appendix_7.E:_Reliability).

### References

Albano A. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36.

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.

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.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J.S. Long (Eds.) *Testing structural equation models* (pp. 136–62)*.* Newbury Park, CA: Sage Publications.

Cai, L. (2017). FlexMIRT®: *Flexible multilevel, multidimensional item analysis and test scoring* (Version 3.51) [computer software]. Chapel Hill, NC: Vector Psychometric Group.

California Department of Education. (2017). *California Spanish Assessment blueprint.* Sacramento, CA: California Department of Education.

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. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. *ETS Research Report Series*, i–38.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–‍66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–65). Hillsdale, NH: Lawrence Erlbaum Associates, Inc.

Dorans, N. J, Moses, T. P., & Eignor, D. R., (2010). *Principles and practices of test score equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: ETS.

Drasgow F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 69–74). New York: Wiley.

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

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–97.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–46). New York: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). *An alternate definition of the ETS delta scale of stem difficulty* (Research Report 85–43). Princeton, NJ: Educational Testing Service.

Huynh, H. (2000, June). Guidelines for Rasch Linking for PACT. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Reference Testing: Practice analysis to Score Reporting.* Maple Grove, MN: JAM Press.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and Praeger.

Kim, S., von Davier, A. A., & Haberman, S. (2006). *An alternate to equating with small samples in the non-equivalent groups anchor test design* (Research Report 06–27). Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement*, *9,* 209–23.

Kolen, M. J., & Brennan, R. L. (2004a). *CIPE*(Version 1) [Computer software and manual] (CASMA Monograph No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004b). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32,* 179–97.

Lord, M. L. (1984). *Standard error of measures at different ability levels* (Research Report, 84–08). Princeton, NJ: Educational Testing Service.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58,* 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute,* *22,* 719–48.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research, 4,* 1332–61.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–‍74.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

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

Olsson, U. (1979) Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443–60*.*

Parshall, C. G., Du Bose, P., Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32,* 37–54.

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.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21,* 365–89.

von Davier, A. A. (2003). *A note on linear equating methods for the nonequivalent-groups design.* (ETS PSRTR-89-83, ETS RR-03-24). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321–44.

### Accessibility Information

#### Alternative Text for Equation 7.1

*p*-value sub dich equals the fraction with the numerator the sum of X sub ic and the denominator N sub I end fraction.

#### Alternative Text for Equation 7.2

*p*-value sub poly equals the fraction with the numerator X sub ij and the denominator N sub i times Max of X sub I end fraction.

#### Alternative Text for Equation 7.3

r sub polyreg equals the fraction Beta sub hat times S tot divided by the square root of Beta sub hat squared times s sub tot squared plus 1.

#### Alternative Text for Equation 7.4

Alpha sub MH equals the numerator open parenthesis the sum sub m of R sub rm times W sub fm divided by N sub tm close parenthesis divided by the denominator open parenthesis the sum sub m of R sub fm times W sub rm divided by N sub tm closed parenthesis.

#### Alternative Text for Equation 7.5

MH D - DIF equals negative 2.35 times the natural logarithm open bracket alpha sub MH close bracket.

#### Alternative Text for Equation 7.6

SMD equals the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub f of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction minus the fraction with numerator the sum from m equals 1 to M of N sub fm times E sub r of Y from X equals m and denominator the sum from m equals 1 to M of N sub fm end fraction equals the fraction with the numerator the sum from m equals 1 to M of D sub fm and the denominator m equals1 to M of N suf fm end fraction.

#### Alternative Text for Equation 7.7

P sub ih of theta sub j equals:

The numerator exp open parenthesis the sum from v equals 1 to h of Da sub i open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis divided by the denominator open parenthesis 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c of Da sub I open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis close parenthesis, if score h equals 1, 2, …, n sub i.

P sub ih of theta sub j equals:

1 divided by the denominator open parenthesis 1 plus the sum from c equals 1 to n sub I exp open parenthesis the sum from v equals 1 to c of Da sub I open parenthesis theta sub j minus b sub I plus d sub iv close parenthesis close parenthesis close parenthesis, if score h equals 0.

#### Alternative Text for Equation 7.8

Z equals the numerator open absolute symbol, D subtracts Md sub D, close absolute symbol, divided by the denominator of 0.74 times IQR.

#### Alternative Text for Equation 7.9

Epsilon of theta equals the sum from i equals 1 to ndich of P sub i of theta plus the sum from j equals 1 to npoly times the sum of x equals 1 to m of s sub xj times P sub xj of theta.

#### Alternative Text for Equation 7.10

Scale score equals intercept a plus slope times theta.

#### Alternative Text for Equation 7.11

Slope *b* equals sigma sub scale score divided by sigma sub theta.

#### Alternative Text for Equation 7.12

Intercept *a* equals scale score at threshold 3 subtract sigma sub scale score divided by sigma sub theta times theta sub threshold 3.

#### Alternative Text for Equation 7.13

Scale score equals scale score at threshold 3 plus sigma sub scale score divided by sigma sub theta times open bracket theta minus theta sub threshold 3 close bracket.

#### Alternative Text for Equation 7.14

x subtract mu of x over sigma of x equals y subtracts mu of y over sigma of y.

#### Alternative Text for Equation 7.15

Ly of x equals y equals sigma of y times open bracket x subtracts mu of x over sigma of x close bracket pluses mu of y.

#### Alternative Text for Equation 7.16

T equals w times P plus open bracket 1 minus Q close bracket times Q.

#### Alternative Text for Equation 7.17

Alpha equals fraction with numerator K and denominator K minus 1 end fraction times open bracket 1 minus fraction with numerator sum from I equals 1 to K of S squared sub X sub I and denominator S squared sub X close bracket.

#### Alternative Text for Equation 7.18

Rho sub strata equals 1 minus fraction with numerator sum of sigma squared sub X sub j times open bracket 1 minus alpha sub j close bracket and denominator sigma squared sub X.

#### Alternative Text for Equation 7.19

SEM equals S sub X times square root of 1 minus rho sub strata.

#### Alternative Text for Equation 7.20

SEM of Theta sub j equals 1 divided by the square root of I of theta sub j.

#### Alternative Text for Equation 7.21

I of theta sub j equals the sum from I equals 1 to n of I sub I of theta sub j.

#### Alternative Text for Equation 7.22

I sub i of theta sub j equals open bracket s sub i2 of theta sub j minus s sub i squared of theta sub j.

#### Alternative Text for Equation 7.23

S sub i of Theta sub j equals the sum from h equals zero to n of h times p sub ih of theta sub j.

#### Alternative Text for Equation 7.24

S sub i2 of Theta sub j equals the sum from h equals zero to n sub i of h squared times p sub ih of theta sub j.

#### Alternative Text for Equation 7.25

CSEM of SS equals 1 divided by the square root of I of theta sub j times b.

### Appendix 7.A: Classical Item Analyses

**Note:** Item information for high school is not available given that item analysis for high school was not conducted because of the extremely small test sample.

Table 7.A.1 Item Difficulty Distributions by Domain

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level | Domain | 0 ≤ *p* < 0.2 | 0.2 ≤ *p* < 0.4 | 0.4 ≤ *p* < 0.6 | 0.6 ≤ *p* < 0.8 | 0.8 ≤ *p* ≤ 1.0 | Total Number of Items | Mean *p-*‍value | Minimum *p-*value | Maximum *p-*value |
| Grade 3 | Reading | 1 | 14 | 8 | 1 | 0 | 24 | 0.38 | 0.17 | 0.61 |
| Grade 3 | Writing Mechanics | 1 | 7 | 8 | 0 | 0 | 16 | 0.39 | 0.19 | 0.56 |
| Grade 3 | Listening | 0 | 2 | 7 | 3 | 0 | 12 | 0.51 | 0.22 | 0.80 |
| Grade 4 | Reading | 2 | 10 | 12 | 0 | 0 | 24 | 0.40 | 0.17 | 0.60 |
| Grade 4 | Writing Mechanics | 0 | 5 | 9 | 2 | 0 | 16 | 0.46 | 0.24 | 0.70 |
| Grade 4 | Listening | 0 | 3 | 5 | 4 | 0 | 12 | 0.52 | 0.20 | 0.76 |
| Grade 5 | Reading | 3 | 9 | 10 | 2 | 0 | 24 | 0.40 | 0.13 | 0.66 |
| Grade 5 | Writing Mechanics | 1 | 6 | 9 | 0 | 0 | 16 | 0.42 | 0.20 | 0.59 |
| Grade 5 | Listening | 0 | 2 | 3 | 6 | 1 | 12 | 0.59 | 0.35 | 0.83 |
| Grade 6 | Reading | 1 | 3 | 18 | 2 | 0 | 24 | 0.48 | 0.17 | 0.68 |
| Grade 6 | Writing Mechanics | 0 | 4 | 6 | 5 | 1 | 16 | 0.53 | 0.28 | 0.80 |
| Grade 6 | Listening | 0 | 1 | 5 | 6 | 0 | 12 | 0.59 | 0.25 | 0.79 |
| Grade 7 | Reading | 1 | 8 | 14 | 1 | 0 | 24 | 0.44 | 0.04 | 0.65 |
| Grade 7 | Writing Mechanics | 0 | 4 | 8 | 4 | 0 | 16 | 0.49 | 0.30 | 0.68 |
| Grade 7 | Listening | 0 | 2 | 5 | 4 | 1 | 12 | 0.58 | 0.37 | 0.82 |
| Grade 8 | Reading | 0 | 8 | 13 | 3 | 0 | 24 | 0.45 | 0.21 | 0.67 |
| Grade 8 | Writing Mechanics | 0 | 3 | 9 | 4 | 0 | 16 | 0.52 | 0.34 | 0.79 |
| Grade 8 | Listening | 0 | 1 | 5 | 4 | 2 | 12 | 0.59 | 0.27 | 0.86 |

Table 7.A.2 Item-Total Correlation Distributions by Domain

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade Level | Domain | r < 0 | 0 ≤ r < 0.2 | 0.2 ≤ r < 0.3 | 0.3 ≤ r < 0.4 | 0.4 ≤ r < 0.5 | r ≥ 0.5 | Total Number of Items | Mean r | Minimum r | Maximum r |
| Grade 3 | Reading | 0 | 2 | 2 | 1 | 12 | 7 | 24 | 0.44 | 0.16 | 0.76 |
| Grade 3 | Writing Mechanics | 0 | 2 | 1 | 5 | 5 | 3 | 16 | 0.40 | 0.15 | 0.64 |
| Grade 3 | Listening | 0 | 0 | 0 | 4 | 0 | 8 | 12 | 0.50 | 0.30 | 0.65 |
| Grade 4 | Reading | 0 | 1 | 5 | 4 | 6 | 8 | 24 | 0.42 | 0.08 | 0.68 |
| Grade 4 | Writing Mechanics | 0 | 0 | 5 | 3 | 5 | 3 | 16 | 0.40 | 0.25 | 0.64 |
| Grade 4 | Listening | 0 | 0 | 1 | 2 | 4 | 5 | 12 | 0.48 | 0.24 | 0.72 |
| Grade 5 | Reading | 0 | 3 | 1 | 9 | 4 | 7 | 24 | 0.39 | 0.07 | 0.69 |
| Grade 5 | Writing Mechanics | 0 | 1 | 4 | 6 | 3 | 2 | 16 | 0.36 | 0.19 | 0.62 |
| Grade 5 | Listening | 0 | 0 | 1 | 3 | 3 | 5 | 12 | 0.49 | 0.27 | 0.77 |
| Grade 6 | Reading | 0 | 2 | 1 | 4 | 10 | 7 | 24 | 0.44 | 0.11 | 0.77 |
| Grade 6 | Writing Mechanics | 0 | 2 | 0 | 3 | 5 | 6 | 16 | 0.44 | 0.10 | 0.61 |
| Grade 6 | Listening | 0 | 0 | 0 | 3 | 3 | 6 | 12 | 0.48 | 0.31 | 0.62 |
| Grade 7 | Reading | 0 | 3 | 6 | 5 | 3 | 7 | 24 | 0.39 | 0.06 | 0.65 |
| Grade 7 | Writing Mechanics | 0 | 1 | 6 | 3 | 5 | 1 | 16 | 0.35 | 0.15 | 0.58 |
| Grade 7 | Listening | 0 | 0 | 1 | 4 | 4 | 3 | 12 | 0.43 | 0.28 | 0.62 |
| Grade 8 | Reading | 0 | 2 | 3 | 6 | 7 | 6 | 24 | 0.41 | 0.09 | 0.68 |
| Grade 8 | Writing Mechanics | 0 | 0 | 3 | 4 | 5 | 4 | 16 | 0.42 | 0.23 | 0.64 |
| Grade 8 | Listening | 0 | 1 | 3 | 3 | 3 | 2 | 12 | 0.38 | 0.17 | 0.60 |

**Notes:**

What follows are the values that can appear in the *Flag* column in table 7.A.3 through table 7.A.8:

* A = low average item score
* Dichotomous
* Item with two choices ≤ less than 0.50
* Item with three choices ≤ less than 0.30
* Item with four choices ≤ less than 0.20
* Polytomous = < 0.20
* D = proportionally more higher-ability test-takers (top 20%) selected a distractor over the key
* H = high average item score (> 0.95)
* L = underrepresented score category (< 3%)
* O = high percent of omits (> 5%)
* P = positive correlation with the criterion score for one or more of the distractors
* Rpoly = low correlation with criterion (< 0.20)

Table 7.A.3 Item Difficulty and Item-Total Correlation by Item for Grade Three

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.46 | 0.38 | [no flag] | 1 | MC |
| 2 | 0.49 | 0.54 | [no flag] | 1 | MC |
| 3 | 0.54 | 0.37 | [no flag] | 1 | InlineChoices |
| 4 | 0.51 | 0.53 | P | 2 | MC |
| 5 | 0.59 | 0.56 | [no flag] | 2 | InlineChoices |
| 6 | 0.45 | 0.58 | [no flag] | 1 | MC |
| 7 | 0.44 | 0.40 | [no flag] | 1 | MC |
| 8 | 0.32 | 0.51 | [no flag] | 1 | MC |
| 9 | 0.53 | 0.16 | P R | 1 | InlineChoices |
| 10 | 0.44 | 0.45 | [no flag] | 1 | MC |
| 11 | 0.45 | 0.43 | P | 1 | MC |
| 12 | 0.36 | 0.60 | [no flag] | 1 | MC |
| 13 | 0.61 | 0.50 | P | 2 | MC |
| 14 | 0.46 | 0.21 | [no flag] | 1 | MC |
| 15 | 0.33 | 0.43 | [no flag] | 1 | MC |
| 16 | 0.27 | 0.47 | [no flag] | 1 | MC |
| 17 | 0.29 | 0.19 | P R | 1 | MC |
| 18 | 0.19 | 0.35 | A | 1 | MC |
| 19 | 0.32 | 0.15 | P R | 1 | InlineChoices |
| 20 | 0.47 | 0.50 | [no flag] | 1 | MC |

Table 7.A.3 *(continuation one)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 21 | 0.52 | 0.43 | [no flag] | 1 | InlineChoices |
| 22 | 0.28 | 0.64 | [no flag] | 1 | MC |
| 23 | 0.33 | 0.29 | [no flag] | 1 | MC |
| 24 | 0.56 | 0.47 | [no flag] | 2 | Composite |
| 25 | 0.32 | 0.30 | [no flag] | 1 | MC |
| 26 | 0.60 | 0.65 | [no flag] | 2 | InlineChoices |
| 27 | 0.52 | 0.57 | [no flag] | 1 | MC |
| 28 | 0.80 | 0.52 | [no flag] | 2 | Zones |
| 29 | 0.62 | 0.63 | [no flag] | 1 | InlineChoices |
| 30 | 0.22 | 0.35 | P | 1 | MC |
| 31 | 0.50 | 0.30 | P | 2 | MC |
| 32 | 0.31 | 0.35 | [no flag] | 2 | Composite |
| 33 | 0.48 | 0.57 | [no flag] | 1 | MC |
| 34 | 0.50 | 0.46 | [no flag] | 1 | MC |
| 35 | 0.27 | 0.46 | [no flag] | 1 | MC |
| 36 | 0.29 | 0.43 | [no flag] | 1 | MC |
| 37 | 0.37 | 0.38 | [no flag] | 1 | MC |
| 38 | 0.42 | 0.16 | P R | 1 | MC |
| 39 | 0.20 | 0.47 | P | 2 | Match |
| 40 | 0.55 | 0.42 | [no flag] | 1 | MC |
| 41 | 0.50 | 0.32 | [no flag] | 1 | InlineChoices |
| 42 | 0.45 | 0.23 | [no flag] | 1 | MC |
| 43 | 0.33 | 0.45 | [no flag] | 1 | MC |
| 44 | 0.32 | 0.50 | [no flag] | 1 | MC |
| 45 | 0.17 | 0.49 | A P | 1 | Match |
| 46 | 0.32 | 0.53 | [no flag] | 1 | Match |
| 47 | 0.39 | 0.40 | D P | 2 | MC |
| 48 | 0.22 | 0.50 | [no flag] | 1 | MC |
| 49 | 0.39 | 0.44 | [no flag] | 1 | InlineChoices |
| 50 | 0.56 | 0.50 | [no flag] | 1 | MC |
| 51 | 0.40 | 0.76 | [no flag] | 1 | MC |
| 52 | 0.34 | 0.62 | [no flag] | 1 | MC |

Table 7.A.4 Item Difficulty and Item-Total Correlation by Item for Grade Four

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.40 | 0.43 | [no flag] | 1 | MC |
| 2 | 0.58 | 0.61 | [no flag] | 1 | MC |
| 3 | 0.76 | 0.43 | [no flag] | 1 | InlineChoices |
| 4 | 0.54 | 0.33 | D P | 2 | Match |
| 5 | 0.40 | 0.46 | [no flag] | 1 | MC |
| 6 | 0.20 | 0.24 | D P | 1 | MC |
| 7 | 0.42 | 0.34 | [no flag] | 1 | MC |
| 8 | 0.27 | 0.42 | P | 1 | Grid |
| 9 | 0.39 | 0.46 | [no flag] | 1 | MC |
| 10 | 0.52 | 0.37 | D P | 2 | InlineChoices |
| 11 | 0.63 | 0.42 | [no flag] | 1 | InlineChoices |
| 12 | 0.55 | 0.34 | [no flag] | 1 | MC |
| 13 | 0.48 | 0.25 | [no flag] | 2 | InlineChoices |
| 14 | 0.70 | 0.64 | [no flag] | 1 | InlineChoices |
| 15 | 0.34 | 0.53 | [no flag] | 1 | MC |
| 16 | 0.55 | 0.30 | P | 2 | MC |
| 17 | 0.36 | 0.22 | P | 1 | MC |
| 18 | 0.59 | 0.29 | P | 2 | MC |
| 19 | 0.24 | 0.43 | [no flag] | 1 | MC |
| 20 | 0.30 | 0.37 | P | 1 | InlineChoices |
| 21 | 0.54 | 0.23 | P | 2 | MC |
| 22 | 0.45 | 0.36 | [no flag] | 1 | MC |
| 23 | 0.28 | 0.38 | [no flag] | 1 | MC |
| 24 | 0.20 | 0.08 | D P R | 1 | InlineChoices |
| 25 | 0.59 | 0.62 | [no flag] | 1 | MC |
| 26 | 0.51 | 0.49 | [no flag] | 1 | MC |
| 27 | 0.34 | 0.26 | [no flag] | 1 | MC |
| 28 | 0.24 | 0.26 | D P | 1 | MC |
| 29 | 0.42 | 0.39 | [no flag] | 1 | MC |
| 30 | 0.49 | 0.60 | [no flag] | 1 | MC |
| 31 | 0.29 | 0.43 | A | 1 | Zones |
| 32 | 0.73 | 0.49 | [no flag] | 2 | Match |
| 33 | 0.73 | 0.55 | L | 2 | Zones |
| 34 | 0.42 | 0.60 | [no flag] | 1 | MC |
| 35 | 0.72 | 0.72 | [no flag] | 1 | MC |
| 36 | 0.47 | 0.26 | [no flag] | 1 | MC |
| 37 | 0.35 | 0.62 | [no flag] | 2 | Composite |
| 38 | 0.17 | 0.42 | A D | 1 | InlineChoices |
| 39 | 0.31 | 0.67 | [no flag] | 1 | MC |

Table 7.A.4 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 40 | 0.44 | 0.68 | [no flag] | 1 | MC |
| 41 | 0.55 | 0.28 | P | 2 | MC |
| 42 | 0.33 | 0.54 | [no flag] | 1 | InlineChoices |
| 43 | 0.35 | 0.33 | [no flag] | 1 | MC |
| 44 | 0.43 | 0.23 | P | 2 | MC |
| 45 | 0.44 | 0.41 | P | 1 | MC |
| 46 | 0.60 | 0.58 | [no flag] | 1 | MC |
| 47 | 0.54 | 0.64 | [no flag] | 1 | MC |
| 48 | 0.19 | 0.59 | A P | 1 | MC |
| 49 | 0.52 | 0.42 | [no flag] | 1 | MC |
| 50 | 0.49 | 0.44 | [no flag] | 1 | MC |
| 51 | 0.39 | 0.42 | [no flag] | 1 | MC |
| 52 | 0.59 | 0.58 | [no flag] | 1 | MC |

Table 7.A.5 Item Difficulty and Item-Total Correlation by Item for Grade Five

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.45 | 0.35 | D P | 2 | InlineChoices |
| 2 | 0.59 | 0.32 | [no flag] | 1 | InlineChoices |
| 3 | 0.68 | 0.54 | [no flag] | 1 | InlineChoices |
| 4 | 0.65 | 0.65 | [no flag] | 1 | MC |
| 5 | 0.83 | 0.27 | [no flag] | 1 | Zones |
| 6 | 0.68 | 0.63 | [no flag] | 1 | MC |
| 7 | 0.66 | 0.49 | P | 2 | MC |
| 8 | 0.35 | 0.31 | [no flag] | 1 | InlineChoices |
| 9 | 0.54 | 0.26 | [no flag] | 1 | InlineChoices |
| 10 | 0.27 | 0.51 | [no flag] | 1 | MC |
| 11 | 0.51 | 0.39 | P | 2 | MC |
| 12 | 0.44 | 0.47 | [no flag] | 1 | MC |
| 13 | 0.43 | 0.51 | [no flag] | 1 | MC |
| 14 | 0.17 | 0.31 | A P | 1 | MC |
| 15 | 0.35 | 0.27 | [no flag] | 1 | InlineChoices |
| 16 | 0.41 | 0.30 | P | 1 | MC |
| 17 | 0.20 | 0.34 | A D | 1 | MC |
| 18 | 0.49 | 0.28 | P | 2 | InlineChoices |
| 19 | 0.45 | 0.37 | P | 2 | InlineChoices |
| 20 | 0.38 | 0.19 | P R | 1 | MC |
| 21 | 0.33 | 0.41 | [no flag] | 1 | MC |
| 22 | 0.36 | 0.27 | [no flag] | 1 | MC |
| 23 | 0.49 | 0.40 | [no flag] | 1 | MC |
| 24 | 0.51 | 0.25 | [no flag] | 1 | MC |
| 25 | 0.35 | 0.53 | P | 2 | MC |
| 26 | 0.36 | 0.38 | P | 1 | MC |
| 27 | 0.49 | 0.47 | [no flag] | 1 | MC |
| 28 | 0.49 | 0.62 | [no flag] | 1 | MC |
| 29 | 0.27 | 0.35 | [no flag] | 1 | MC |
| 30 | 0.41 | 0.42 | D P | 2 | InlineChoices |
| 31 | 0.58 | 0.49 | [no flag] | 1 | MC |
| 32 | 0.59 | 0.30 | P | 2 | MC |
| 33 | 0.66 | 0.77 | [no flag] | 1 | MC |
| 34 | 0.61 | 0.66 | [no flag] | 1 | MC |
| 35 | 0.37 | 0.34 | [no flag] | 1 | MC |
| 36 | 0.31 | 0.36 | [no flag] | 1 | MC |
| 37 | 0.39 | 0.36 | [no flag] | 1 | MC |
| 38 | 0.13 | 0.15 | A D P R | 1 | MC |

Table 7.A.5 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 39 | 0.37 | 0.50 | P | 1 | MC |
| 40 | 0.48 | 0.36 | [no flag] | 1 | Match |
| 41 | 0.58 | 0.40 | [no flag] | 1 | InlineChoices |
| 42 | 0.28 | 0.37 | [no flag] | 1 | MC |
| 43 | 0.53 | 0.43 | [no flag] | 1 | InlineChoices |
| 44 | 0.66 | 0.07 | R | 1 | Match |
| 45 | 0.32 | 0.10 | D P R | 2 | InlineChoices |
| 46 | 0.38 | 0.51 | [no flag] | 1 | MC |
| 47 | 0.45 | 0.57 | [no flag] | 1 | MC |
| 48 | 0.37 | 0.50 | [no flag] | 1 | MC |
| 49 | 0.43 | 0.54 | [no flag] | 1 | MC |
| 50 | 0.60 | 0.69 | P | 2 | MC |
| 51 | 0.18 | 0.34 | A | 1 | MC |
| 52 | 0.48 | 0.44 | [no flag] | 1 | MC |

Table 7.A.6 Item Difficulty and Item-Total Correlation by Item for Grade Six

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.44 | 0.19 | R | 1 | InlineChoices |
| 2 | 0.49 | 0.46 | [no flag] | 1 | MC |
| 3 | 0.48 | 0.53 | [no flag] | 2 | Composite |
| 4 | 0.68 | 0.58 | [no flag] | 1 | MC |
| 5 | 0.78 | 0.56 | [no flag] | 1 | Zones |
| 6 | 0.76 | 0.45 | [no flag] | 1 | MC |
| 7 | 0.79 | 0.32 | L | 2 | MC |
| 8 | 0.57 | 0.42 | [no flag] | 1 | MC |
| 9 | 0.17 | 0.38 | A P | 1 | InlineChoices |
| 10 | 0.42 | 0.53 | P | 2 | Match |
| 11 | 0.50 | 0.21 | [no flag] | 1 | MC |
| 12 | 0.65 | 0.46 | P | 2 | MC |
| 13 | 0.71 | 0.45 | [no flag] | 1 | InlineChoices |
| 14 | 0.54 | 0.35 | [no flag] | 1 | MC |
| 15 | 0.80 | 0.54 | [no flag] | 2 | InlineChoices |
| 16 | 0.38 | 0.61 | [no flag] | 1 | MC |
| 17 | 0.34 | 0.40 | P | 1 | MC |
| 18 | 0.49 | 0.59 | [no flag] | 1 | MC |
| 19 | 0.74 | 0.40 | [no flag] | 2 | InlineChoices |
| 20 | 0.65 | 0.62 | [no flag] | 1 | MC |
| 21 | 0.50 | 0.52 | [no flag] | 1 | MC |
| 22 | 0.46 | 0.52 | [no flag] | 1 | MC |
| 23 | 0.48 | 0.46 | [no flag] | 1 | MC |
| 24 | 0.25 | 0.31 | P | 1 | MC |
| 25 | 0.77 | 0.59 | [no flag] | 1 | MC |
| 26 | 0.43 | 0.31 | [no flag] | 1 | MC |
| 27 | 0.29 | 0.10 | P R | 1 | MC |
| 28 | 0.69 | 0.59 | [no flag] | 1 | MC |
| 29 | 0.46 | 0.50 | [no flag] | 1 | MC |
| 30 | 0.65 | 0.38 | P | 2 | InlineChoices |
| 31 | 0.28 | 0.49 | [no flag] | 1 | MC |
| 32 | 0.51 | 0.56 | [no flag] | 1 | MC |
| 33 | 0.52 | 0.54 | [no flag] | 1 | MC |
| 34 | 0.44 | 0.42 | [no flag] | 1 | Match |
| 35 | 0.68 | 0.49 | [no flag] | 2 | MC |
| 36 | 0.58 | 0.57 | P | 2 | MC |
| 37 | 0.47 | 0.57 | [no flag] | 1 | InlineChoices |
| 38 | 0.61 | 0.67 | [no flag] | 1 | MC |

Table 7.A.6 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 39 | 0.51 | 0.36 | P | 2 | MC |
| 40 | 0.31 | 0.61 | [no flag] | 1 | MC |
| 41 | 0.54 | 0.44 | [no flag] | 1 | MC |
| 42 | 0.53 | 0.45 | [no flag] | 1 | MC |
| 43 | 0.52 | 0.31 | [no flag] | 1 | MC |
| 44 | 0.57 | 0.77 | [no flag] | 1 | MC |
| 45 | 0.48 | 0.11 | P R | 2 | MC |
| 46 | 0.44 | 0.15 | D P R | 2 | InlineChoices |
| 47 | 0.44 | 0.42 | [no flag] | 1 | MC |
| 48 | 0.42 | 0.37 | [no flag] | 1 | MC |
| 49 | 0.53 | 0.48 | [no flag] | 1 | MC |
| 50 | 0.38 | 0.43 | [no flag] | 1 | MC |
| 51 | 0.45 | 0.45 | [no flag] | 2 | Composite |
| 52 | 0.38 | 0.46 | [no flag] | 1 | MC |

Table 7.A.7 Item Difficulty and Item-Total Correlation by Item for Grade Seven

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.60 | 0.15 | R | 1 | MC |
| 2 | 0.37 | 0.36 | P | 2 | InlineChoices |
| 3 | 0.58 | 0.38 | P | 2 | MC |
| 4 | 0.82 | 0.62 | [no flag] | 1 | MC |
| 5 | 0.56 | 0.46 | [no flag] | 1 | MC |
| 6 | 0.40 | 0.28 | [no flag] | 1 | MC |
| 7 | 0.67 | 0.37 | [no flag] | 1 | MC |
| 8 | 0.36 | 0.15 | R | 1 | InlineChoices |
| 9 | 0.51 | 0.64 | [no flag] | 1 | Grid |
| 10 | 0.38 | 0.25 | P | 1 | MC |
| 11 | 0.52 | 0.22 | P | 2 | MC |
| 12 | 0.45 | 0.37 | [no flag] | 1 | MC |
| 13 | 0.32 | 0.36 | P | 1 | MC |
| 14 | 0.37 | 0.29 | P | 1 | MC |
| 15 | 0.64 | 0.47 | P | 2 | MC |
| 16 | 0.57 | 0.36 | P | 2 | MC |
| 17 | 0.57 | 0.43 | [no flag] | 1 | InlineChoices |
| 18 | 0.62 | 0.58 | [no flag] | 1 | MC |
| 19 | 0.39 | 0.25 | P | 1 | MC |
| 20 | 0.55 | 0.45 | P | 1 | MC |
| 21 | 0.33 | 0.26 | P | 1 | MC |
| 22 | 0.45 | 0.48 | [no flag] | 1 | MC |
| 23 | 0.53 | 0.42 | [no flag] | 1 | MC |
| 24 | 0.48 | 0.44 | [no flag] | 1 | MC |
| 25 | 0.57 | 0.35 | [no flag] | 1 | MC |
| 26 | 0.61 | 0.55 | [no flag] | 1 | MC |
| 27 | 0.68 | 0.44 | [no flag] | 2 | MC |
| 28 | 0.69 | 0.54 | [no flag] | 2 | MC |
| 29 | 0.30 | 0.21 | P | 1 | MC |
| 30 | 0.38 | 0.23 | D P | 2 | InlineChoices |
| 31 | 0.52 | 0.46 | [no flag] | 1 | MC |
| 32 | 0.43 | 0.26 | P | 1 | MC |
| 33 | 0.42 | 0.22 | [no flag] | 1 | MC |
| 34 | 0.43 | 0.38 | [no flag] | 1 | MC |
| 35 | 0.68 | 0.40 | [no flag] | 2 | InlineChoices |
| 36 | 0.30 | 0.60 | [no flag] | 1 | MC |
| 37 | 0.47 | 0.52 | [no flag] | 1 | MC |
| 38 | 0.38 | 0.21 | P | 2 | MC |

Table 7.A.7 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 39 | 0.45 | 0.47 | [no flag] | 1 | MC |
| 40 | 0.56 | 0.16 | R | 1 | InlineChoices |
| 41 | 0.40 | 0.22 | P | 1 | MC |
| 42 | 0.04 | 0.39 | A D P | 1 | Grid |
| 43 | 0.32 | 0.44 | P | 1 | MC |
| 44 | 0.45 | 0.64 | [no flag] | 1 | MC |
| 45 | 0.65 | 0.62 | [no flag] | 2 | InlineChoices |
| 46 | 0.55 | 0.65 | [no flag] | 1 | Match |
| 47 | 0.53 | 0.63 | [no flag] | 1 | MC |
| 48 | 0.51 | 0.37 | [no flag] | 1 | MC |
| 49 | 0.47 | 0.24 | P | 2 | MC |
| 50 | 0.53 | 0.34 | P | 2 | MC |
| 51 | 0.58 | 0.44 | [no flag] | 1 | MC |
| 52 | 0.35 | 0.06 | D P R | 1 | InlineChoices |

Table 7.A.8 Item Difficulty and Item-Total Correlation by Item for Grade Eight

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 1 | 0.79 | 0.46 | [no flag] | 1 | MC |
| 2 | 0.65 | 0.56 | [no flag] | 2 | MC |
| 3 | 0.55 | 0.33 | [no flag] | 1 | MC |
| 4 | 0.84 | 0.60 | [no flag] | 1 | MC |
| 5 | 0.68 | 0.24 | L P | 2 | MC |
| 6 | 0.62 | 0.38 | [no flag] | 1 | MC |
| 7 | 0.53 | 0.50 | [no flag] | 1 | MC |
| 8 | 0.48 | 0.45 | [no flag] | 1 | MC |
| 9 | 0.56 | 0.32 | [no flag] | 1 | MC |
| 10 | 0.67 | 0.43 | P | 2 | MC |
| 11 | 0.50 | 0.30 | [no flag] | 1 | MC |
| 12 | 0.45 | 0.30 | [no flag] | 1 | MC |
| 13 | 0.67 | 0.36 | [no flag] | 2 | InlineChoices |
| 14 | 0.44 | 0.33 | [no flag] | 1 | MC |
| 15 | 0.52 | 0.24 | P | 1 | MC |
| 16 | 0.21 | 0.33 | [no flag] | 1 | MC |
| 17 | 0.27 | 0.18 | P R | 1 | MC |
| 18 | 0.39 | 0.30 | [no flag] | 1 | MC |
| 19 | 0.56 | 0.27 | P | 2 | MC |
| 20 | 0.27 | 0.09 | A D P R | 1 | InlineChoices |
| 21 | 0.46 | 0.26 | D P | 2 | InlineChoices |
| 22 | 0.86 | 0.47 | L | 2 | MC |
| 23 | 0.55 | 0.17 | D P R | 2 | InlineChoices |
| 24 | 0.65 | 0.36 | [no flag] | 1 | MC |
| 25 | 0.27 | 0.28 | P | 1 | MC |
| 26 | 0.46 | 0.22 | [no flag] | 1 | MC |
| 27 | 0.52 | 0.56 | [no flag] | 1 | MC |
| 28 | 0.38 | 0.54 | [no flag] | 1 | MC |
| 29 | 0.40 | 0.41 | P | 1 | Match |
| 30 | 0.39 | 0.23 | [no flag] | 1 | MC |
| 31 | 0.69 | 0.62 | [no flag] | 1 | Match |
| 32 | 0.34 | 0.43 | P | 1 | MC |
| 33 | 0.58 | 0.49 | P | 2 | MC |
| 34 | 0.45 | 0.64 | [no flag] | 1 | MC |
| 35 | 0.47 | 0.62 | [no flag] | 1 | MC |
| 36 | 0.51 | 0.61 | [no flag] | 1 | MC |
| 37 | 0.50 | 0.39 | [no flag] | 1 | MC |
| 38 | 0.67 | 0.43 | L P | 2 | MC |

Table 7.A.8 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item Sequence | *p*-value | r | Flag | Maximum Score Points | Item Type |
| 39 | 0.53 | 0.68 | [no flag] | 1 | MC |
| 40 | 0.38 | 0.41 | [no flag] | 1 | MC |
| 41 | 0.60 | 0.64 | [no flag] | 1 | InlineChoices |
| 42 | 0.56 | 0.45 | [no flag] | 1 | MC |
| 43 | 0.45 | 0.34 | P | 2 | InlineChoices |
| 44 | 0.50 | 0.39 | [no flag] | 2 | Composite |
| 45 | 0.28 | 0.47 | P | 1 | Grid |
| 46 | 0.63 | 0.50 | [no flag] | 1 | InlineChoices |
| 47 | 0.55 | 0.35 | P | 2 | MC |
| 48 | 0.35 | 0.65 | [no flag] | 1 | MC |
| 49 | 0.46 | 0.48 | [no flag] | 1 | MC |
| 50 | 0.48 | 0.43 | [no flag] | 1 | MC |
| 51 | 0.47 | 0.40 | [no flag] | 1 | MC |
| 52 | 0.35 | 0.22 | P | 1 | MC |

### Appendix 7.B: Item Response Theory Analyses Results

**Note:** Because of small test sample, no IRT analysis was conducted for high school. Therefore, there is no IRT item difficulty for high school.

Table 7.B.1 IRT Item Difficulty for Grade Three

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 1 | -0.1139 | 0.0722 | N/A | N/A |
| 2 | -0.1703 | 0.0703 | N/A | N/A |
| 3 | -0.4296 | 0.0718 | N/A | N/A |
| 4 | -0.2819 | 0.0613 | 1.1459, -1.1459 | 0.0611, 0.0611 |
| 5 | -0.6161 | 0.0542 | 0.3369, -0.3369 | 0.0638, 0.0638 |
| 6 | -0.0127 | 0.0749 | N/A | N/A |
| 7 | -0.0088 | 0.0706 | N/A | N/A |
| 8 | 0.6241 | 0.0789 | N/A | N/A |
| 9 | -0.4214 | 0.0715 | N/A | N/A |
| 10 | 0.0028 | 0.0719 | N/A | N/A |
| 11 | -0.0030 | 0.0705 | N/A | N/A |
| 12 | 0.4919 | 0.0770 | N/A | N/A |
| 13 | -0.7822 | 0.0582 | 0.9462, -0.9462 | 0.0632, 0.0632 |
| 14 | -0.0758 | 0.0713 | N/A | N/A |
| 15 | 0.5478 | 0.0765 | N/A | N/A |
| 16 | 0.8255 | 0.0796 | N/A | N/A |
| 17 | 0.7546 | 0.0778 | N/A | N/A |
| 18 | 1.2745 | 0.0842 | N/A | N/A |
| 19 | 0.5872 | 0.0776 | N/A | N/A |
| 20 | -0.1367 | 0.0722 | N/A | N/A |
| 21 | -0.3557 | 0.0717 | N/A | N/A |
| 22 | 0.8495 | 0.0837 | N/A | N/A |
| 23 | 0.5784 | 0.0749 | N/A | N/A |
| 24 | -0.5191 | 0.0550 | 0.7089, -0.7089 | 0.0607, 0.0607 |
| 25 | 0.5609 | 0.0763 | N/A | N/A |
| 26 | -0.6488 | 0.0551 | 0.4669, -0.4669 | 0.0642, 0.0642 |
| 27 | -0.3482 | 0.0719 | N/A | N/A |
| 28 | -1.9784 | 0.0927 | 1.0939, -1.0939 | 0.0989, 0.0989 |
| 29 | -0.7443 | 0.0747 | N/A | N/A |
| 30 | 1.1433 | 0.0823 | N/A | N/A |
| 31 | -0.2552 | 0.0669 | 1.7637, -1.7637 | 0.0653, 0.0653 |
| 32 | 0.5290 | 0.0595 | 0.3552, -0.3552 | 0.0701, 0.0701 |
| 33 | -0.1900 | 0.0715 | N/A | N/A |
| 34 | -0.2316 | 0.0730 | N/A | N/A |

Table 7.B.1 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 35 | 0.8272 | 0.0789 | N/A | N/A |
| 36 | 0.7985 | 0.0808 | N/A | N/A |
| 37 | 0.3750 | 0.0754 | N/A | N/A |
| 38 | 0.0280 | 0.0734 | N/A | N/A |
| 39 | 0.9367 | 0.0663 | -0.1756, 0.1756 | 0.0915, 0.0915 |
| 40 | -0.4690 | 0.0723 | N/A | N/A |
| 41 | -0.3177 | 0.0719 | N/A | N/A |
| 42 | -0.0054 | 0.0732 | N/A | N/A |
| 43 | 0.5340 | 0.0770 | N/A | N/A |
| 44 | 0.6579 | 0.0791 | N/A | N/A |
| 45 | 1.5574 | 0.0947 | N/A | N/A |
| 46 | 0.6060 | 0.0812 | N/A | N/A |
| 47 | 0.4380 | 0.0687 | 1.4761, -1.4761 | 0.0692, 0.0692 |
| 48 | 1.1479 | 0.0883 | N/A | N/A |
| 49 | 0.2312 | 0.0736 | N/A | N/A |
| 50 | -0.5079 | 0.0726 | N/A | N/A |
| 51 | 0.2798 | 0.0817 | N/A | N/A |
| 52 | 0.5707 | 0.0805 | N/A | N/A |

Table 7.B.2 IRT Item Difficulty for Grade Four

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 1 | 0.2799 | 0.0774 | N/A | N/A |
| 2 | -0.4855 | 0.0778 | N/A | N/A |
| 3 | -1.4065 | 0.0825 | N/A | N/A |
| 4 | -0.5561 | 0.0768 | 1.9804, -1.9804 | 0.0764, 0.0764 |
| 5 | 0.3165 | 0.0807 | N/A | N/A |
| 6 | 1.3008 | 0.0911 | N/A | N/A |
| 7 | 0.1939 | 0.0766 | N/A | N/A |
| 8 | 0.9590 | 0.0846 | N/A | N/A |
| 9 | 0.2880 | 0.0794 | N/A | N/A |
| 10 | -0.2934 | 0.0624 | 1.1876, -1.1876 | 0.0652, 0.0652 |
| 11 | -0.7075 | 0.0768 | N/A | N/A |
| 12 | -0.3960 | 0.0767 | N/A | N/A |
| 13 | -0.1107 | 0.0659 | 1.3956, -1.3956 | 0.0664, 0.0664 |
| 14 | -1.0456 | 0.0829 | N/A | N/A |
| 15 | 0.6408 | 0.0839 | N/A | N/A |
| 16 | -0.5002 | 0.0698 | 1.6010, -1.6010 | 0.0696, 0.0696 |
| 17 | 0.4735 | 0.0793 | N/A | N/A |
| 18 | -0.6968 | 0.0654 | 1.2643, -1.2643 | 0.0679, 0.0679 |
| 19 | 1.1281 | 0.0888 | N/A | N/A |
| 20 | 0.7471 | 0.0810 | N/A | N/A |
| 21 | -0.4461 | 0.0685 | 1.5991, -1.5991 | 0.0688, 0.0688 |
| 22 | 0.0098 | 0.0750 | N/A | N/A |
| 23 | 0.8804 | 0.0828 | N/A | N/A |
| 24 | 1.3194 | 0.0900 | N/A | N/A |
| 25 | -0.5632 | 0.0802 | N/A | N/A |
| 26 | -0.2499 | 0.0781 | N/A | N/A |
| 27 | 0.5828 | 0.0795 | N/A | N/A |
| 28 | 1.1446 | 0.0875 | N/A | N/A |
| 29 | 0.2069 | 0.0773 | N/A | N/A |
| 30 | -0.1185 | 0.0792 | N/A | N/A |
| 31 | 0.7990 | 0.0848 | N/A | N/A |
| 32 | -1.5413 | 0.0853 | 1.1925, -1.1925 | 0.0893, 0.0893 |
| 33 | -1.7228 | 0.0984 | 1.5628, -1.5628 | 0.1028, 0.1028 |
| 34 | 0.1707 | 0.0796 | N/A | N/A |
| 35 | -1.1628 | 0.0841 | N/A | N/A |
| 36 | -0.0500 | 0.0763 | N/A | N/A |
| 37 | 0.3999 | 0.0635 | 0.1967, -0.1967 | 0.0753, 0.0753 |
| 38 | 1.5867 | 0.0986 | N/A | N/A |

Table 7.B.2 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 39 | 0.7235 | 0.0924 | N/A | N/A |
| 40 | 0.1509 | 0.0849 | N/A | N/A |
| 41 | -0.4568 | 0.0615 | 0.9475, -0.9475 | 0.0657, 0.0657 |
| 42 | 0.6259 | 0.0840 | N/A | N/A |
| 43 | 0.3724 | 0.0795 | N/A | N/A |
| 44 | 0.1722 | 0.0704 | 1.5308, -1.5308 | 0.0701, 0.0701 |
| 45 | 0.0996 | 0.0774 | N/A | N/A |
| 46 | -0.6429 | 0.0774 | N/A | N/A |
| 47 | -0.3198 | 0.0805 | N/A | N/A |
| 48 | 1.3994 | 0.1009 | N/A | N/A |
| 49 | -0.3549 | 0.0762 | N/A | N/A |
| 50 | -0.1664 | 0.0763 | N/A | N/A |
| 51 | 0.2634 | 0.0768 | N/A | N/A |
| 52 | -0.5135 | 0.0803 | N/A | N/A |

Table 7.B.3 IRT Item Difficulty for Grade Five

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d-*parameters Standard Error |
| 1 | 0.0828 | 0.0645 | 0.9081, -0.9081 | 0.0716, 0.0716 |
| 2 | -0.5594 | 0.0810 | N/A | N/A |
| 3 | -0.8799 | 0.0841 | N/A | N/A |
| 4 | -0.7742 | 0.0837 | N/A | N/A |
| 5 | -1.8393 | 0.1005 | N/A | N/A |
| 6 | -0.8959 | 0.0895 | N/A | N/A |
| 7 | -1.4633 | 0.1060 | 1.9223, -1.9223 | 0.1102, 0.1102 |
| 8 | 0.5670 | 0.0823 | N/A | N/A |
| 9 | -0.4161 | 0.0825 | N/A | N/A |
| 10 | 0.9978 | 0.0939 | N/A | N/A |
| 11 | -0.1978 | 0.0661 | 1.3029, -1.3029 | 0.0724, 0.0724 |
| 12 | 0.0774 | 0.0806 | N/A | N/A |
| 13 | 0.1474 | 0.0847 | N/A | N/A |
| 14 | 1.5555 | 0.1028 | N/A | N/A |
| 15 | 0.5279 | 0.0848 | N/A | N/A |
| 16 | 0.2488 | 0.0815 | N/A | N/A |
| 17 | 1.3768 | 0.0977 | N/A | N/A |
| 18 | -0.0080 | 0.0686 | 1.3502, -1.3502 | 0.0714, 0.0714 |
| 19 | 0.1760 | 0.0684 | 1.2018, -1.2018 | 0.0719, 0.0719 |
| 20 | 0.4152 | 0.0823 | N/A | N/A |
| 21 | 0.6217 | 0.0839 | N/A | N/A |
| 22 | 0.4578 | 0.0841 | N/A | N/A |
| 23 | -0.0226 | 0.0806 | N/A | N/A |
| 24 | -0.2312 | 0.0815 | N/A | N/A |
| 25 | 0.6328 | 0.0738 | 0.8820, -0.8820 | 0.0805, 0.0805 |
| 26 | 0.5484 | 0.0846 | N/A | N/A |
| 27 | -0.0973 | 0.0797 | N/A | N/A |
| 28 | -0.0376 | 0.0857 | N/A | N/A |
| 29 | 0.9615 | 0.0885 | N/A | N/A |
| 30 | 0.2754 | 0.0638 | 0.8428, -0.8428 | 0.0732, 0.0732 |
| 31 | -0.4476 | 0.0823 | N/A | N/A |
| 32 | -0.7185 | 0.0720 | 1.3008, -1.3008 | 0.0758, 0.0758 |
| 33 | -0.7766 | 0.0902 | N/A | N/A |
| 34 | -0.5794 | 0.0898 | N/A | N/A |
| 35 | 0.4026 | 0.0809 | N/A | N/A |
| 36 | 0.6766 | 0.0876 | N/A | N/A |
| 37 | 0.3521 | 0.0829 | N/A | N/A |
| 38 | 1.8521 | 0.1130 | N/A | N/A |

Table 7.B.3 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d-*parameters Standard Error |
| 39 | 0.4808 | 0.0843 | N/A | N/A |
| 40 | -0.0417 | 0.0797 | N/A | N/A |
| 41 | -0.5118 | 0.0824 | N/A | N/A |
| 42 | 0.9297 | 0.0890 | N/A | N/A |
| 43 | -0.2909 | 0.0811 | N/A | N/A |
| 44 | -0.8339 | 0.0847 | N/A | N/A |
| 45 | 0.7736 | 0.0771 | 1.0521, -1.0521 | 0.0829, 0.0829 |
| 46 | 0.4483 | 0.0861 | N/A | N/A |
| 47 | 0.1193 | 0.0839 | N/A | N/A |
| 48 | 0.4483 | 0.0860 | N/A | N/A |
| 49 | 0.1244 | 0.0829 | N/A | N/A |
| 50 | -0.5943 | 0.0652 | 0.7517, -0.7517 | 0.0744, 0.0744 |
| 51 | 1.5237 | 0.1029 | N/A | N/A |
| 52 | -0.0317 | 0.0805 | N/A | N/A |

Table 7.B.4 IRT Item Difficulty for Grade Six

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d-*parameters | *d*-parameters Standard Error |
| 1 | 0.2479 | 0.0816 | N/A | N/A |
| 2 | 0.0552 | 0.0829 | N/A | N/A |
| 3 | 0.1424 | 0.0623 | 0.5930, -0.5930 | 0.0692, 0.0692 |
| 4 | -0.8175 | 0.0844 | N/A | N/A |
| 5 | -1.2852 | 0.0913 | N/A | N/A |
| 6 | -1.2105 | 0.0919 | N/A | N/A |
| 7 | -1.8567 | 0.1164 | 1.2589, -1.2589 | 0.1264, 0.1264 |
| 8 | -0.3429 | 0.0810 | N/A | N/A |
| 9 | 1.7741 | 0.1018 | N/A | N/A |
| 10 | 0.3750 | 0.0650 | 0.7746, -0.7746 | 0.0698, 0.0698 |
| 11 | -0.0125 | 0.0803 | N/A | N/A |
| 12 | -0.8312 | 0.0738 | 1.2249, -1.2249 | 0.0775, 0.0775 |
| 13 | -0.9715 | 0.0857 | N/A | N/A |
| 14 | -0.2675 | 0.0810 | N/A | N/A |
| 15 | -1.3359 | 0.0795 | 0.3193, -0.3193 | 0.0937, 0.0937 |
| 16 | 0.6593 | 0.0859 | N/A | N/A |
| 17 | 0.7171 | 0.0835 | N/A | N/A |
| 18 | 0.0644 | 0.0840 | N/A | N/A |
| 19 | -1.0802 | 0.0715 | 0.5852, -0.5852 | 0.0837, 0.0837 |
| 20 | -0.5916 | 0.0855 | N/A | N/A |
| 21 | 0.0918 | 0.0819 | N/A | N/A |
| 22 | 0.1435 | 0.0825 | N/A | N/A |
| 23 | 0.1484 | 0.0801 | N/A | N/A |
| 24 | 1.1731 | 0.0891 | N/A | N/A |
| 25 | -1.2501 | 0.0929 | N/A | N/A |
| 26 | 0.3425 | 0.0811 | N/A | N/A |
| 27 | 0.9847 | 0.0883 | N/A | N/A |
| 28 | -0.9409 | 0.0887 | N/A | N/A |
| 29 | 0.1945 | 0.0822 | N/A | N/A |
| 30 | -0.8038 | 0.0723 | 1.0679, -1.0679 | 0.0781, 0.0781 |
| 31 | 1.0494 | 0.0889 | N/A | N/A |
| 32 | 0.0483 | 0.0833 | N/A | N/A |
| 33 | -0.0918 | 0.0821 | N/A | N/A |
| 34 | 0.2605 | 0.0824 | N/A | N/A |
| 35 | -1.1746 | 0.0882 | 1.3670, -1.3670 | 0.0944, 0.0944 |
| 36 | -0.3345 | 0.0657 | 0.8020, -0.8020 | 0.0696, 0.0696 |
| 37 | 0.2210 | 0.0847 | N/A | N/A |
| 38 | -0.4772 | 0.0862 | N/A | N/A |

Table 7.B.4 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d-*parameters | *d*-parameters Standard Error |
| 39 | -0.0522 | 0.0667 | 1.1426, -1.1426 | 0.0690, 0.0690 |
| 40 | 0.9329 | 0.0894 | N/A | N/A |
| 41 | -0.1357 | 0.0811 | N/A | N/A |
| 42 | -0.1162 | 0.0814 | N/A | N/A |
| 43 | -0.0468 | 0.0797 | N/A | N/A |
| 44 | -0.2920 | 0.0901 | N/A | N/A |
| 45 | 0.1107 | 0.0705 | 1.3392, -1.3392 | 0.0709, 0.0709 |
| 46 | 0.2953 | 0.0681 | 0.9760, -0.9760 | 0.0694, 0.0694 |
| 47 | 0.3391 | 0.0828 | N/A | N/A |
| 48 | 0.3441 | 0.0837 | N/A | N/A |
| 49 | -0.1529 | 0.0821 | N/A | N/A |
| 50 | 0.5115 | 0.0828 | N/A | N/A |
| 51 | 0.1903 | 0.0648 | 0.8508, -0.8508 | 0.0701, 0.0701 |
| 52 | 0.5477 | 0.0848 | N/A | N/A |

Table 7.B.5 IRT Item Difficulty for Grade Seven

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 1 | -0.3769 | 0.0891 | N/A | N/A |
| 2 | 0.7077 | 0.0700 | 0.5954, -0.5954 | 0.0854, 0.0854 |
| 3 | -0.2916 | 0.0774 | 1.3775, -1.3775 | 0.0842, 0.0842 |
| 4 | -1.4441 | 0.1141 | N/A | N/A |
| 5 | -0.1361 | 0.0916 | N/A | N/A |
| 6 | 0.6760 | 0.0915 | N/A | N/A |
| 7 | -0.6684 | 0.0924 | N/A | N/A |
| 8 | 0.7510 | 0.0926 | N/A | N/A |
| 9 | 0.0773 | 0.0931 | N/A | N/A |
| 10 | 0.7335 | 0.0907 | N/A | N/A |
| 11 | -0.0619 | 0.0731 | 1.1873, -1.1873 | 0.0805, 0.0805 |
| 12 | 0.2844 | 0.0885 | N/A | N/A |
| 13 | 0.8960 | 0.0920 | N/A | N/A |
| 14 | 0.7137 | 0.0899 | N/A | N/A |
| 15 | -0.5698 | 0.0805 | 1.2621, -1.2621 | 0.0894, 0.0894 |
| 16 | -0.1719 | 0.0733 | 1.1954, -1.1954 | 0.0818, 0.0818 |
| 17 | -0.2171 | 0.0904 | N/A | N/A |
| 18 | -0.2986 | 0.0962 | N/A | N/A |
| 19 | 0.5841 | 0.0907 | N/A | N/A |
| 20 | -0.0562 | 0.0902 | N/A | N/A |
| 21 | 0.7753 | 0.0951 | N/A | N/A |
| 22 | 0.5136 | 0.0911 | N/A | N/A |
| 23 | -0.0173 | 0.0909 | N/A | N/A |
| 24 | 0.2412 | 0.0878 | N/A | N/A |
| 25 | -0.2344 | 0.0899 | N/A | N/A |
| 26 | -0.3418 | 0.0942 | N/A | N/A |
| 27 | -1.1713 | 0.1132 | 1.6705, -1.6705 | 0.1178, 0.1178 |
| 28 | -0.8855 | 0.0852 | 1.0596, -1.0596 | 0.0962, 0.0962 |
| 29 | 1.0262 | 0.0972 | N/A | N/A |
| 30 | 0.5943 | 0.0725 | 0.7292, -0.7292 | 0.0827, 0.0827 |
| 31 | 0.0334 | 0.0909 | N/A | N/A |
| 32 | 0.4032 | 0.0893 | N/A | N/A |
| 33 | 0.5057 | 0.0908 | N/A | N/A |
| 34 | 0.4415 | 0.0907 | N/A | N/A |
| 35 | -1.0768 | 0.1008 | 1.4532, -1.4532 | 0.1067, 0.1067 |
| 36 | 0.9625 | 0.1000 | N/A | N/A |
| 37 | 0.2321 | 0.0919 | N/A | N/A |
| 38 | 0.7750 | 0.0830 | 1.2944, -1.2944 | 0.0900, 0.0900 |

Table 7.B.5 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b*-parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 39 | 0.3457 | 0.0905 | N/A | N/A |
| 40 | -0.1192 | 0.0890 | N/A | N/A |
| 41 | 0.5449 | 0.0913 | N/A | N/A |
| 42 | 3.6333 | 0.2326 | N/A | N/A |
| 43 | 0.8599 | 0.0972 | N/A | N/A |
| 44 | 0.2984 | 0.0966 | N/A | N/A |
| 45 | -0.4948 | 0.0678 | -0.0896, 0.0896 | 0.0972, 0.0972 |
| 46 | -0.1673 | 0.0955 | N/A | N/A |
| 47 | -0.0598 | 0.0944 | N/A | N/A |
| 48 | 0.0658 | 0.0891 | N/A | N/A |
| 49 | 0.2073 | 0.0718 | 1.1505, -1.1505 | 0.0809, 0.0809 |
| 50 | -0.0203 | 0.0672 | 0.6562, -0.6562 | 0.0801, 0.0801 |
| 51 | -0.3082 | 0.0932 | N/A | N/A |
| 52 | 0.6745 | 0.0928 | N/A | N/A |

Table 7.B.6 IRT Item Difficulty for Grade Eight

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b-*parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 1 | -1.4555 | 0.1206 | N/A | N/A |
| 2 | -0.6899 | 0.0875 | 0.7966, -0.7966 | 0.1003, 0.1003 |
| 3 | -0.1708 | 0.1042 | N/A | N/A |
| 4 | -1.7727 | 0.1381 | N/A | N/A |
| 5 | -1.4720 | 0.1506 | 1.8971, -1.8971 | 0.1566, 0.1566 |
| 6 | -0.4994 | 0.1055 | N/A | N/A |
| 7 | -0.1708 | 0.1003 | N/A | N/A |
| 8 | 0.1193 | 0.1015 | N/A | N/A |
| 9 | -0.2657 | 0.1024 | N/A | N/A |
| 10 | -0.6585 | 0.0807 | 0.5393, -0.5393 | 0.0976, 0.0976 |
| 11 | 0.0293 | 0.1017 | N/A | N/A |
| 12 | 0.2341 | 0.1019 | N/A | N/A |
| 13 | -0.7448 | 0.0845 | 0.6906, -0.6906 | 0.1011, 0.1011 |
| 14 | 0.3432 | 0.1012 | N/A | N/A |
| 15 | 0.0027 | 0.1008 | N/A | N/A |
| 16 | 1.4903 | 0.1230 | N/A | N/A |
| 17 | 1.0850 | 0.1105 | N/A | N/A |
| 18 | 0.5810 | 0.1052 | N/A | N/A |
| 19 | -0.4487 | 0.1008 | 1.8261, -1.8261 | 0.1074, 0.1074 |
| 20 | 1.0918 | 0.1119 | N/A | N/A |
| 21 | 0.2396 | 0.0761 | 0.6553, -0.6553 | 0.0907, 0.0907 |
| 22 | -1.9141 | 0.1507 | 0.8067, -0.8067 | 0.1697, 0.1697 |
| 23 | -0.1944 | 0.0846 | 1.0876, -1.0876 | 0.0921, 0.0921 |
| 24 | -0.5327 | 0.1060 | N/A | N/A |
| 25 | 1.1569 | 0.1168 | N/A | N/A |
| 26 | 0.1646 | 0.1013 | N/A | N/A |
| 27 | -0.0323 | 0.1026 | N/A | N/A |
| 28 | 0.5601 | 0.1100 | N/A | N/A |
| 29 | 0.3854 | 0.1040 | N/A | N/A |
| 30 | 0.5096 | 0.1050 | N/A | N/A |
| 31 | -0.8231 | 0.1183 | N/A | N/A |
| 32 | 0.7504 | 0.1078 | N/A | N/A |
| 33 | -0.3627 | 0.0812 | 0.8426, -0.8426 | 0.0964, 0.0964 |
| 34 | 0.2204 | 0.1104 | N/A | N/A |
| 35 | 0.1127 | 0.1105 | N/A | N/A |
| 36 | -0.0345 | 0.1078 | N/A | N/A |
| 37 | -0.0345 | 0.1014 | N/A | N/A |
| 38 | -1.3206 | 0.1388 | 1.8491, -1.8491 | 0.1422, 0.1422 |

Table 7.B.6 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item Sequence | *b*-parameter | *b-*parameter Standard Error | *d*-parameters | *d*-parameters Standard Error |
| 39 | -0.0908 | 0.1123 | N/A | N/A |
| 40 | 0.4877 | 0.1044 | N/A | N/A |
| 41 | -0.4509 | 0.1140 | N/A | N/A |
| 42 | -0.1393 | 0.1053 | N/A | N/A |
| 43 | 0.3076 | 0.0876 | 1.2188, -1.2188 | 0.0974, 0.0974 |
| 44 | 0.0147 | 0.0827 | 1.0256, -1.0256 | 0.0918, 0.0918 |
| 45 | 0.9909 | 0.1118 | N/A | N/A |
| 46 | -0.4995 | 0.1072 | N/A | N/A |
| 47 | -0.5188 | 0.1074 | 1.9809, -1.9809 | 0.1138, 0.1138 |
| 48 | 0.6946 | 0.1116 | N/A | N/A |
| 49 | 0.1158 | 0.1036 | N/A | N/A |
| 50 | 0.1158 | 0.1076 | N/A | N/A |
| 51 | 0.1564 | 0.1034 | N/A | N/A |
| 52 | 0.6418 | 0.1086 | N/A | N/A |

### Appendix 7.C: Classical Equating Analysis for High School

**Note:** Correlation coefficients in table 7.C.1 represent the correlation between the overall scores and the anchor scores from the same administration.

Table 7.C.1 Summary of High School Forms in 2018–2019 and 2020–2021

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | 2021 Total | 2021 Anchor | 2019 Total | 2019 Anchor |
| Mean | 30.61 | 18.41 | 30.37 | 20.32 |
| SD | 10.24 | 6.14 | 8.72 | 5.97 |
| N | 226 | 226 | 3942 | 3942 |
| Possible Max Score | 62 | 39 | 64 | 39 |
| Correlation | 0.96 | N/A | 0.95 | N/A |

Table 7.C.2 Standard Error of Classical Equating

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Total Score | Mean: Tucker | Mean: Levine | Linear: Tucker | Linear: Levine Observed Score | Linear: Levine True Score | Linear: Chain |
| 0 | 0.24 | 0.23 | 0.49 | 0.58 | 0.61 | 0.59 |
| 1 | 0.24 | 0.23 | 0.47 | 0.56 | 0.59 | 0.57 |
| 2 | 0.24 | 0.23 | 0.46 | 0.55 | 0.57 | 0.56 |
| 3 | 0.24 | 0.23 | 0.45 | 0.53 | 0.55 | 0.54 |
| 4 | 0.24 | 0.23 | 0.43 | 0.52 | 0.54 | 0.52 |
| 5 | 0.24 | 0.23 | 0.42 | 0.50 | 0.52 | 0.51 |
| 6 | 0.24 | 0.23 | 0.41 | 0.48 | 0.50 | 0.49 |
| 7 | 0.24 | 0.23 | 0.39 | 0.47 | 0.48 | 0.48 |
| 8 | 0.24 | 0.23 | 0.38 | 0.45 | 0.47 | 0.46 |
| 9 | 0.24 | 0.23 | 0.37 | 0.44 | 0.45 | 0.44 |
| 10 | 0.24 | 0.23 | 0.35 | 0.42 | 0.43 | 0.43 |
| 11 | 0.24 | 0.23 | 0.34 | 0.41 | 0.42 | 0.41 |
| 12 | 0.24 | 0.23 | 0.33 | 0.39 | 0.40 | 0.40 |
| 13 | 0.24 | 0.23 | 0.31 | 0.38 | 0.38 | 0.38 |
| 14 | 0.24 | 0.23 | 0.30 | 0.36 | 0.37 | 0.36 |
| 15 | 0.24 | 0.23 | 0.29 | 0.35 | 0.35 | 0.35 |
| 16 | 0.24 | 0.23 | 0.28 | 0.33 | 0.33 | 0.33 |
| 17 | 0.24 | 0.23 | 0.27 | 0.32 | 0.32 | 0.32 |
| 18 | 0.24 | 0.23 | 0.25 | 0.30 | 0.30 | 0.30 |
| 19 | 0.24 | 0.23 | 0.24 | 0.29 | 0.29 | 0.29 |
| 20 | 0.24 | 0.23 | 0.23 | 0.28 | 0.27 | 0.28 |
| 21 | 0.24 | 0.23 | 0.22 | 0.26 | 0.26 | 0.26 |
| 22 | 0.24 | 0.23 | 0.21 | 0.25 | 0.25 | 0.25 |
| 23 | 0.24 | 0.23 | 0.20 | 0.24 | 0.23 | 0.24 |
| 24 | 0.24 | 0.23 | 0.20 | 0.23 | 0.22 | 0.22 |
| 25 | 0.24 | 0.23 | 0.19 | 0.22 | 0.21 | 0.21 |

Table 7.C.2 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Total Score | Mean: Tucker | Mean: Levine | Linear: Tucker | Linear: Levine Observed Score | Linear: Levine True Score | Linear: Chain |
| 26 | 0.24 | 0.23 | 0.18 | 0.21 | 0.20 | 0.20 |
| 27 | 0.24 | 0.23 | 0.18 | 0.20 | 0.19 | 0.19 |
| 28 | 0.24 | 0.23 | 0.17 | 0.19 | 0.19 | 0.18 |
| 29 | 0.24 | 0.23 | 0.17 | 0.18 | 0.18 | 0.18 |
| 30 | 0.24 | 0.23 | 0.16 | 0.18 | 0.18 | 0.17 |
| 31 | 0.24 | 0.23 | 0.16 | 0.17 | 0.17 | 0.16 |
| 32 | 0.24 | 0.23 | 0.16 | 0.17 | 0.18 | 0.16 |
| 33 | 0.24 | 0.23 | 0.17 | 0.17 | 0.18 | 0.16 |
| 34 | 0.24 | 0.23 | 0.17 | 0.17 | 0.18 | 0.16 |
| 35 | 0.24 | 0.23 | 0.17 | 0.17 | 0.19 | 0.16 |
| 36 | 0.24 | 0.23 | 0.18 | 0.17 | 0.19 | 0.17 |
| 37 | 0.24 | 0.23 | 0.19 | 0.18 | 0.20 | 0.17 |
| 38 | 0.24 | 0.23 | 0.19 | 0.19 | 0.21 | 0.18 |
| 39 | 0.24 | 0.23 | 0.20 | 0.19 | 0.22 | 0.19 |
| 40 | 0.24 | 0.23 | 0.21 | 0.20 | 0.24 | 0.20 |
| 41 | 0.24 | 0.23 | 0.22 | 0.21 | 0.25 | 0.21 |
| 42 | 0.24 | 0.23 | 0.23 | 0.22 | 0.26 | 0.22 |
| 43 | 0.24 | 0.23 | 0.24 | 0.24 | 0.28 | 0.23 |
| 44 | 0.24 | 0.23 | 0.25 | 0.25 | 0.29 | 0.25 |
| 45 | 0.24 | 0.23 | 0.26 | 0.26 | 0.31 | 0.26 |
| 46 | 0.24 | 0.23 | 0.27 | 0.27 | 0.32 | 0.27 |
| 47 | 0.24 | 0.23 | 0.29 | 0.29 | 0.34 | 0.29 |
| 48 | 0.24 | 0.23 | 0.30 | 0.30 | 0.35 | 0.30 |
| 49 | 0.24 | 0.23 | 0.31 | 0.32 | 0.37 | 0.32 |
| 50 | 0.24 | 0.23 | 0.32 | 0.33 | 0.39 | 0.33 |
| 51 | 0.24 | 0.23 | 0.34 | 0.34 | 0.40 | 0.35 |
| 52 | 0.24 | 0.23 | 0.35 | 0.36 | 0.42 | 0.36 |
| 53 | 0.24 | 0.23 | 0.36 | 0.37 | 0.44 | 0.38 |
| 54 | 0.24 | 0.23 | 0.37 | 0.39 | 0.45 | 0.39 |
| 55 | 0.24 | 0.23 | 0.39 | 0.40 | 0.47 | 0.41 |
| 56 | 0.24 | 0.23 | 0.40 | 0.42 | 0.49 | 0.42 |
| 57 | 0.24 | 0.23 | 0.41 | 0.44 | 0.50 | 0.44 |
| 58 | 0.24 | 0.23 | 0.43 | 0.45 | 0.52 | 0.46 |
| 59 | 0.24 | 0.23 | 0.44 | 0.47 | 0.54 | 0.47 |
| 60 | 0.24 | 0.23 | 0.46 | 0.48 | 0.56 | 0.49 |
| 61 | 0.24 | 0.23 | 0.47 | 0.50 | 0.57 | 0.50 |
| 62 | 0.24 | 0.23 | 0.48 | 0.51 | 0.59 | 0.52 |

Figure 7.C.1, which was created using data derived from table 7.C.2, presents the magnitudes of standard errors of equating by six equating methods, where the x-axis shows the raw scores from 0 to 62 in intervals of 3 and the y-axis shows the corresponding values of standard error from 0 to 0.7 in intervals of 0.1.

The six lines shown in the plot represent the following values:

* The solid straight line represents the standard error of Levine mean equating.
* The dotted straight line represents the standard error of Tucker mean equating.
* The solid line with a V-shaped curve represents the standard error of Tucker linear equating.
* The solid gray line with a V-shaped curve represents the standard error of Levine linear equating.
* The double-dashed line with a V-shaped curve represents the standard error of Levine true score equating.
* The double-dashed, dotted line with a V-shaped curve represents the standard error of chain linear equating.

All V-shaped curves are lower in the middle of raw score ranges from 21 to 42 and higher at the two ends of raw scores from 0 or 62.

The straight lines indicate that the mean equating produces consistent standard errors of equating across all raw score points. The V-shaped lines indicate that the linear equating produces smaller standard error in the middle of raw score range and higher standard errors at the two ends of raw scores. Figure 7.C.1 also indicates that the mean equating produced smaller standard errors, at about 0.25 overall.

Figure 7.C.1 Standard error of classical equating

### Appendix 7.D: Testing Time Analysis

**Notes:**

* Total raw scores were used to partition students into quartiles.
* All students who completed testing with a valid reporting scale score are included.
* Form 1 represents the general form, while Form A represents the accommodated form.

Table 7.D.1 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Three

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–18 | 288 | 56.31 | 42.51 | 9.15 | 281.85 | 10.33 | 18.67 | 26.43 | 44.24 | 74.72 | 110.75 | 250.56 |
| 1 | 19–22 | 272 | 63.06 | 46.35 | 7.90 | 318.88 | 9.80 | 18.18 | 29.23 | 49.85 | 88.29 | 117.55 | 238.65 |
| 1 | 23–31 | 314 | 82.28 | 50.18 | 8.09 | 424.43 | 13.84 | 35.54 | 50.47 | 69.33 | 101.67 | 139.02 | 230.89 |
| 1 | 32–62 | 322 | 95.09 | 44.29 | 25.02 | 359.63 | 32.88 | 51.33 | 66.55 | 85.21 | 114.76 | 145.54 | 243.88 |
| A | 0–15 | 3 | 115.92 | 47.32 | 73.74 | 167.09 | 73.74 | 73.74 | 73.74 | 106.94 | 167.09 | 167.09 | 167.09 |
| A | 16–18 | 5 | 90.40 | 54.32 | 38.80 | 182.69 | 38.80 | 38.80 | 72.72 | 74.22 | 83.57 | 182.69 | 182.69 |
| A | 19–20 | 4 | 107.58 | 37.05 | 72.23 | 149.22 | 72.23 | 72.23 | 76.53 | 104.44 | 138.63 | 149.22 | 149.22 |
| A | 21–60 | 5 | 141.61 | 128.19 | 52.57 | 367.09 | 52.57 | 52.57 | 74.79 | 100.76 | 112.82 | 367.09 | 367.09 |

Table 7.D.2 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Four

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–21 | 256 | 54.28 | 32.04 | 8.67 | 233.32 | 9.99 | 18.50 | 29.49 | 50.16 | 69.48 | 95.11 | 157.31 |
| 1 | 22–25 | 226 | 79.61 | 51.00 | 9.73 | 343.27 | 13.67 | 30.07 | 43.44 | 70.44 | 103.83 | 128.00 | 268.94 |
| 1 | 26–34 | 289 | 90.87 | 50.52 | 14.22 | 310.10 | 19.21 | 39.32 | 55.28 | 79.33 | 113.23 | 159.87 | 274.73 |
| 1 | 35–63 | 269 | 106.55 | 42.98 | 31.63 | 278.57 | 38.23 | 60.24 | 74.79 | 96.56 | 131.06 | 173.76 | 230.61 |
| A | 0–18 | 3 | 88.57 | 23.36 | 64.73 | 111.41 | 64.73 | 64.73 | 64.73 | 89.57 | 111.41 | 111.41 | 111.41 |
| A | 19–22 | 3 | 123.09 | 32.59 | 90.76 | 155.94 | 90.76 | 90.76 | 90.76 | 122.58 | 155.94 | 155.94 | 155.94 |
| A | 23–23 | 1 | 34.77 | N/A | 34.77 | 34.77 | 34.77 | 34.77 | 34.77 | 34.77 | 34.77 | 34.77 | 34.77 |
| A | 24–61 | 5 | 115.70 | 88.03 | 55.40 | 268.89 | 55.40 | 55.40 | 64.44 | 80.76 | 108.99 | 268.89 | 268.89 |

Table 7.D.3 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Five

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–19 | 182 | 65.10 | 44.33 | 5.87 | 268.97 | 8.13 | 20.95 | 32.54 | 55.89 | 88.26 | 120.41 | 239.17 |
| 1 | 20–25 | 256 | 85.69 | 56.99 | 8.77 | 314.22 | 17.46 | 28.93 | 44.41 | 72.14 | 116.98 | 163.10 | 264.66 |
| 1 | 26–32 | 218 | 104.58 | 65.43 | 13.68 | 528.40 | 18.33 | 41.49 | 62.90 | 87.04 | 128.62 | 190.53 | 296.16 |
| 1 | 33–62 | 246 | 109.45 | 51.53 | 28.23 | 303.86 | 34.43 | 58.57 | 73.99 | 96.62 | 131.81 | 178.59 | 299.03 |
| A | 0–19 | 3 | 80.94 | 74.77 | 14.15 | 161.72 | 14.15 | 14.15 | 14.15 | 66.96 | 161.72 | 161.72 | 161.72 |
| A | 20–20 | 4 | 78.51 | 70.28 | 24.11 | 180.41 | 24.11 | 24.11 | 32.84 | 54.75 | 124.17 | 180.41 | 180.41 |
| A | 21–25 | 2 | 66.09 | 8.48 | 60.09 | 72.08 | 60.09 | 60.09 | 60.09 | 66.09 | 72.08 | 72.08 | 72.08 |
| A | 26–63 | 5 | 183.44 | 100.74 | 99.32 | 331.78 | 99.32 | 99.32 | 100.21 | 146.01 | 239.89 | 331.78 | 331.78 |

Table 7.D.4 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Six

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–25 | 232 | 81.15 | 68.11 | 4.09 | 371.42 | 5.72 | 22.25 | 34.25 | 62.23 | 102.64 | 163.45 | 304.27 |
| 1 | 26–32 | 216 | 93.99 | 55.50 | 7.60 | 350.32 | 13.40 | 36.38 | 57.92 | 81.02 | 115.56 | 169.51 | 277.48 |
| 1 | 33–41 | 238 | 123.18 | 77.79 | 27.01 | 503.04 | 37.21 | 50.28 | 72.68 | 104.74 | 144.82 | 215.27 | 416.18 |
| 1 | 42–65 | 251 | 106.72 | 57.60 | 30.65 | 530.49 | 40.71 | 58.35 | 71.87 | 91.21 | 122.39 | 169.67 | 306.47 |
| A | 0–16 | 1 | 70.06 | N/A | 70.06 | 70.06 | 70.06 | 70.06 | 70.06 | 70.06 | 70.06 | 70.06 | 70.06 |
| A | 17–22 | 4 | 100.30 | 48.05 | 39.83 | 156.72 | 39.83 | 39.83 | 67.97 | 102.32 | 132.63 | 156.72 | 156.72 |
| A | 23–25 | 3 | 66.57 | 24.79 | 40.06 | 89.18 | 40.06 | 40.06 | 40.06 | 70.47 | 89.18 | 89.18 | 89.18 |
| A | 26–64 | 3 | 136.42 | 31.21 | 116.21 | 172.36 | 116.21 | 116.21 | 116.21 | 120.69 | 172.36 | 172.36 | 172.36 |

Table 7.D.5 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Seven

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–24 | 175 | 96.60 | 84.73 | 7.16 | 598.44 | 11.01 | 33.24 | 47.45 | 79.59 | 108.36 | 178.68 | 522.88 |
| 1 | 25–30 | 159 | 103.46 | 48.61 | 14.78 | 329.73 | 21.61 | 49.12 | 72.60 | 96.87 | 127.06 | 163.64 | 270.06 |
| 1 | 31–38 | 203 | 118.64 | 55.35 | 40.20 | 415.26 | 44.88 | 63.12 | 82.87 | 107.40 | 137.22 | 192.80 | 294.80 |
| 1 | 39–65 | 180 | 116.89 | 51.85 | 33.41 | 365.44 | 41.66 | 62.67 | 79.22 | 105.50 | 144.23 | 186.26 | 272.69 |
| A | 0–19 | 3 | 115.36 | 23.25 | 99.60 | 142.06 | 99.60 | 99.60 | 99.60 | 104.41 | 142.06 | 142.06 | 142.06 |
| A | 20–24 | 2 | 136.83 | 30.63 | 115.17 | 158.49 | 115.17 | 115.17 | 115.17 | 136.83 | 158.49 | 158.49 | 158.49 |
| A | 25–25 | 3 | 102.96 | 46.04 | 67.99 | 155.12 | 67.99 | 67.99 | 67.99 | 85.76 | 155.12 | 155.12 | 155.12 |
| A | 26–63 | 4 | 86.53 | 53.14 | 14.07 | 140.28 | 14.07 | 14.07 | 50.55 | 95.89 | 122.52 | 140.28 | 140.28 |

Table 7.D.6 Total Testing Time (in Minutes) at Each Raw Score Interval—Grade Eight

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–26 | 134 | 96.78 | 58.68 | 8.18 | 307.09 | 9.04 | 29.65 | 51.91 | 81.01 | 137.90 | 174.72 | 299.74 |
| 1 | 27–32 | 133 | 101.10 | 46.72 | 15.16 | 248.21 | 23.93 | 50.99 | 67.99 | 92.93 | 118.15 | 168.87 | 244.44 |
| 1 | 33–39 | 142 | 113.36 | 55.57 | 36.76 | 357.24 | 37.01 | 60.16 | 73.47 | 100.17 | 132.58 | 190.04 | 331.23 |
| 1 | 40–65 | 151 | 125.13 | 52.48 | 40.54 | 325.02 | 47.40 | 71.59 | 88.50 | 114.20 | 153.86 | 201.78 | 309.24 |
| A | 0–15 | 3 | 87.96 | 37.40 | 45.24 | 114.79 | 45.24 | 45.24 | 45.24 | 103.84 | 114.79 | 114.79 | 114.79 |
| A | 16–19 | 3 | 109.49 | 35.61 | 73.92 | 145.15 | 73.92 | 73.92 | 73.92 | 109.41 | 145.15 | 145.15 | 145.15 |
| A | 20–23 | 2 | 192.88 | 33.36 | 169.29 | 216.47 | 169.29 | 169.29 | 169.29 | 192.88 | 216.47 | 216.47 | 216.47 |
| A | 24–58 | 4 | 133.64 | 82.12 | 33.08 | 215.56 | 33.08 | 33.08 | 67.91 | 142.96 | 199.36 | 215.56 | 215.56 |

Table 7.D.7 Total Testing Time (in Minutes) at Each Raw Score Interval—High School, All Grade Levels

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Form | Raw Score Interval | N | Mean | SD | Minimum | Maximum | % Pt. 1 | % Pt. 10 | % Pt. 25 | % Pt. 50 | % Pt. 75 | % Pt. 90 | % Pt. 99 |
| 1 | 0–20 | 61 | 59.17 | 24.41 | 14.84 | 138.09 | 14.84 | 25.94 | 41.00 | 61.60 | 76.93 | 84.28 | 138.09 |
| 1 | 21–29 | 62 | 68.15 | 35.17 | 14.11 | 207.60 | 14.11 | 32.53 | 49.44 | 60.82 | 76.94 | 110.98 | 207.60 |
| 1 | 30–37 | 62 | 81.66 | 29.34 | 39.45 | 156.27 | 39.45 | 48.19 | 56.23 | 74.07 | 105.09 | 118.67 | 156.27 |
| 1 | 38–62 | 68 | 101.01 | 39.47 | 45.26 | 303.29 | 45.26 | 56.40 | 73.09 | 98.21 | 119.55 | 133.36 | 303.29 |

### Appendix 7.E: Reliability Analyses

**Notes:**

* The reliabilities are reported only for samples that comprise 11 or more examinees.
* In some cases in [appendix 7.E](#_Appendix_7.E:_Reliability), score reliabilities were not estimable and are presented in the tables as “N/A.”
* Results based on samples that contain 50 or fewer examinees should be interpreted with caution because of small sample sizes.
* Nonbinary is not included in table 7.E.1 because of insufficient sample size.

Table 7.E.1 Reliabilities and SEMs by Gender

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Male N | Male Reliability | Male SEM | Female N | Female Reliability | Female SEM |
| Grade 3 | 566 | 0.86 | 3.50 | 617 | 0.87 | 3.53 |
| Grade 4 | 492 | 0.85 | 3.44 | 528 | 0.86 | 3.45 |
| Grade 5 | 436 | 0.83 | 3.50 | 454 | 0.85 | 3.50 |
| Grade 6 | 454 | 0.86 | 3.62 | 464 | 0.88 | 3.58 |
| Grade 7 | 328 | 0.82 | 3.67 | 363 | 0.82 | 3.64 |
| Grade 8 | 267 | 0.83 | 3.61 | 274 | 0.85 | 3.56 |
| High school | 104 | 0.86 | 3.52 | 120 | 0.86 | 3.52 |

Table 7.E.2 Reliabilities and SEMs by Economic Status

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Not Disadvantaged N | Not Disadvantaged Reliability | Not Disadvantaged SEM | Disadvantaged N | Disadvantaged Reliability | Disadvantaged SEM |
| Grade 3 | 426 | 0.89 | 3.52 | 757 | 0.82 | 3.51 |
| Grade 4 | 358 | 0.87 | 3.45 | 662 | 0.85 | 3.44 |
| Grade 5 | 311 | 0.84 | 3.51 | 579 | 0.84 | 3.49 |
| Grade 6 | 357 | 0.89 | 3.55 | 561 | 0.86 | 3.63 |
| Grade 7 | 308 | 0.83 | 3.64 | 383 | 0.81 | 3.66 |
| Grade 8 | 203 | 0.84 | 3.56 | 338 | 0.84 | 3.60 |
| High school | 31 | 0.82 | 3.47 | 193 | 0.87 | 3.53 |

Table 7.E.3 Reliabilities and SEMs by Special Education Services

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | No Special Services N | No Special Services Reliability | No Special Services SEM | Special Services N | Special Services Reliability | Special Services SEM |
| Grade 3 | 1,096 | 0.87 | 3.52 | 87 | 0.80 | 3.52 |
| Grade 4 | 944 | 0.86 | 3.45 | 76 | 0.81 | 3.41 |
| Grade 5 | 833 | 0.84 | 3.50 | 57 | 0.66 | 3.48 |
| Grade 6 | 865 | 0.87 | 3.60 | 53 | 0.74 | 3.65 |
| Grade 7 | 671 | 0.82 | 3.66 | 20 | 0.75 | 3.52 |
| Grade 8 | 526 | 0.84 | 3.59 | 15 | 0.75 | 3.48 |
| High school | 213 | 0.86 | 3.54 | 11 | 0.88 | 3.12 |

Table 7.E.4 Reliabilities and SEMs by Attendance in US Schools

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Less than 12 Months N | Less than 12 Months Reliability | Less than 12 Months SEM | 12 Months or More N | 12 Months or More Reliability | 12 Months or More SEM |
| Grade 3 | 9 | N/A | N/A | 1,174 | 0.87 | 3.52 |
| Grade 4 | 7 | N/A | N/A | 1,013 | 0.86 | 3.45 |
| Grade 5 | 2 | N/A | N/A | 888 | 0.84 | 3.50 |
| Grade 6 | 6 | N/A | N/A | 912 | 0.87 | 3.61 |
| Grade 7 | 1 | N/A | N/A | 690 | 0.82 | 3.66 |
| Grade 8 | 2 | N/A | N/A | 539 | 0.84 | 3.59 |
| High school | 32 | 0.86 | 3.45 | 192 | 0.87 | 3.53 |

Table 7.E.5 Reliabilities and SEMs by Spanish Instruction Status

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | Received Instruction in Spanish N | Received Instruction in Spanish Reliability | Received Instruction in Spanish SEM | Not Received Instruction in Spanish N | Not Received Instruction in Spanish Reliability | Not Received Instruction in Spanish SEM |
| Grade 3 | 1,109 | 0.87 | 3.52 | 74 | 0.75 | 3.46 |
| Grade 4 | 972 | 0.86 | 3.46 | 48 | 0.73 | 3.25 |
| Grade 5 | 862 | 0.84 | 3.51 | 28 | 0.79 | 3.41 |
| Grade 6 | 875 | 0.87 | 3.60 | 43 | 0.86 | 3.60 |
| Grade 7 | 680 | 0.83 | 3.66 | 11 | 0.77 | 3.46 |
| Grade 8 | 528 | 0.84 | 3.59 | 13 | 0.85 | 3.47 |
| High school | 188 | 0.86 | 3.53 | 36 | 0.87 | 3.46 |

Table 7.E.6 Reliabilities and SEMs by English Proficiency

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | English Only N | English Only Reliability | English Only SEM | Initial Fluent English Proficient (IFEP) N | IFEP Reliability | IFEP SEM |
| Grade 3 | 378 | 0.88 | 3.52 | 56 | 0.89 | 3.46 |
| Grade 4 | 311 | 0.84 | 3.46 | 45 | 0.90 | 3.36 |
| Grade 5 | 294 | 0.85 | 3.48 | 40 | 0.87 | 3.40 |
| Grade 6 | 271 | 0.89 | 3.57 | 37 | 0.86 | 3.52 |
| Grade 7 | 225 | 0.84 | 3.65 | 25 | 0.84 | 3.58 |
| Grade 8 | 125 | 0.84 | 3.61 | 30 | 0.79 | 3.38 |
| High school | 16 | 0.77 | 3.36 | 3 | N/A | N/A |

Table 7.E.7 Reliabilities and SEMs by English Proficiency (Continued)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade Level or Grade Band | EL N | EL Reliability | EL SEM | RFEP N | RFEP Reliability | RFEP SEM |
| Grade 3 | 604 | 0.83 | 3.50 | 143 | 0.89 | 3.51 |
| Grade 4 | 450 | 0.83 | 3.43 | 213 | 0.88 | 3.43 |
| Grade 5 | 337 | 0.75 | 3.51 | 219 | 0.83 | 3.47 |
| Grade 6 | 284 | 0.81 | 3.66 | 325 | 0.86 | 3.55 |
| Grade 7 | 147 | 0.73 | 3.62 | 294 | 0.79 | 3.63 |
| Grade 8 | 95 | 0.76 | 3.59 | 291 | 0.82 | 3.57 |
| High school | 81 | 0.88 | 3.51 | 120 | 0.86 | 3.53 |

Table 7.E.8 Reliability of Classification for Grade Three: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 300–348 | 0.67 | 0.04 | 0.00 | 0.71 |
| 349–359 | 0.05 | 0.13 | 0.01 | 0.19 |
| 360–399 | 0.00 | 0.02 | 0.08 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.88

Table 7.E.9 Reliability of Classification for Grade Three: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 300–348 | 0.65 | 0.06 | 0.00 | 0.71 |
| 349–359 | 0.06 | 0.11 | 0.02 | 0.19 |
| 360–399 | 0.00 | 0.02 | 0.07 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.83

Table 7.E.10 Reliability of Classification for Grade Four: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 400–448 | 0.62 | 0.05 | 0.00 | 0.68 |
| 449–459 | 0.05 | 0.17 | 0.01 | 0.22 |
| 460–499 | 0.00 | 0.03 | 0.07 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.86

Table 7.E.11 Reliability of Classification for Grade Four: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 400–448 | 0.59 | 0.08 | 0.01 | 0.68 |
| 449–459 | 0.06 | 0.14 | 0.02 | 0.22 |
| 460–499 | 0.00 | 0.03 | 0.06 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.80

Table 7.E.12 Reliability of Classification for Grade Five: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 500–545 | 0.50 | 0.07 | 0.01 | 0.57 |
| 546–559 | 0.05 | 0.27 | 0.01 | 0.33 |
| 560–599 | 0.00 | 0.03 | 0.07 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.83

Table 7.E.13 Reliability of Classification for Grade Five: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 500–545 | 0.47 | 0.10 | 0.01 | 0.57 |
| 546–559 | 0.07 | 0.23 | 0.02 | 0.33 |
| 560–599 | 0.00 | 0.03 | 0.06 | 0.10 |

Estimated Proportion Correctly Classified: Total = 0.76

Table 7.E.14 Reliability of Classification for Grade Six: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 600–647 | 0.41 | 0.06 | 0.01 | 0.48 |
| 648–659 | 0.04 | 0.25 | 0.02 | 0.30 |
| 660–699 | 0.01 | 0.05 | 0.16 | 0.22 |

Estimated Proportion Correctly Classified: Total = 0.82

Table 7.E.15 Reliability of Classification for Grade Six: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 600–647 | 0.39 | 0.08 | 0.01 | 0.48 |
| 648–659 | 0.06 | 0.2 | 0.04 | 0.3 |
| 660–699 | 0.01 | 0.06 | 0.15 | 0.22 |

Estimated Proportion Correctly Classified: Total = 0.74

Table 7.E.16 Reliability of Classification for Grade Seven: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 700‒743 | 0.45 | 0.09 | 0.01 | 0.55 |
| 744‒759 | 0.03 | 0.33 | 0.00 | 0.36 |
| 760‒799 | 0.01 | 0.07 | 0.01 | 0.09 |

Estimated Proportion Correctly Classified: Total = 0.79

Table 7.E.17 Reliability of Classification for Grade Seven: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 700‒743 | 0.42 | 0.11 | 0.02 | 0.55 |
| 744‒759 | 0.06 | 0.27 | 0.03 | 0.36 |
| 760‒799 | 0.01 | 0.05 | 0.03 | 0.09 |

Estimated Proportion Correctly Classified: Total = 0.72

Table 7.E.18 Reliability of Classification for Grade Eight: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 800–847 | 0.51 | 0.06 | 0.00 | 0.57 |
| 848–859 | 0.06 | 0.23 | 0.02 | 0.31 |
| 860–899 | 0.00 | 0.04 | 0.07 | 0.11 |

Estimated Proportion Correctly Classified: Total = 0.81

Table 7.E.19 Reliability of Classification for Grade Eight: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 800–847 | 0.48 | 0.09 | 0.00 | 0.57 |
| 848–859 | 0.08 | 0.19 | 0.04 | 0.31 |
| 860–899 | 0.00 | 0.04 | 0.07 | 0.11 |

Estimated Proportion Correctly Classified: Total = 0.74

Table 7.E.20 Reliability of Classification for High School: Decision Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 900–949 | 0.67 | 0.04 | 0.00 | 0.70 |
| 950–959 | 0.07 | 0.14 | 0.03 | 0.24 |
| 960–999 | 0.00 | 0.01 | 0.05 | 0.06 |

Estimated Proportion Correctly Classified: Total = 0.85

Table 7.E.21 Reliability of Classification for High School: Decision Consistency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reporting Score | Score Reporting Range 1 | Score Reporting Range 2 | Score Reporting Range 3 | Category Total |
| 900–949 | 0.64 | 0.06 | 0.00 | 0.70 |
| 950–959 | 0.08 | 0.11 | 0.05 | 0.24 |
| 960–999 | 0.00 | 0.01 | 0.05 | 0.06 |

Estimated Proportion Correctly Classified: Total = 0.80

## Quality Control Procedures

The California Department of Education (CDE) and ETS implemented rigorous quality control procedures throughout the test development, administration, scoring, and analyses processes for the California Spanish Assessment (CSA). 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 in evaluating those products and services. This chapter highlights the quality control processes used at various stages of administration.

### Quality Control of Item Development

ETS’ goal is to provide the best standards-based and innovative items for the CSA. Items developed for the CSA were subject to an extensive item review process. The item writers hired to develop CSA items and tasks, some of whom are current California educators, were trained in California Assessment of Student Performance and Progress (CAASPP) and ETS policies on quality control of item content, sensitivity and bias guidelines, and guidelines for accessibility to ensure that the items allow the widest possible range of students to demonstrate their content knowledge.

Once a written item was accepted for authoring—that is, once it was entered into ETS’ item bank and formatted for use in an assessment—ETS employed a series of internal and external reviews. These reviews used established criteria and specifications to judge the quality of item content and to ensure that each item measured what it is intended to measure. These reviews also examined the overall quality of the test items before presentation to the CDE and item reviewers. To finish the process for item development, a group of California educators reviewed the items for accessibility, bias and sensitivity, and content, and made recommendations for item enhancement. The details on item development processes for quality control purposes are described in section [*3.5 ETS Item Review Process*](#_Item_Review_Process) of [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1).

### Quality Control of Test Form Development

The assembly of all test forms must conform to blueprints that represent a set of constraints and specifications. ETS conducted multiple levels of quality assurance (QA) checks on each constructed test form to ensure it met the form-building specifications. Both ETS Assessment and Learning Technology Research & Development (ALTRD) and Psychometric Analysis & Research (PAR) staff reviewed and signed off on the accuracy of forms before the test forms were posted for CDE review. Detailed information related to test assembly can be found in section [*3.8 Test Assembly and Length*](#_Test_Assembly_and).

In particular, the assembly of all test forms went through a certification process that included various checks, including verifying that

* all answers were correct,
* items were scored correctly in the item bank and incorrect answers were scored as incorrect,
* all items assessed the standard,
* all content in the item was correct,
* distractors were plausible,
* multiple-choice (MC) item options were parallel in structure,
* language was grade-level appropriate,
* no more than three MC items in a row had the same key,
* all graphics were correct (copyright, spelling, relevance, etc.),
* there were no errors in spelling or grammar, and
* items adhered to the approved style guide.

Reviews were also conducted for functionality and sequencing during the user acceptance testing (UAT) process to ensure all items functioned as expected.

### Quality Control of Test Materials

Brief descriptions of other types of materials used for and during testing appear in the following subsections.

#### Developing Computer-based Assessments

The steps taken to develop and ensure the quality of the computer-based assessments are described in [*Chapter 3: Item Development and Form Assembly*](#_Item_Development_and_1).

#### Developing Test Administration Instructions

ETS staff consulted with internal subject matter experts and conducted validation checks to verify that test instructions accurately matched the testing processes. Copy editors and content editors reviewed each document for spelling, grammar, accuracy, and adherence to CDE style and usage requirements as well as the CDE accessibility standards. Instructions for the CSA were written in Spanish to be read to students in Spanish.

CSA content was incorporated to fit the CAASPP System specifications. All CAASPP documents were approved by the CDE before they could be published to the CAASPP website. Only nonsecure documents were posted to this website.

#### Processing Test Materials

Computer-based tests that were submitted by students were transmitted from Cambium Assessment, Inc. (CAI) to ETS each day. Each system was checked for the completeness of the student record, and records that were identified as having an error were flagged for review.

### Quality Control of Test Administration

The quality of test administration for the CSA, and all assessments administered as part of the CAASPP System, was monitored and controlled through several strategies. A fully staffed support center, the California Technical Assistance Center (CalTAC), supported all local educational agencies (LEAs) in the administration of CAASPP assessments. In addition to providing guidance and answering questions, CalTAC regularly conducted outreach campaigns on particular administration topics to ensure all LEAs understood correct test administration procedures. CalTAC was guided by a core group of LEA outreach and advocacy staff that managed communications to LEAs; provided regional and web-based trainings; and hosted a website, [the](https://www.caaspp.org/) CAASPP 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 CAASPP assessments, including the CSA. 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 QA services for all testing programs managed by ETS. The detailed procedures the OTI developed and applied in quality control are described in subsection [*4.6.1 ETS’ Office of Testing Integrity*](#_ETS’_Office_of).

### Quality Control of Scoring

#### 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 reporting specifications approved by the CDE. These specifications contain detailed scoring procedures, along with the procedures for determining whether a student has attempted a test and whether that student’s response data should be included in the statistical analysis and calculation for computing summary data.

Prior to the test administration, ETS ALTRD staff reviewed and verified the keys and scoring rubrics for each item. Then, these keys and rubrics were provided to CAI for implementing machine-scoring of the item responses. In addition, the student’s original response string was stored for data verification and auditing purposes. 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.

#### Quality Control of Machine-Scoring Procedures

To ensure valid item-level scoring for the CSA, 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 CSA forms by taking sample tests. Responses to the test forms 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 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 a test 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 test contained information for each item, keys for MC 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.

#### Enterprise Score Key Management System Processing

Prior to the start of the test administration, test-level scores were defined in a scoring model configured in ETS’ Enterprise Score Key Management (eSKM) system.

After the administration started, and after CAI completed machine-scoring, item scores and responses were delivered to ETS. ETS’ Centralized Repository Distribution System and Enterprise Service Bus departments collected and parsed .xml files that contained student response data from CAI. The eSKM system collected and calculated individual students’ overall scores and generated student scores in the approved statistical extract format. The data extracts were sent to ETS’ Data Quality Services for data validation.

Following successful validation, the student response statistical extracts were made available to the PAR team. The eSKM system implemented scoring procedures specified by the PAR team.

ETS developed two parallel scoring systems to produce and verify student scores:

1. The eSKM scoring system received an individual student’s item scores and item responses from CAI and calculated individual student scores for ETS’ reporting systems.
2. The ETS PAR team computed individual student scores based on item scores delivered by CAI.

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 to ensure the quality and accuracy of scoring and to support the transfer of scores into the Test Operations Management System (TOMS), the database of the student records scoring system.

### Quality Control of Psychometric Processes

#### Development of Scoring Specifications

ETS scoring specifications for the CSA were completed, reviewed, approved, and checked in advance of the receipt of student response data. Before psychometric analysis, PAR developed a psychometric analysis plan and road map, describing each step of psychometric analyses, procedures, and schedules. This plan was submitted to the CDE for review and approval. After that, psychometric specifications were developed for ETS data analysts conducting all analyses. Psychometric specifications contained detailed scoring procedures as well as the procedures for determining whether a student attempted a test and whether that student’s response data should be included in the statistical analyses and calculations for computing summary data.

#### Psychometric Analyses

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 CSA results data, including classical item analyses, differential item functioning (DIF), item response theory (IRT) calibration, linking, and scaling.

Detailed checklists were developed by members of the team for each of the statistical procedures performed on the CSA data. Classical item analyses and DIF analyses were run and confirmed by independent analysts. Results were then reviewed by the ETS psychometricians to compile a list of flagged items. Items that were flagged for questionable statistical attributes were sent to ETS ALTRD staff for review; their comments were reviewed by the psychometricians before the review by the CDE. The ETS ALTRD and PAR teams worked together to evaluate and make recommendations to the CDE about any problematic items that should be removed from IRT calibration.

In the calibration and equating process, checks were made to ascertain that the input files were established accurately. Checks were also made on the number of items, number of examinees with valid scores, IRT item difficulty estimates, standard errors for the item difficulty estimates, and the linking and scaling process. Two psychometricians conducted parallel calibration, robust-z, mean-to-mean linking, scaling, and classical linear equating 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 fit the data.

After the calibration and equating analyses were complete, ETS’ PAR team documented all options and results from the calibration and equating process for the director and manager reviews. After the director and manager reviewed, ETS summarized the calibration and equating results and proposed recommendations for the CDE to review and approve. After the CDE approved the equating decisions, raw-to-scale score conversion tables were generated.

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 score reporting range 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 the CSA 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 ETS’ Data Quality Services and Center of Reporting & Scoring Services teams, as well as running of all the Student Score Reports (SSRs) through ETS’ patented Quality Control 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’s 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’s 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 the accuracy of 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 the CDE’s approval of the reports generated for the quality control LEAs, ETS continued with the first batch of report production. The quality control LEAs comprised CDE‑selected LEAs to validate a subset of LEAs that contained key reporting characteristics (e.g., academic achievement) and demographics of the state. The reviewed set of reports incorporated CDE-selected LEAs and provided the final check prior to generating the reports and making them electronically available for download from 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 tests that, for example, identified students who tested but responded to no items, who were not tested because of parent/guardian request, or who did not complete the test because of illness. The methods for handling other anomalies were also covered in the specifications. These anomalies are described in more detail in subsection [*6.3.2 Special Cases*](#_Special_Cases).

### 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 was to verify that all systems, processes, and resources were ready for the test administration. ETS employed a number of approaches to verify ongoing systems performance, including monitoring of system availability and online 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 mirrored the final production environment, for testing and UAT. Final approval by the CDE triggered the final deployment of the system.

### Reference

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

## Continuous and Systematic Improvement

The California Spanish Assessment (CSA) had its first operational administration in spring 2019. Since its inception, continuous efforts have been made to improve the CSA in various ways. This chapter documents the processes whereby ETS ensures continuous improvements and the results of this process in the current year in the areas of test delivery and administration, item pool diversity, and accessibility.

### Feedback for Continuous Improvement Survey

The California Assessment of Student Performance and Progress (CAASPP) program solicits feedback annually through the CAASPP and English Language Proficiency Assessments for California Feedback for Continuous Improvement Survey (formerly the “Post-Test Survey”). The feedback received from this survey is one of the key sources of information ETS and the California Department of Education (CDE) use to inform continuous growth of the CSA and CAASPP as a whole. Local educational agency (LEA) and test site staff, as well as test administrators and test examiners, were invited to respond.

A total of 1,615 California educators provided specific, actionable insights about their testing experience; in a more typical test administration year, 8,000 or more responses are generally received. Overall, California educators continue to express positive experiences in their preparations for the CAASPP. Although the 2020–2021 school year included the daunting task of remote testing, educators also felt that the resources and training materials they were given were useful in preparing them and their students for test administration. Their feedback generally described smooth preparation, training, support, and assessment administration experiences.

Only 11 responses to questions asked in the CSA section of the survey were received. Results showed that respondents who planned to administer the CSA in 2020–2021 would administer it in 2021–2022 as well. Respondents were asked whether their decision to administer the CSA would be impacted if the CSA were to include writing prompts that required local scoring. Eighty-two percent reported that their LEA would administer the CSA to fewer students. About 64 percent (seven respondents) reported that they expected to register more students for the CSA in the 2021–2022 administration.

### Administration and Test Delivery

#### Training and Communication

ETS and the CDE have fully integrated the CSA into the CAASPP System of assessments by including it in the overall CAASPP training and resources, with communication as a focal point. Because the CSA is one of the newer assessments and is voluntary, ETS continues to provide statewide training with specific information on the CSA to LEA staff and test administrators to help LEAs understand and interpret CSA scores and to communicate the availability of the CSA.

ETS has developed practice and training tests that are available to the general public online. Refer to section [*4.5 Practice and Training Tests*](#_Practice_and_Training) for additional information about these products.

Both the practice and training tests include all accessibility resources that the operational assessment provides. A student’s experience with the practice and training tests can help inform the decision about whether or not a student takes the operational CSA. LEAs are encouraged to use the practice and training tests to help students become more familiar with using the technology and technology-enhanced items prior to taking the operational assessment.

### Accessibility

ETS continues to increase the number of accessibility resources available for the future assessment. In the interest of increasing the number of items that are “born accessible”—i.e., items that are as universally accessible as possible by all populations—ETS continues to investigate the construct-irrelevant use of item types that have no discernable difference from traditional test questions. As a result, ETS is reducing the development of both Match items and text-based Zone items, because they can be less accessible for students with sensory disabilities to discern.

Operational items were reviewed by teachers of students with visual impairment (TVIs) after their appearance on the CSA field test. Future item development was adjusted by incorporating TVIs’ feedback. To ensure items’ accessibility to students with sensory disabilities, CSA items are reviewed on an ongoing basis by TVIs.

### Item and Form Development

The 2020–2021 test administration year featured new online capabilities of the Item Banking Information System (IBIS) that the CDE could use for its initial review of new items and passages. The interface gives a view similar to what students experience during test taking, and metadata and other attachments are also available for review and use.

Additionally, the October 2020 meeting with California educators took place during the novel coronavirus disease 2019 pandemic. To accommodate travel restrictions, ETS conducted the meetings online using Zoom and the IBIS Content Review Tool (CRT). When educators were reviewing passages and items, they worked independently in the CRT to read, comment, and vote on passages and items. Once the set of items for a passage had been reviewed independently, an ETS staff member took notes about the discussions, using Zoom to share the CRT screen with participants. The notetaker cast a vote based on the group consensus, sometimes capturing a split in voting, if relevant. Thus, the CRT provided the individual and group records for the meetings. Reconciliation with the CDE took place subsequently, also using the CRT and Zoom.

1. Data for 2020–2021 was retrieved from the *CalEdFacts* web page on the CDE website. [↑](#footnote-ref-2)
2. This definition was retrieved from the California Longitudinal Pupil Achievement Data System (CALPADS) web page on the CDE website. [↑](#footnote-ref-3)
3. This technical report is based on the version of the *Usability, Accessibility, and Accommodations Guidelines* that was available during the 2020–2021 CAASPP administration. [↑](#footnote-ref-4)
4. No item review meetings were held for the 2020–2021 administration because reused forms were administered in 2020–2021. No new items were developed for the 2020–2021 administration. [↑](#footnote-ref-5)
5. For students in grades six through eight and high school [↑](#footnote-ref-6)
6. The *Crosswalk* has since been replaced with the Accessibility Strategies web page on the Tools for Teachers website. [↑](#footnote-ref-7)
7. In several applications of the Bookmark method, a target probability of two-thirds is used to define “most likely.” Refer, for example, to Cizek (2007). [↑](#footnote-ref-8)
8. The language standards, which focus on vocabulary, can be seen as an integral support of each of the four skills. [↑](#footnote-ref-9)