

# Informing the California Science Test (CAST) Blueprint Improvements: Results from the Psychometric Studies

**Contract #CN150012**

**Prepared for the California Department of Education by Educational Testing Service**

**Presented November 25, 2019**

Table of Contents

[Executive Summary 8](#_Toc25318770)

[Chapter 1: Overview 10](#_Toc25318771)

[1.1. Overview of the Purpose 10](#_Toc25318772)

[1.2. Overview of the Current Test Design and Blueprint 10](#_Toc25318773)

[1.2.1. Approved High-Level Test Design 10](#_Toc25318774)

[1.2.2. Test Administration Phases 11](#_Toc25318775)

[1.3. Overview of the Psychometric Studies 12](#_Toc25318776)

[1.3.1. Response-Time Analysis and Test Length 12](#_Toc25318777)

[1.3.2. Test Dimensionality Study 12](#_Toc25318778)

[1.3.3. Multistage Test Practicality Study 12](#_Toc25318779)

[1.3.4. Content Screener Study 13](#_Toc25318780)

[1.4. Overview of the Operational Year Data 13](#_Toc25318781)

[1.4.1. Test Design for the 2018–19 Operational Administration 13](#_Toc25318782)

[1.4.2. Testing Population 14](#_Toc25318783)

[Appendix 1.A: Testing Population by Student Group 15](#_Toc25318784)

[Chapter 2: Response-Time Analysis and Test Length 21](#_Toc25318785)

[2.1. Study Purpose 21](#_Toc25318786)

[2.2. Testing Time for the Total Test and by Segment 21](#_Toc25318787)

[2.3. Proposed Test Length 23](#_Toc25318788)

[2.4. Evaluation of the Proposed Test Length 24](#_Toc25318789)

[2.4.1. Estimated Response Time for the Proposed Blueprint 24](#_Toc25318790)

[2.4.2. Content Coverage 26](#_Toc25318791)

[2.4.3. Test Reliability 27](#_Toc25318792)

[2.5. Summary 28](#_Toc25318793)

[Appendix 2.A. Testing Time by Student Group 29](#_Toc25318794)

[Chapter 3: Test Dimensionality 45](#_Toc25318795)

[3.1. Study Purpose 45](#_Toc25318796)

[3.2. Study Design 45](#_Toc25318797)

[3.3. Methods 48](#_Toc25318798)

[3.3.1. Model Specification 48](#_Toc25318799)

[3.3.2. Local Dependency Evaluation 51](#_Toc25318800)

[3.4. Model Evaluation Criteria 53](#_Toc25318801)

[3.5. Results 53](#_Toc25318802)

[3.5.1. Test Dimensionality by Content Domain 54](#_Toc25318803)

[3.5.2. Test Dimensionality by SEP 56](#_Toc25318804)

[3.5.3. Test Dimensionality by CCC 59](#_Toc25318805)

[3.5.4. Test Dimensionality by Item Type 61](#_Toc25318806)

[3.5.5. Test Dimensionality by Task Type 63](#_Toc25318807)

[3.6. Implications on Calibration and Score Reporting 64](#_Toc25318808)

[References 66](#_Toc25318809)

[Appendix 3.A: Factor Loading Matrix 67](#_Toc25318810)

[Chapter 4: The Practicality of a Multistage Adaptive Test Design 87](#_Toc25318811)

[4.1. Study Purpose 87](#_Toc25318812)

[4.2. Methods 88](#_Toc25318813)

[4.2.1. Assemble MST Panels with Two and Three Levels of Difficulty 88](#_Toc25318814)

[4.2.2. Evaluate Panel Performance 92](#_Toc25318815)

[4.3. Item Pools 95](#_Toc25318816)

[4.4. Results 99](#_Toc25318817)

[4.4.1. Information Functions 99](#_Toc25318818)

[4.4.2. Conditional Standard Error of Measurement (CSEM) 104](#_Toc25318819)

[4.4.3. Conditional Bias 106](#_Toc25318820)

[4.4.4. Relative Efficiency 108](#_Toc25318821)

[4.4.5. Reliability 110](#_Toc25318822)

[4.5. Implications for Test Blueprint Change 110](#_Toc25318823)

[References 111](#_Toc25318824)

[Appendix 4.A: Table Data for Graphs 112](#_Toc25318825)

[Data for Scatter Plots of Discrimination and Difficulty Parameters 112](#_Toc25318826)

[Data for Information Curves 130](#_Toc25318827)

[Data for Conditional Standard Errors of Measurement (CSEM) Graphs 142](#_Toc25318828)

[Data for Conditional Bias Graphs 148](#_Toc25318829)

[Data for Relative Efficiency Graphs 154](#_Toc25318830)

[Chapter 5: The Utility of a Content Screener 160](#_Toc25318831)

[5.1. Study Purpose 160](#_Toc25318832)

[5.2. Methods 161](#_Toc25318833)

[Step 1: Data Preparation 162](#_Toc25318834)

[Step 2: Comparable Scores 162](#_Toc25318835)

[Step 3: Descriptive Analyses 163](#_Toc25318836)

[Step 4: The Alignment Index 163](#_Toc25318837)

[Step 5: Fitted Linear Models 165](#_Toc25318838)

[5.3. Study Data 166](#_Toc25318839)

[5.4. Results 166](#_Toc25318840)

[5.4.1. Descriptive Statistics 166](#_Toc25318841)

[5.4.2. Model Results 174](#_Toc25318842)

[5.4.3. Follow-Up Analyses 175](#_Toc25318843)

[5.5. Limitations 179](#_Toc25318844)

[5.6. Implications for Test Blueprint Change 179](#_Toc25318845)

[Chapter 6: Summary of Recommendations 180](#_Toc25318846)

List of Tables

[Table 1.1 Number of Available Blocks in the 2018–19 Administration 13](#_Toc25318983)

[Table 1.2 Testing Population 14](#_Toc25318984)

[Table 1.A.1 CAST Operational Test Taking Rates for Grade Five by Student Group 15](#_Toc25318985)

[Table 1.A.2 Operational Test Taking Rates for Grade Eight by Student Group 16](#_Toc25318986)

[Table 1.A.3 Operational Test Taking Rates for Grade Ten by Student Group 17](#_Toc25318987)

[Table 1.A.4 CAST Operational Test Taking Rates for Grade Eleven by Student Group 18](#_Toc25318988)

[Table 1.A.5 Operational Test Taking Rates for Grade Twelve by Student Group 19](#_Toc25318989)

[Table 1.A.6 Operational Test Taking Rates for High School by Student Group 20](#_Toc25318990)

[Table 2.1 Time Spent (in Minutes) on the CAST for Grade Five 21](#_Toc25318991)

[Table 2.2 Time Spent (in Minutes) on the CAST for Grade Eight 21](#_Toc25318992)

[Table 2.3 Time Spent (in Minutes) on the CAST for High School 22](#_Toc25318993)

[Table 2.4 Proposed Test Length 23](#_Toc25318994)

[Table 2.5 Selected Item Type Categories in the CAST Operational Administration 24](#_Toc25318995)

[Table 2.6 Median Response Time by Item Type 25](#_Toc25318996)

[Table 2.7 Response Time Comparison for the Current and Recommended Blueprints 26](#_Toc25318997)

[Table 2.8 Anticipated Number of Unique PEs Assessed on Segments A and B of the CAST 27](#_Toc25318998)

[Table 2.9 A Comparison of the Reliability of the Current and Recommended Blueprints 27](#_Toc25318999)

[Table 2.10 A Comparison of the Reliability of the Current and Recommended Blueprints by Domain 28](#_Toc25319000)

[Table 2.A.1 Time Spent (in Minutes) on the CAST for Grade Five for Segment A by Student Group 29](#_Toc25319001)

[Table 2.A.2 Time Spent (in Minutes) on the CAST for Grade Five for Segment B by Student Group 30](#_Toc25319002)

[Table 2.A.3 Time Spent (in Minutes) on the CAST for Grade Five for Segment C Discrete Block by Student Group 31](#_Toc25319003)

[Table 2.A.4 Time Spent (in Minutes) on the CAST for Grade Five for Segment C Performance Task by Student Group 32](#_Toc25319004)

[Table 2.A.5 Time Spent (in Minutes) on the CAST for Grade Five for Total Test with Discrete Field Test Block by Student Group 33](#_Toc25319005)

[Table 2.A.6 Time Spent (in Minutes) on the CAST for Grade Five for Total Test with Performance Task Field Test Block by Student Group 34](#_Toc25319006)

[Table 2.A.7 Time Spent (in Minutes) on the CAST for Grade Eight for Segment A by StudentGroup 34](#_Toc25319007)

[Table 2.A.8 Time Spent (in Minutes) on the CAST for Grade Eight for Segment B by Student Group 35](#_Toc25319008)

[Table 2.A.9 Time Spent (in Minutes) on the CAST for Grade Eight for Segment C Discrete Block by Student Group 36](#_Toc25319009)

[Table 2.A.10 Time Spent (in Minutes) on the CAST for Grade Eight for Segment C Performance Task by Student Group 37](#_Toc25319010)

[Table 2.A.11 Time Spent (in Minutes) on the CAST for Grade Eight for Total Test with Discrete Field Test Block by Student Group 38](#_Toc25319011)

[Table 2.A.12 Time Spent (in Minutes) on the CAST for Grade Eight for Total Test with Performance Task Field Test Block by Student Group 39](#_Toc25319012)

[Table 2.A.13 Time Spent (in Minutes) on the CAST for High School for Segment A by Student Group 40](#_Toc25319013)

[Table 2.A.14 Time Spent (in Minutes) on the CAST for High School for Segment B by Student Group 41](#_Toc25319014)

[Table 2.A.15 Time Spent (in Minutes) on the CAST for High School for Segment C Discrete Block by Student Group 42](#_Toc25319015)

[Table 2.A.16 Time Spent (in Minutes) on the CAST for High School for Segment C Performance Task by Student Group 42](#_Toc25319016)

[Table 2.A.17 Time Spent (in Minutes) on the CAST for High School for Total Test with Discrete Field Test Block by Student Group 43](#_Toc25319017)

[Table 2.A.18 Time Spent (in Minutes) on the CAST for High School for Total Test with Performance Task Field Test Block by Student Group 44](#_Toc25319018)

[Table 3.1 Hypothesized Dimensional Structures in the Study 46](#_Toc25319019)

[Table 3.2 Forms Used in the Test Dimensionality Study 48](#_Toc25319020)

[Table 3.3 Evaluation Indices for the Bifactor Model by Content Domain (Two PTs) 54](#_Toc25319021)

[Table 3.4 Evaluation Indices for the Bifactor Model by Content Domain (Three PTs) 55](#_Toc25319022)

[Table 3.5 Correlations Among the Latent Content Domain Scores (Two PTs) 56](#_Toc25319023)

[Table 3.6 Correlations Among the Latent Content Domain Scores (Three PTs) 56](#_Toc25319024)

[Table 3.7 Evaluation Indices for the Bifactor Model by SEP (Two PTs) 57](#_Toc25319025)

[Table 3.8 Evaluation Indices for the Bifactor Model by SEP (Three PTs) 57](#_Toc25319026)

[Table 3.9 Correlations Among the Latent SEP Scores (Two PTs) 58](#_Toc25319027)

[Table 3.10 Correlations Among the Latent SEP Scores (Three PTs) 58](#_Toc25319028)

[Table 3.11 Evaluation Indices for the Bifactor Model by CCC (Two PTs) 59](#_Toc25319029)

[Table 3.12 Evaluation Indices for the Bifactor Model by CCC (Three PTs) 60](#_Toc25319030)

[Table 3.13 Correlations Among Latent CCC Scores (Two PTs) 61](#_Toc25319031)

[Table 3.14 Correlations Among Latent CCC Scores (Three PTs) 61](#_Toc25319032)

[Table 3.15 Evaluation Indices for the Bifactor Model by Item Type (Two PTs) 62](#_Toc25319033)

[Table 3.16 Evaluation Indices for the Bifactor Model by Item Type (Three PTs) 62](#_Toc25319034)

[Table 3.17 Correlations Among the Latent Item-Type Scores (Two PTs) 63](#_Toc25319035)

[Table 3.18 Correlations Among the Latent Item-Type Scores (Three PTs) 63](#_Toc25319036)

[Table 3.19 Evaluation Indices for the Bifactor Model by Task Type (Two PTs) 64](#_Toc25319037)

[Table 3.20 Evaluation Indices for the Bifactor Model by Task Type (Three PTs) 64](#_Toc25319038)

[Table 3.A.1 Factor Loading Matrix by Content Domain—Grade Five, Form One, Two PTs 67](#_Toc25319039)

[Table 3.A.2 Factor Loading Matrix by Content Domain—Grade Five, Form One, Three PTs 69](#_Toc25319040)

[Table 3.A.3 Factor Loading Matrix by Science and Engineering Practices (SEP)—Grade Five, Form One, Two PTs 71](#_Toc25319041)

[Table 3.A.4 Factor Loading Matrix by SEP—Grade Five, Form One, Three PTs 73](#_Toc25319042)

[Table 3.A.5 Factor Loading Matrix by Crosscutting Concept (CCC)—Grade Five, Form One, Two PTs 75](#_Toc25319043)

[Table 3.A.6 Factor Loading Matrix by Crosscutting Concept (CCC)—Grade Five, Form One, Three PTs 77](#_Toc25319044)

[Table 3.A.7 Factor Loading Matrix by Item Type—Grade Five, Form One, Two PTs 79](#_Toc25319045)

[Table 3.A.8 Factor Loading Matrix by Item Type—Grade Five, Form One, Three PTs 81](#_Toc25319046)

[Table 3.A.9 Factor Loading Matrix by Task Type—Grade Five, Form One, Two PTs 83](#_Toc25319047)

[Table 3.A.10 Factor Loading Matrix by Task Type—Grade Five, Form One, Three PTs 85](#_Toc25319048)

[Table 4.1 Content Specification and Conformation for Grade Five 89](#_Toc25319049)

[Table 4.2 Content Specification and Conformation for Grade Eight 90](#_Toc25319050)

[Table 4.3 Content Specification and Conformation for High School 91](#_Toc25319051)

[Table 4.4 Summary of Difficulty Parameters for Grade Five 96](#_Toc25319052)

[Table 4.5 Summary of Difficulty Parameters for Grade Eight 96](#_Toc25319053)

[Table 4.6 Summary of Difficulty Parameters for High School 96](#_Toc25319054)

[Table 4.7 Marginal Test Reliabilities 110](#_Toc25319055)

[Table 4.A.1 Scatter Plot of Discrimination and Difficulty Parameters for Grade Five Data 112](#_Toc25319056)

[Table 4.A.2 Scatter Plot of Discrimination and Difficulty Parameters for Grade Eight Data 118](#_Toc25319057)

[Table 4.A.3 Scatter Plot of Discrimination and Difficulty Parameters for High School Data 124](#_Toc25319058)

[Table 4.A.4 Information Curves of the Assembled MST 1-2 Design Panel for Grade Five Data 130](#_Toc25319059)

[Table 4.A.5 Information Curves of the Assembled MST 1-3 Design Panel for Grade Five Data 132](#_Toc25319060)

[Table 4.A.6 Information Curves of the Assembled MST 1-2 Design Panel for Grade Eight Data 134](#_Toc25319061)

[Table 4.A.7 Information Curves of the Assembled MST 1-3 Design Panel for Grade Eight Data 136](#_Toc25319062)

[Table 4.A.8 Information Curves of the Assembled MST 1-2 Design Panel for High School Data 138](#_Toc25319063)

[Table 4.A.9 Information Curves of the Assembled MST 1-3 Design Panel for High School Data 140](#_Toc25319064)

[Table 4.A.10 CSEMs for Grade Five Data 142](#_Toc25319065)

[Table 4.A.11 CSEM for Grade Eight Data 144](#_Toc25319066)

[Table 4.A.12 CSEMs for High School Data 146](#_Toc25319067)

[Table 4.A.13 Conditional Bias for Grade Five Data 148](#_Toc25319068)

[Table 4.A.14 Conditional Bias for Grade Eight Data 150](#_Toc25319069)

[Table 4.A.15 Conditional Bias for High School Data 152](#_Toc25319070)

[Table 4.A.16 Relative Efficiency for Grade Five Data 154](#_Toc25319071)

[Table 4.A.17 Relative Efficiency for Grade Eight Data 156](#_Toc25319072)

[Table 4.A.18 Relative Efficiency for High School Data 158](#_Toc25319073)

[Table 5.1 Score Reliabilities and Correlations Between Scores, Grade Five 167](#_Toc25319074)

[Table 5.2 Score Reliabilities and Correlations Between Scores, Grade Eight 167](#_Toc25319075)

[Table 5.3 Score Reliabilities and Correlations Between Scores, High School 168](#_Toc25319076)

[Table 5.4 Summary Statistics for the Alignment Index Across Student Groups, Grade Five 171](#_Toc25319077)

[Table 5.5 Summary Statistics for the Alignment Index Across Student Groups, Grade Eight 172](#_Toc25319078)

[Table 5.6 Summary Statistics for the Alignment Index Across Student Groups, High School 173](#_Toc25319079)

[Table 5.7 R2 and the Adjusted-R2 Values for the Fitted Linear Models 174](#_Toc25319080)

[Table 5.8 Average Total Segment B Scores for Students by the Students’ Weakest Domain 176](#_Toc25319081)

[Table 6.1 Summary of CAST Blueprint Improvements 180](#_Toc25319082)

List of Figures

[Figure 3.1. The Path Diagram for a Bifactor Model 49](#_Toc25310889)

[Figure 3.2 The Path Diagram for a Correlated Factor MIRT Model 50](#_Toc25310890)

[Figure 4.1 Typical MST Panel 87](#_Toc25310891)

[Figure 4.2 Sample Block Information Targets 92](#_Toc25310892)

[Figure 4.3 Example Conditional Standard Errors of Measurement 94](#_Toc25310893)

[Figure 4.4 Scatter Plot of Discrimination and Difficulty Parameters for Grade Five 97](#_Toc25310894)

[Figure 4.5 Scatter Plot of Discrimination and Difficulty Parameters for Grade Eight 98](#_Toc25310895)

[Figure 4.6 Scatter Plot of Discrimination and Difficulty Parameters for High School 99](#_Toc25310896)

[Figure 4.7 Information Curves of the Assembled MST 1-2 Design Panel for Grade Five 100](#_Toc25310897)

[Figure 4.8 Information Curves of the Assembled MST 1-3 Design Panel for Grade Five 100](#_Toc25310898)

[Figure 4.9  Information Curves of the Assembled MST 1-2 Design Panel for Grade Eight 101](#_Toc25310899)

[Figure 4.10 Information Curves of the Assembled MST 1-3 Design Panel for Grade Eight 102](#_Toc25310900)

[Figure 4.11 Information Curves of the Assembled MST 1-2 Design Panel for High School 103](#_Toc25310901)

[Figure 4.12 Information Curves of the Assembled MST 1-3 Design Panel for High School 103](#_Toc25310902)

[Figure 4.13 Conditional Standard Error of Measurement for Grade Five 104](#_Toc25310903)

[Figure 4.14 Conditional Standard Error of Measurement for Grade Eight 105](#_Toc25310904)

[Figure 4.15 Conditional Standard Error of Measurement for High School 105](#_Toc25310905)

[Figure 4.16 Conditional Bias for Grade Five 106](#_Toc25310906)

[Figure 4.17 Conditional Bias for Grade Eight 107](#_Toc25310907)

[Figure 4.18 Conditional Bias for High School 107](#_Toc25310908)

[Figure 4.19 Relative Efficiency for Grade Five 108](#_Toc25310909)

[Figure 4.20 Relative Efficiency for Grade Eight 109](#_Toc25310910)

[Figure 4.21 Relative Efficiency for High School 109](#_Toc25310911)

## Executive Summary

In March 2016, the California State Board of Education (SBE) approved the original test design for a general summative assessment developed to the California Next Generation Science Standards. In November 2017, the SBE approved the California Science Test (CAST) Blueprint. Since the 2016 and 2017 test design and blueprint approvals, Educational Testing Service (ETS) has developed and administered a census pilot (2016–‍17), a census field test (2017–18), and the first operational administration of CAST in 2018–19.

Special psychometric studies were conducted using the data from the 2017–18 field test and the 2018–19 first operational administration. The purpose was to evaluate the tests so as to inform the blueprint improvement recommendations.

The results of the special psychometric studies from the two administrations were very similar, so only results from the 2018–19 operational administration are included in this report. These psychometric analyses included the following:

* The response time analysis
* The test dimensionality
* Transitioning to a multistage adaptive test (MST) design
* The utility of implementing a content screener

Test response-time analysis suggests that the test length is likely too long for grade five and grade eight. For these two grades, the time students spent on Segment A was longer than the expected 60 minutes. For all three grades, the time students spent on the two performance tasks (PTs) in Segment B was shorter than the expected 40 minutes. It took longer for students who received the discrete block in Segment C to complete the test than those who received the PT.

The results from the test dimensionality study suggest that the CAST is essentially unidimensional, and that it is appropriate to calibrate items from different domains jointly and report a single total score that is based on all items in the test form.

The multistage adaptive practicality study suggests modest improvement in terms of measurement precision, with the benefits most noticeable for the lowest performing and highest performing students at grade eight and for the lowest performing students at grade five and high school. However, the item pool is still not robust enough to fully realize the benefit of the MST design. For grade five, the pool lacks enough difficult items to build a hard block that can be separated from the easy and the router blocks.

The results from the content screener study were inconclusive in that students who were assigned PTs in the students’ poorest performing domain did not consistently perform less well than students who were randomly assigned PTs in other domains. Therefore, it is inconclusive that a screener would be useful for the CAST.

Based on the results from these studies and feedback from California educators and the California Department of Education (CDE), ETS has made the following recommendations for changes to the test blueprint that will better meet the program’s needs:

* Reduce the number of discrete items in segments A and C
* Add a third PT to Segment B
* Discontinue the plan to implement a screener
* Postpone the implementation of the MST design until the pool can be expanded to fully realize the potential of the adaptive design

The proposed test length is presented in Table 2.4. The following benefits are realized with these recommended changes to the test blueprint:

* The number of students in grades five and eight who can complete the test within two hours increases.
* Each student will now receive three PTs in Segment B representing each of three science domains.
* The time needed to complete Segment C discrete items is reduced and more in line with the time needed to complete a Segment C PT.

## Overview

### Overview of the Purpose

In March 2016, the California State Board of Education (SBE) approved the original test design for a general summative assessment developed to the California Next Generation Science Standards (CA NGSS). In November 2017, the SBE approved the California Science Test (CAST) Blueprint. Since the 2016 and 2017 test design and blueprint approvals, Educational Testing Service (ETS) has developed and administered a census pilot (2016–17), a census field test (2017–18), and the first operational administration of CAST in 2018–19.

ETS analyzed results from the 2017–18 field test and the 2018–19 first operational administration. Based on the results from these studies and feedback from California educators and the California Department of Education (CDE), ETS has made recommendations for several changes to the test blueprint that will better meet the program’s needs. This report summarizes the results from the analyses using the 2018–19 operational administration data that were used to inform the blueprint improvement recommendations.

### Overview of the Current Test Design and Blueprint

California adopted the CA NGSS in September 2013. The CAST is an online assessment aligned with the CA NGSS that is intended to measure a full range of the CA NGSS performance expectations (PEs). Each PE is an integration of a disciplinary core idea (DCI), science and engineering practice (SEP), and crosscutting concept (CCC).

#### Approved High-Level Test Design

In the approved test design and blueprint, the CAST includes three major domains—‍Physical Sciences, Life Sciences, and Earth and Space Sciences—as well as one associated domain (Engineering, Technology, and Applications of Science). For scoring and reporting purposes, each of the three major science domains constitutes one-third of the test; items written to assess PEs associated with the fourth, associated domain are assigned to one of the three major science domains, depending upon the context of their stimulus.

Based on the approved high-level test design, the CAST, when fully implemented, would consist of three segments:

1. Segment A would be a two-stage adaptive section that presents items that are substantively equivalent in content but which may differ in difficulty so as to most appropriately match each student’s level of performance. This segment contains 32 to 34 discrete items with 42 to 44 points. (The adaptive aspect of the design was not implemented during the first and second operational administrations.)
2. Segment B would contain two performance tasks (PTs), each with four to six items set in a content domain context. The content presented in this segment is adapted to the student’s performance in Segment A. Where a student demonstrates conspicuously poor performance in a particular domain in Segment A, the student will not be presented with items based on the same domain since to do so would yield little or no additional measure of student performance. In this way, Segment A serves the role to screen out content assigned to a student in Segment B. Where a student performs with relative comparability in *all* science domains in Segment A, the assignment of items presented in Segment B will be random. (The role of Segment A as a content screener was not implemented.)
3. Segment C would contain either 13 discrete items or one PT. Segment C is intended to support embedded field testing of new items or the administration of operational items that could contribute additional, unique information at the group level.

Items in segments A and B contribute to reporting individual student scores. Items in all three segments—exclusive of field test items—are intended for reporting group-level scores.

A variety of item types are used to measure CA NGSS PEs. Different item types better measure targeted CA NGSS PEs. These item types range from the traditional multiple-choice (MC) and constructed-response (CR) items, to the new technology-enhanced items (TEIs) that involve the use of dynamic stimuli and other types of new media such as simulations.

#### Test Administration Phases

##### 2016–17 Pilot

The CAST pilot occurred during the 2016–17 administration. Its goal was to provide information on the performance of the newly developed CA NGSS–aligned items and item types, in particular the TEIs. Each grade had two forms, where each form consisted of 12 to 15 discrete items and one PT. Pilot analyses focused on the item performance rather than the form properties.

##### 2017–18 Field Test

In spring 2018, the CAST had a field test administration where a much larger number of items were field-tested. Items were assembled into blocks that met the test blueprint specifications. This enabled not only the evaluation of item performance, but also the evaluation of form properties and how students responded to the test.

The field test administration featured a large number of new items. The field test forms were built to meet the blueprint to the extent possible with the existing item pool.

The field test was taken by all students in grades five, eight, and twelve, as well as high school students in grades ten and eleven who were assigned by the local educational agency (LEA).

Field test data provided a rich source of information for researchers on the performance of the newly developed CA NGSS–aligned items, item types, and test form properties, including whether the CAST would benefit from potential blueprint changes.

The data collected from the field test were used as a first attempt to inform how student scores should be reported.

##### 2018–19 Operational Assessment

In spring 2019, the operational CAST was administered to all students in grades five and eight. Students in grade twelve who had not yet taken the science test in the field test year were automatically registered to take the test. LEAs could elect to administer the science test to students in grade ten or grade eleven who did not take the 2017–18 science test. Scores will be reported for each student who tested. The forms were built using the items field-tested in the 2017–18 administration that had passed the review by the data-review panel and the CDE.

The forms were built to conform as much as possible to the test blueprint, with the exception that some operational PTs included embedded FT items. Extra items were added to the PTs to reduce the chance of losing a full PT if some items performed poorly.

The multistage adaptive test (MST) test design and the content screener were implemented in neither the first nor second operational year. The plan was for ETS to use the data from the first operational year to inform the practicality and the utility of these design features.

### Overview of the Psychometric Studies

The psychometric analyses addressed the following questions:

1. Is the current test length appropriate?
2. If there is a need to change the blueprint based on the analyses, does the test dimensionality change under the proposed blueprint and will score reporting be affected?
3. Is there enough evidence to support the implementation of an MST design and to screen out a PT based on a student’s performance on Segment A items?

This report includes the analyses conducted to address these questions to inform the test blueprint improvement recommendations. These analyses include a response-time analysis, a test dimensionality study, an MST practicality study, and a content screener study.

#### Response-Time Analysis and Test Length

In the approved high-level test design and blueprint, the CAST is an untimed test and is expected to take approximately two hours to administer all three segments.

The response-time analysis summarized the time students spent on the total test, by test segment and by item type. This analysis was to address the question of whether the current test length is appropriate for students. If the test appeared to be too long, what would be the appropriate test length for students to complete within two hours?

#### Test Dimensionality Study

The dimensionality study provided evidence of whether the CAST measures a single integrated science construct or several distinctive content domains. The test dimensionality study has implications on how items are calibrated and on how scores are reported. If the test predominantly measures a single, latent variable—that is, the test is unidimensional or essentially unidimensional—then a unidimensional item response theory (IRT) model can be used to calibrate the items and score students. If the test measures multiple latent variables, the test is multidimensional and a multidimensional IRT model should be used.

#### Multistage Test Practicality Study

The MST practicality study used the item pool that passed the data review and the CDE review to assemble the MST forms for Segment A. The constructed MST forms were compared against a linear, nonadaptive Segment A form to evaluate the extent to which the MST form improves the measurement precision. This study was used to inform whether the item pool was adequate enough to implement an MST design for the CAST.

#### Content Screener Study

The content screener study evaluated the approach to selecting Segment B PTs for each student based on Segment A performance. Students who performed conspicuously poorly in one content domain, relative to the other domains, would not be administered a PT in that domain. They would receive two PTs in the other two content domains. This study evaluated whether a content screener would be useful to improve students’ performance on the PTs. The results of the study were used to inform the implementation of this design feature in future administrations.

### Overview of the Operational Year Data

#### Test Design for the 2018–19 Operational Administration

The CAST was administered as an online assessment from January 9 through July 15, 2019, to students in grade five, grade eight, and high school (grade ten, eleven, or twelve). All students in grades five and eight were required to take the test. All grade twelve students who hadn’t taken the field test were required to take the test, as well as students in grades ten and eleven who were assigned by the LEA.

Table 1.1 provides the number of blocks included for each segment in each grade.

Table 1.1 Number of Available Blocks in the 2018–19 Administration

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade | Segment A | Segment B | Segment C Discrete Blocks | Segment C Performance Tasks Blocks |
| Five | 2 | 5 | 15 | 6 |
| Eight | 2 | 6 | 15 | 5 |
| High School | 2 | 3 | 15 | 5 |

Each Segment A block contains 16 to 17 items. Each PT in Segment B contains 6 to 7 items. Each Segment C discrete block contains 13 items. Each Segment C PT block contains 7 to 8 items. Each student received the two Segment A blocks and two randomly selected Segment B PTs from different domains from the available blocks. Then, about half of the students were randomly assigned to one discrete block and the remaining students to one PT block in Segment C.

The delivery order of the blocks within a segment is random. For example, for Segment A, some students received the blocks in A1–A2 order and some received them in A2–A1 order. Students always received the discrete item blocks—blocks in segments A or C—first, and then the PTs. The discrete Segment C block can be placed either after the first A block or the second A block. The PT Segment C block can be placed either after the first or the second Segment B PT.

##### Scoring

The scoring included both human scoring and machine scoring. The selected-response (SR) items—MC items and TEIs—were machine-scored. The CR items, and some composite CR items (a two-part item with a CR component), were human- or artificial intelligence (AI)-scored. All responses for the CR or composite CR items in Segment B were scored. For the CR items in Segment C, not all responses were scored because these items are field test items and will not contribute to reporting individual student scores. A random sample of responses that is representative of the population was selected for scoring for each CR item to support the item analyses, IRT calibration, and the building of the AI scoring models.

#### Testing Population

Table 1.2 provides the number of registered students and the number of students who took the test. Students who are enrolled in grade five and grade eight are automatically registered for either the CAST or the California Alternate Assessment for Science (CAA for Science). Students enrolled in grade twelve are registered for one of the science assessments if they have not already completed a science test while in high school. Table 1.A.1 through Table 1.A.6 in appendix 1.A provide the testing population information by student group.

Table 1.2 Testing Population

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Grade Five | Grade Eight | High School—Grade Ten | High School—Grade Eleven | High School—Grade Twelve | High School—Total |
| Number of Registered Students | 464,158 | 474,971 | 28,124 | 272,744 | 323,200 | 624,443 |
| Number of Students Taking Test | 456,692 | 463,310 | 23,362 | 257,323 | 276,687 | 557,372 |
| Percentage of Students Taking the Test | 98.4 | 97.5 | 83.1 | 94.3 | 85.6 | 89.3 |

Across the grades, 83.1 percent to 98.4 percent of registered students took the test. For high school, grade ten and grade twelve had a slightly lower percentage, but overall, the percent of students who took the test was high. For the high school assessment, the percent of students from each grade who took the test is uneven—the majority of students who took the test were from grades eleven and twelve (4.2 percent from grade ten, 46.2 percent from grade eleven, and 49.7 percent from grade twelve). The number of registered students and the number of students taking the test in grade ten were low, given students have three years in which they could take the CAST.

### Appendix 1.A: Testing Population by Student Group

Table 1.A.1 CAST Operational Test Taking Rates for Grade Five by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 464,008 | 456,692 | 98.4 |
| Male | 238,699 | 234,626 | 98.3 |
| Female | 225,309 | 222,066 | 98.6 |
| English learner | 89,985 | 88,693 | 98.6 |
| English only | 269,034 | 263,658 | 98.0 |
| Reclassified fluent English proficient | 86,026 | 85,713 | 99.6 |
| Initial fluent English proficient | 18,065 | 17,934 | 99.3 |
| To be determined | 482 | 372 | 77.2 |
| English proficiency unknown | 416 | 322 | 77.4 |
| Economically disadvantaged | 288,110 | 284,275 | 98.7 |
| Not economically disadvantaged | 175,898 | 172,417 | 98.0 |
| American Indian or Alaska Native | 2,250 | 2,186 | 97.2 |
| Asian | 43,035 | 42,661 | 99.1 |
| Native Hawaiian or Other Pacific Islander | 2,115 | 2,076 | 98.2 |
| Filipino | 9,285 | 9,209 | 99.2 |
| Hispanic or Latino | 256,465 | 253,621 | 98.9 |
| Black or African American | 24,928 | 24,305 | 97.5 |
| White | 103,490 | 100,807 | 97.4 |
| Two or more races | 19,106 | 18,719 | 98.0 |
| Special education services | 59,420 | 57,163 | 96.2 |
| No special education services | 404,588 | 399,529 | 98.7 |
| Migrant | 3,870 | 3,819 | 98.7 |
| Nonmigrant | 460,138 | 452,873 | 98.4 |

Table 1.A.2 Operational Test Taking Rates for Grade Eight by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 474,860 | 463,310 | 97.6 |
| Male | 242,317 | 236,230 | 97.5 |
| Female | 232,543 | 227,080 | 97.7 |
| English learner | 59,048 | 57,420 | 97.2 |
| English only | 261,394 | 253,164 | 96.9 |
| Reclassified fluent English proficient | 133,258 | 132,030 | 99.1 |
| Initial fluent English proficient | 20,246 | 19,969 | 98.6 |
| To be determined | 524 | 441 | 84.2 |
| English proficiency unknown | 390 | 286 | 73.3 |
| Economically disadvantaged | 287,140 | 280,734 | 97.8 |
| Not economically disadvantaged | 187,720 | 182,576 | 97.3 |
| American Indian or Alaska Native | 2,504 | 2,386 | 95.3 |
| Asian | 44,167 | 43,644 | 98.8 |
| Native Hawaiian or Other Pacific Islander | 2,261 | 2,194 | 97.0 |
| Filipino | 11,054 | 10,856 | 98.2 |
| Hispanic or Latino | 259,754 | 254,829 | 98.1 |
| Black or African American | 25,427 | 24,363 | 95.8 |
| White | 109,402 | 105,546 | 96.5 |
| Two or more races | 17,498 | 16,916 | 96.7 |
| Special education services | 55,307 | 52,392 | 94.7 |
| No special education services | 419,553 | 410,918 | 97.9 |
| Migrant | 3,780 | 3,738 | 98.9 |
| Nonmigrant | 471,080 | 459,572 | 97.6 |

Table 1.A.3 Operational Test Taking Rates for Grade Ten by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 28,252 | 23,362 | 82.7 |
| Male | 14,334 | 11,944 | 83.3 |
| Female | 13,918 | 11,418 | 82.0 |
| English learner | 2,419 | 2,008 | 83.0 |
| English only | 16,394 | 13,358 | 81.5 |
| Reclassified fluent English proficient | 8,236 | 6,967 | 84.6 |
| Initial fluent English proficient | 1,153 | 987 | 85.6 |
| To be determined | 22 | 19 | 86.4 |
| English proficiency unknown | 28 | 23 | 82.1 |
| Economically disadvantaged | 16,718 | 13,876 | 83.0 |
| Not economically disadvantaged | 11,534 | 9,486 | 82.2 |
| American Indian or Alaska Native | 201 | 163 | 81.1 |
| Asian | 1,349 | 1,218 | 90.3 |
| Native Hawaiian or Other Pacific Islander | 102 | 75 | 73.5 |
| Filipino | 678 | 586 | 86.4 |
| Hispanic or Latino | 15,277 | 12,722 | 83.3 |
| Black or African American | 1,716 | 1,510 | 88.0 |
| White | 7,695 | 6,059 | 78.7 |
| Two or more races | 1,028 | 859 | 83.6 |
| Special education services | 3,220 | 2,643 | 82.1 |
| No special education services | 25,032 | 20,719 | 82.8 |
| Migrant | 243 | 174 | 71.6 |
| Nonmigrant | 28,009 | 23,188 | 82.8 |

Table 1.A.4 CAST Operational Test Taking Rates for Grade Eleven by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 272,747 | 257,323 | 94.3 |
| Male | 138,232 | 129,957 | 94.0 |
| Female | 135,515 | 127,366 | 94.0 |
| English learner | 24,874 | 22,456 | 90.3 |
| English only | 144,505 | 134,857 | 93.3 |
| Reclassified fluent English proficient | 89,857 | 86,236 | 96.0 |
| Initial fluent English proficient | 14,158 | 13,485 | 95.2 |
| To be determined | 201 | 176 | 87.6 |
| English proficiency unknown | 152 | 113 | 74.3 |
| Economically disadvantaged | 158,395 | 148,791 | 93.9 |
| Not economically disadvantaged | 115,352 | 108,532 | 94.1 |
| American Indian or Alaska Native | 1,517 | 1,390 | 91.6 |
| Asian | 24,978 | 24,210 | 96.9 |
| Native Hawaiian or Other Pacific Islander | 1,178 | 1,094 | 92.9 |
| Filipino | 7,594 | 7,315 | 96.3 |
| Hispanic or Latino | 148,019 | 139,689 | 94.4 |
| Black or African American | 13,542 | 12,242 | 90.4 |
| White | 66,734 | 62,097 | 93.1 |
| Two or more races | 8,368 | 7,659 | 91.5 |
| Special education services | 26,457 | 23,401 | 88.4 |
| No special education services | 247,290 | 233,922 | 94.6 |
| Migrant | 1,836 | 1,746 | 95.1 |
| Nonmigrant | 271,911 | 255,577 | 94.0 |

Table 1.A.5 Operational Test Taking Rates for Grade Twelve by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 323,219 | 276,687 | 85.6 |
| Male | 165,215 | 140,381 | 85.0 |
| Female | 158,004 | 136,306 | 86.3 |
| English learner | 28,517 | 21,998 | 77.1 |
| English only | 176,202 | 148,936 | 84.5 |
| Reclassified fluent English proficient | 100,384 | 90,436 | 90.1 |
| Initial fluent English proficient | 16,701 | 14,860 | 89.0 |
| To be determined | 338 | 215 | 63.6 |
| English proficiency unknown | 1,077 | 242 | 22.5 |
| Economically disadvantaged | 179,040 | 152,874 | 85.4 |
| Not economically disadvantaged | 144,179 | 123,813 | 85.9 |
| American Indian or Alaska Native | 1,754 | 1,354 | 77.2 |
| Asian | 32,016 | 29,237 | 91.3 |
| Native Hawaiian or Other Pacific Islander | 1,680 | 1,430 | 85.1 |
| Filipino | 9,763 | 8,926 | 91.4 |
| Hispanic or Latino | 165,665 | 141,675 | 85.5 |
| Black or African American | 19,726 | 15,729 | 79.7 |
| White | 79,232 | 67,421 | 85.1 |
| Two or more races | 10,719 | 9,068 | 84.6 |
| Special education services | 36,651 | 25,222 | 68.8 |
| No special education services | 286,568 | 251,465 | 87.8 |
| Migrant | 2,219 | 1,939 | 87.4 |
| Nonmigrant | 321,000 | 274,748 | 85.6 |

Table 1.A.6 Operational Test Taking Rates for High School by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | Number of Registered Students | Number of Students Taking CAST | Percent Taking CAST |
| All students | 625,214 | 557,372 | 89.1 |
| Male | 317,779 | 282,282 | 88.8 |
| Female | 307,435 | 275,090 | 89.5 |
| English learner | 55,809 | 46,462 | 83.3 |
| English only | 337,101 | 297,151 | 88.1 |
| Reclassified fluent English proficient | 198,474 | 183,639 | 92.5 |
| Initial fluent English proficient | 32,012 | 29,332 | 91.6 |
| To be determined | 561 | 410 | 73.1 |
| English proficiency unknown | 1,257 | 378 | 30.1 |
| Economically disadvantaged | 354,150 | 315,541 | 89.1 |
| Not economically disadvantaged | 271,064 | 241,831 | 89.2 |
| American Indian or Alaska Native | 3,472 | 2,907 | 83.7 |
| Asian | 58,342 | 54,665 | 93.7 |
| Native Hawaiian or Other Pacific Islander | 2,960 | 2,599 | 87.8 |
| Filipino | 18,035 | 16,827 | 93.3 |
| Hispanic or Latino | 328,958 | 294,086 | 89.4 |
| Black or African American | 34,984 | 29,481 | 84.3 |
| White | 153,661 | 135,577 | 88.2 |
| Two or more races | 20,115 | 17,586 | 87.4 |
| Special education services | 66,328 | 51,266 | 77.3 |
| No special education services | 558,886 | 506,106 | 90.6 |
| Migrant | 4,298 | 3,859 | 89.8 |
| Nonmigrant | 620,916 | 553,513 | 89.1 |

## Response-Time Analysis and Test Length

### Study Purpose

The California Science Test (CAST) was designed to be a two-hour test, with 60 minutes for the two blocks in Segment A, 40 minutes for the two performance tasks (PTs) in Segment B, and 20 minutes for either the discrete block or the performance task in Segment C. With the test administered as an online administration, the response time students spent on each item or page was collected. This analysis uses this information to evaluate whether the current test length adheres to the prescribed testing time. If the analysis reveals that the test exceeds the prescribed testing time, the item timing data could be used to inform potential modifications to the test that could be enacted to reduce the testing time.

### Testing Time for the Total Test and by Segment

Table 2.1 through Table 2.3 provides the amount of time students spent on each segment of the test and on the total test from the 2018–19 administration for grade five, grade eight, and high school respectively. For example, the 65 minutes listed for the twenty-fifth percentile for Segment A means that 25 percent of the students spent 65 minutes or less on Segment A. Table 2.1 provides this information for grade five.

Table 2.1 Time Spent (in Minutes) on the CAST for Grade Five

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | 25th Percentile | 50th Percentile | 75th Percentile |
| Segment A | 65 | 88 | 122 |
| Segment B | 16 | 23 | 32 |
| Segment C Discrete Block | 15 | 22 | 32 |
| Segment C Performance Task | 8 | 13 | 19 |
| Total Test with Discrete Field Test Block | 100 | 134 | 181 |
| Total Test with Performance Task Field Test Block | 97 | 130 | 177 |

Table 2.2 provides similar information for grade eight.

Table 2.2 Time Spent (in Minutes) on the CAST for Grade Eight

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | 25th Percentile | 50th Percentile | 75th Percentile |
| Segment A | 58 | 77 | 104 |
| Segment B | 10 | 16 | 23 |
| Segment C Discrete Block | 14 | 21 | 30 |
| Segment C Performance Task | 5 | 8 | 12 |
| Total Test with Discrete Field Test Block | 86 | 114 | 150 |
| Total Test with Performance Task Field Test Block | 81 | 107 | 140 |

Table 2.3 provides similar information for high school.

Table 2.3 Time Spent (in Minutes) on the CAST for High School

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | 25th Percentile | 50th Percentile | 75th Percentile |
| Segment A | 32 | 46 | 61 |
| Segment B | 6 | 9 | 14 |
| Segment C Discrete Block | 7 | 12 | 18 |
| Segment C Performance Task | 2 | 5 | 8 |
| Total Test with Discrete Field Test Block | 47 | 67 | 89 |
| Total Test with Performance Task Field Test Block | 44 | 63 | 85 |

Table 2.A.1 through Table 2.A.18 in appendix 2.A provide testing time information by student group. There was substantial variation in the testing times across the student groups with the largest differences in grade five and grade eight. Among students assigned a CAST with a discrete field test block, students who were not economically disadvantaged, Asian, White, and from two or more races had median testing times within the targeted testing time of two hours in grade five.

In grade eight, all student groups except reclassified fluent English proficient (RFEP), economically disadvantaged, Hispanic or Latino, and migrant had median testing times within two hours. For students that were assigned a CAST with a performance task field test block, students from two or more races, Asian, White, and students who require special education services had median testing times within two hours in grade five. In grade eight, all student groups except migrants and RFEP had median testing times of two hours or less.

The three major findings reflected in Table 2.1 through Table 2.3 include the following:

1. With a similar number of items on the test, it took longer for students in grade five and grade eight to complete the total test than it took students from high school. The median testing time (i.e., fiftieth percentile) for grade five is 134 minutes for students who received the discrete Segment C block and 130 minutes for students who received the PT Segment C block. For grade eight, the median time is 114 minutes for discrete Segment C and 107 for PT Segment C block. For high school, more than 75 percent of students completed the test within two hours.
2. The median time spent on Segment A in grade five and grade eight exceeded 60 minutes. On Segment B, the median time spent was shorter than the expected 40 minutes. For example, half of the grade five students took longer than 89 minutes to complete Segment A, while half of the students completed two PTs in Segment B in 23 minutes.
3. Completion time on Segment C was longer for those students who received the discrete item block versus those who received a PT. This is most likely due to the difference in the number of items in the discrete block (13 items) versus a PT (five to eight items).

These results suggest that the test length is likely too long for grade five and grade eight. A recommendation is made to reduce the number of items for grades five and eight while still providing reliable scores and appropriate content coverage to meet a revised test blueprint. Given that the numbers of PEs assessed in grade five and grade eight are lower than in high school (45 for grade five, 59 for grade eight, and 71 for high school), shortening the test for grades five and eight will still provide PE coverage within a three-year period. (The blueprint specifies that all PEs be assessed within a three-year period.)

The original test blueprint was based on an estimate of 20 minutes to complete one PT in Segment B. Therefore, a student could only receive a maximum of two PTs from two of the three science domains in Segment B, which results in a variance in the number of items reported for each domain. Given that students did not take as long as the expected 40 minutes on two PTs for all grades, it is reasonable to add a third PT to Segment B, which will provide students balance across each of the three science domains. With this improvement, all students’ score reporting will be based on the same representation of items from each of the science domains.

The number of discrete items in Segment A can be reduced for grades five and eight while still preserving the test reliability and content coverage. The number of discrete items in Segment A should not be reduced for high school due to the number of PEs that need to be assessed in a three-year period. The reduction of discrete items in Segment C for all grade levels can also help balance the response time for students who receive the discrete Segment C block and those who receive the PT Segment C block.

Even though the response time results suggest there were no testing time concerns for the high school assessment, with the addition of a third PT in Segment B, a recommendation is made to reduce the number of items in Segment C to be consistent with grade five and grade eight, and to keep the time similar for students who receive the discrete and those who receive the PT in Segment C.

### Proposed Test Length

Based on the response time results from the 2018–19 administration, an effort was made to evaluate a change in the current test length that meets the following guidelines:

1. Test length is appropriate so that most students can finish the test within two hours.
2. The content coverage is not impacted, (i.e., all PEs can still be covered with a three-year cycle as described in the current test blueprint).
3. Test reliability is acceptable.

Using these guidelines and the response time by segments and by item type, ETS estimated the time for students to complete the test under several blueprint change options and recommended the test length as described in Table 2.4.

Table 2.4 Proposed Test Length

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Information Type | Current Blueprint | Proposed Blueprint Grade Five | Proposed Blueprint Grade Eight | Proposed Blueprint High School |
| Segment A | 32–34 items | 26 items | 28 items | 32 items |
| Segment B | 2 PTs  (8–12 items) | 3 PTs  (12–18 items) | 3 PTs  (12–18 items) | 3 PTs  (12–18 items) |

Table 2.4 *(continuation*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Information Type | Current Blueprint | Proposed Blueprint Grade Five | Proposed Blueprint Grade Eight | Proposed Blueprint High School |
| Segment C Discrete | 12–14 items | 6 items | 6 items | 6 items |
| Segment C PT | 1 PT  4–7 items | 1 PT  6 items | 1 PT  6 items | 1 PT  6 items |
| Total Number of Operational items | 40–46 | 38–44 | 40–46 | 44–50 |
| Total Number of Items Discrete C Block | 52–60 | 44–50 | 46–52 | 50–56 |
| Total Number of Items PT C Block | 44–53 | 44–50 | 46–52 | 50–56 |

### Evaluation of the Proposed Test Length

The evaluation of the proposed test length is focused on estimating time for students to complete the test with the proposed test length, the content coverage, and the test reliability under the proposed blueprint.

#### Estimated Response Time for the Proposed Blueprint

Because the response time can vary significantly by item type, to estimate the time for students to complete the test with the proposed blueprint, the response time information by item type from the 2018–19 administration was used.

Table 2.5 outlines the major categories of item types that were included in the CAST operational administration. This includes item types ranging from traditional multiple choice (MC) and constructed response (CR) (i.e., extended text) to new technology-enhanced item (TEI) types (the rest of the item types). In addition to the item types included in Table 2.6, CAST also used composite items in the test. The composite items usually include two parts, possibly with different item types.

Table 2.5 Selected Item Type Categories in the CAST Operational Administration

|  |  |
| --- | --- |
| Feature | Description |
| **Choice** | Traditional single-select or multiple-select MC items or grid items |
| **Extended Text** | Traditional essay or other CR items, where the student provides a text response |
| **Hot Spot** | Items that present a graphic—such as an anatomical diagram or a drawing of laboratory equipment—where a student selects a part of the graphic as the response (Example: Zone) |

Table 2.5 *(continuation)*

|  |  |
| --- | --- |
| Feature | Description |
| **Match** | Items that present multiple pieces of evidence for a student to match to each of various alternate conclusions, and items that present a grid with row and column headings (e.g., representing alternate experimental designs to address alternate hypotheses), where a student selects table cells as the response to indicate which experimental design is appropriate to test each hypothesis |
| **Inline Choice** | Items that provide multiple choices for filling in one or more blanks within a sentence or paragraph |
| **Composite** | Two-part items that include one traditional MC item part and one non-MC, non-CR item (such as an inline choice list or grid) part |
| **Custom** | Items where a student manipulates an object, such as a scale, a histogram, a clock, or an arrangement of laboratory materials; a collection of interactive items and custom interactive stimuli in a set with multiple-scored interactive components, such as simulations (Examples: bar-picturegraph, interactive) |

Table 2.6 shows the median response time for all the item types used in grade five, grade eight, and high school. A composite CR item refers to an item that has two parts, one part being the CR item.

Table 2.6 Median Response Time by Item Type

|  |  |  |  |
| --- | --- | --- | --- |
| Item Type | Grade Five | Grade Eight | High School |
| Composite CR | 6.14 | 3.73 | NA |
| ExtendedText (i.e., CR Items) | 4.47 | 4.58 | 2.35 |
| Composite | 2.60 | 2.26 | 1.06 |
| Match Multiple-Selection | 1.63 | 0.95 | 0.70 |
| Multiple-Choice Single-Selection | 1.37 | 1.16 | 0.76 |
| Zone Single-Selection | 1.30 | 1.60 | 0.80 |
| Zone Multiple-Selection | 1.28 | 1.20 | 0.94 |
| Grid Multiple-Selection | 1.25 | 1.19 | 0.54 |
| Multiple-Choice Multiple-Selection | 1.10 | 1.28 | 0.90 |
| Bar-Picturegraph Multiple-Selection | 1.07 | 0.99 | NA |
| InlineChoice Single-Selection | 1.04 | 1.01 | 0.98 |
| InlineChoice Multiple-Selection | 0.98 | 1.00 | 1.02 |
| Match Single-Selection | 0.76 | 1.13 | 0.65 |
| Interactive | NA | 1.60 | 0.67 |

As shown in Table 2.6, the composite CR item, the extended text (i.e., the CR item), and the composite items took much longer than other item types, especially for lower grades. These results are as expected because the composite items include two parts, and it is expected to take longer to respond to these. The number of composite, composite CR, and CR items in a test form is usually limited and does not vary much from form to form. Most items in a test form are from the other item types, and the number of items from each of these other item types may vary from form to form. Because the response time for these other item types were very similar, to make the estimation more robust, they were grouped into one category, and the average median response time was used as the estimated time to finish an item in this category.

For each segment, the reduction (or addition) of time can then be estimated by the reduction (or addition) of the number of composite items, composite CR items, extended text items, and items of other item types. Table 2.7 compares the time students spent on the test from the 2018–19 administration to the estimated time for the proposed blueprint.

Table 2.7 Response Time Comparison for the Current and Proposed Blueprints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade and Blueprint | 50th Percentile (in Minutes) | 60th Percentile (in Minutes) | 75th Percentile (in Minutes) | Estimated Percentage of Students to Complete Test in Two Hours |
| Grade Five Current | 132 | 148 | 179 | 42 |
| Grade Five Proposed | 114 | 128 | 154 | 54 |
| Grade Eight Current | 111 | 122 | 145 | 58 |
| Grade Eight Proposed | 94 | 105 | 124 | 72 |
| High School Current | 65 | 73 | 87 | 92 |
| High School Proposed | 60 | 68 | 82 | 94 |

As shown in Table 2.7, there is a significant reduction in time to complete the test. For example, with the current blueprint, 75 percent of the grade five students completed the test within 179 minutes or less. With the proposed blueprint, 75 percent of the students can finish the test within 154 minutes, a reduction of 25 minutes. For grade eight, 75 percent of the students can finish the test within 124 minutes or less, a reduction of 21 minutes. Moreover, as the percentage of students who can finish the test within two hours is expected to increase in all grades, with substantial increases in grade five and grade eight. (a 12 percent increase for grade five, a 14 percent increase for grade eight, and a 2 percent increase for high school). As CA NGSS implementation gains momentum, ETS will continue to monitor the time it takes students to complete the test.

#### Content Coverage

The number of PEs to cover varies by grade. There are 45 PEs for grade five, 59 for grade eight, and 71 for high school. The proposed blueprint includes 38 to 44 operational items for grade five, 40 to 46 items for grade eight, and 44 to 50 items for high school. The proposed test length is enough to cover all PEs across a three-year cycle as proposed in the high‑level test design.

Table 2.8 summarizes the number of the unique PEs covered in 2018–19 forms, the anticipated number of PEs covered in the 2019–20 forms[[1]](#footnote-2), and the planned 2020–21 forms.

Table 2.8 Anticipated Number of Unique PEs Assessed on Segments A and B of the CAST

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade Level and Number of Unique PEs | 2018–19 Forms | 2019–20 Forms | 2020–21 Forms | Total for Three Operational Administrations |
| Grade Five (45 PEs) | 34 | 10 | 1 | 45 |
| Grade Eight (59 PEs) | 39 | 14 | 6 | 59 |
| High School (71 PEs) | 35 | 16 | 20 | 71 |

#### Test Reliability

Table 2.9 provides reliabilities for the 2018–19 operational forms and the estimated reliability of forms developed with the number of discrete items and PTs proposed in the blueprint change in Table 2.4. As shown in Table 2.9, the change in the number of discrete items and PTs results in either negligible or no reduction of test reliability for each grade.

Table 2.9 A Comparison of the Reliability of the Current and Proposed Blueprints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade | Average Number of OP Items—Current Blueprint | Average Number of OP Items—Proposed Blueprint | Reliability—Current Blueprint | Reliability—Proposed Blueprint |
| Five | 46 | 41 | 0.91 | 0.90 |
| Eight | 44 | 43 | 0.89 | 0.89 |
| High School | 46 | 47 | 0.88 | 0.88 |

Table 2.10 provides domain-specific reliabilities for the current and proposed blueprints. We note that the reliabilities for Earth and Space Sciences are slightly lower for all grades. The number of items in Earth and Space Sciences is smaller than other domains; reliability largely depends on the number of items. Compared to the current blueprint, the reliabilities for the proposed blueprint are slightly lower in grade five and remain about the same for grade eight and high school for all domains.

Table 2.10 A Comparison of the Reliability of the Current and Recommended Blueprints by Domain

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Domain | Grade Five Reliability—Current Blueprint | Grade Five Reliability—Proposed Blueprint | Grade Eight Reliability—Current Blueprint | Grade Eight Reliability—Proposed Blueprint | High School Reliability—Current Blueprint | High School Reliability—Proposed Blueprint |
| Life Sciences | 0.75 | 0.73 | 0.71 | 0.70 | 0.71 | 0.71 |
| Physical Sciences | 0.79 | 0.77 | 0.77 | 0.77 | 0.72 | 0.72 |
| Earth and Space Sciences | 0.76 | 0.74 | 0.66 | 0.65 | 0.68 | 0.69 |

### Summary

Based on the response time information from the 2018–19 operational administration, ETS recommends a reduction in the number of discrete items in Segment A (for grade five and grade eight) and Segment C (for all three grades), and that a third PT in Segment B be added for all grades. There are three key benefits with this proposed blueprint:

* The number of students in grade five and grade eight who can complete the test within two hours is expected to increase.
* Each student will now receive three PTs in Segment B representing each of three science domains.
* The time needed to complete Segment C discrete items is reduced and more comparable with the time needed to complete a Segment C PT.

### Appendix 2.A. Testing Time by Student Group

Table 2.A.1 Time Spent (in Minutes) on the CAST for Grade Five for Segment A by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 65 | 88 | 122 |
| Male | 61 | 84 | 117 |
| Female | 69 | 94 | 128 |
| English learner | 71 | 100 | 138 |
| English only | 62 | 83 | 114 |
| Reclassified fluent English proficient | 74 | 98 | 134 |
| Initial fluent English proficient | 60 | 80 | 111 |
| To be determined | 51 | 77 | 109 |
| English proficiency unknown | 52 | 76 | 99 |
| Economically disadvantaged | 71 | 97 | 133 |
| Not economically disadvantaged | 59 | 77 | 104 |
| American Indian or Alaska Native | 62 | 85 | 114 |
| Asian | 57 | 75 | 100 |
| Native Hawaiian or Other Pacific Islander | 64 | 87 | 119 |
| Filipino | 64 | 84 | 112 |
| Hispanic or Latino | 73 | 99 | 136 |
| Black or African American | 64 | 89 | 125 |
| White | 58 | 76 | 101 |
| Two or more races | 58 | 76 | 103 |
| Special education services | 55 | 80 | 115 |
| No special education services | 66 | 89 | 123 |
| Migrant | 80 | 108 | 149 |
| Nonmigrant | 65 | 88 | 122 |

Table 2.A.2 Time Spent (in Minutes) on the CAST for Grade Five for Segment B by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 16 | 23 | 32 |
| Male | 14 | 21 | 30 |
| Female | 17 | 24 | 34 |
| English learner | 14 | 22 | 32 |
| English only | 15 | 22 | 31 |
| Reclassified fluent English proficient | 18 | 26 | 36 |
| Initial fluent English proficient | 17 | 23 | 32 |
| To be determined | 10 | 18 | 29 |
| English proficiency unknown | 12 | 18 | 27 |
| Economically disadvantaged | 15 | 23 | 33 |
| Not economically disadvantaged | 16 | 22 | 30 |
| American Indian or Alaska Native | 13 | 20 | 30 |
| Asian | 16 | 22 | 30 |
| Native Hawaiian or Other Pacific Islander | 14 | 21 | 30 |
| Filipino | 17 | 23 | 32 |
| Hispanic or Latino | 16 | 24 | 34 |
| Black or African American | 12 | 20 | 31 |
| White | 15 | 21 | 29 |
| Two or more races | 15 | 21 | 29 |
| Special education services | 11 | 19 | 29 |
| No special education services | 16 | 23 | 33 |
| Migrant | 16 | 25 | 37 |
| Nonmigrant | 15 | 23 | 32 |

Table 2.A.3 Time Spent (in Minutes) on the CAST for Grade Five for Segment C Discrete Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 15 | 22 | 32 |
| Male | 14 | 21 | 30 |
| Female | 16 | 23 | 33 |
| English learner | 15 | 24 | 36 |
| English only | 15 | 21 | 30 |
| Reclassified fluent English proficient | 18 | 25 | 35 |
| Initial fluent English proficient | 15 | 20 | 29 |
| To be determined | 9 | 16 | 29 |
| English proficiency unknown | 12 | 18 | 25 |
| Economically disadvantaged | 16 | 24 | 35 |
| Not economically disadvantaged | 14 | 20 | 28 |
| American Indian or Alaska Native | 13 | 20 | 29 |
| Asian | 14 | 19 | 27 |
| Native Hawaiian or Other Pacific Islander | 14 | 21 | 32 |
| Filipino | 16 | 22 | 30 |
| Hispanic or Latino | 16 | 24 | 35 |
| Black or African American | 13 | 21 | 32 |
| White | 14 | 19 | 27 |
| Two or more races | 14 | 19 | 27 |
| Special education services | 12 | 19 | 30 |
| No special education services | 16 | 22 | 32 |
| Migrant | 17 | 26 | 40 |
| Nonmigrant | 15 | 22 | 32 |

Table 2.A.4 Time Spent (in Minutes) on the CAST for Grade Five for Segment C Performance Task by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 8 | 13 | 19 |
| Male | 8 | 12 | 18 |
| Female | 9 | 14 | 20 |
| English learner | 7 | 13 | 19 |
| English only | 8 | 12 | 18 |
| Reclassified fluent English proficient | 10 | 15 | 22 |
| Initial fluent English proficient | 9 | 13 | 19 |
| To be determined | 4 | 10 | 17 |
| English proficiency unknown | 6 | 10 | 15 |
| Economically disadvantaged | 8 | 13 | 20 |
| Not economically disadvantaged | 9 | 12 | 17 |
| American Indian or Alaska Native | 7 | 11 | 18 |
| Asian | 9 | 12 | 17 |
| Native Hawaiian or Other Pacific Islander | 7 | 11 | 17 |
| Filipino | 9 | 13 | 19 |
| Hispanic or Latino | 9 | 14 | 21 |
| Black or African American | 6 | 11 | 18 |
| White | 8 | 12 | 17 |
| Two or more races | 8 | 12 | 17 |
| Special education services | 6 | 11 | 17 |
| No special education services | 9 | 13 | 19 |
| Migrant | 9 | 15 | 22 |
| Nonmigrant | 8 | 13 | 19 |

Table 2.A.5 Time Spent (in Minutes) on the CAST for Grade Five for Total Test with Discrete Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 100 | 134 | 181 |
| Male | 94 | 127 | 172 |
| Female | 107 | 142 | 190 |
| English learner | 105 | 146 | 199 |
| English only | 95 | 126 | 169 |
| Reclassified fluent English proficient | 114 | 149 | 200 |
| Initial fluent English proficient | 94 | 125 | 168 |
| To be determined | 80 | 112 | 161 |
| English proficiency unknown | 80 | 111 | 150 |
| Economically disadvantaged | 106 | 145 | 195 |
| Not economically disadvantaged | 92 | 120 | 157 |
| American Indian or Alaska Native | 91 | 124 | 167 |
| Asian | 89 | 116 | 152 |
| Native Hawaiian or Other Pacific Islander | 97 | 131 | 173 |
| Filipino | 101 | 130 | 169 |
| Hispanic or Latino | 110 | 148 | 199 |
| Black or African American | 94 | 132 | 180 |
| White | 90 | 117 | 153 |
| Two or more races | 90 | 118 | 156 |
| Special education services | 82 | 120 | 169 |
| No special education services | 102 | 136 | 182 |
| Migrant | 115 | 157 | 214 |
| Nonmigrant | 100 | 134 | 180 |

Table 2.A.6 Time Spent (in Minutes) on the CAST for Grade Five for Total Test with Performance Task Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 97 | 130 | 177 |
| Male | 91 | 123 | 168 |
| Female | 104 | 138 | 186 |
| English learner | 102 | 142 | 194 |
| English only | 92 | 123 | 165 |
| Reclassified fluent English proficient | 110 | 146 | 196 |
| Initial fluent English proficient | 92 | 122 | 164 |
| To be determined | 70 | 111 | 150 |
| English proficiency unknown | 81 | 113 | 138 |
| Economically disadvantaged | 103 | 141 | 191 |
| Not economically disadvantaged | 89 | 116 | 153 |
| American Indian or Alaska Native | 91 | 123 | 167 |
| Asian | 86 | 113 | 150 |
| Native Hawaiian or Other Pacific Islander | 95 | 126 | 172 |
| Filipino | 96 | 124 | 164 |
| Hispanic or Latino | 107 | 144 | 194 |
| Black or African American | 92 | 129 | 177 |
| White | 87 | 114 | 149 |
| Two or more races | 86 | 114 | 149 |
| Special education services | 80 | 116 | 164 |
| No special education services | 99 | 132 | 178 |
| Migrant | 118 | 160 | 210 |
| Nonmigrant | 96 | 130 | 177 |

Table 2.A.7 Time Spent (in Minutes) on the CAST for Grade Eight for Segment A by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 58 | 77 | 104 |
| Male | 54 | 74 | 100 |
| Female | 61 | 81 | 107 |
| English learner | 61 | 87 | 117 |
| English only | 54 | 71 | 93 |
| Reclassified fluent English proficient | 67 | 88 | 116 |
| Initial fluent English proficient | 57 | 74 | 97 |
| To be determined | 48 | 72 | 95 |
| English proficiency unknown | 46 | 61 | 84 |
| Economically disadvantaged | 62 | 85 | 113 |
| Not economically disadvantaged | 53 | 68 | 88 |
| American Indian or Alaska Native | 51 | 71 | 96 |
| Asian | 56 | 72 | 92 |
| Native Hawaiian or Other Pacific Islander | 56 | 76 | 100 |
| Filipino | 60 | 75 | 96 |
| Hispanic or Latino | 64 | 86 | 115 |
| Black or African American | 56 | 78 | 106 |
| White | 50 | 65 | 83 |
| Two or more races | 52 | 67 | 86 |
| Special education services | 49 | 73 | 103 |
| No special education services | 58 | 78 | 104 |
| Migrant | 70 | 95 | 122 |
| Nonmigrant | 57 | 77 | 103 |

Table 2.A.8 Time Spent (in Minutes) on the CAST for Grade Eight for Segment B by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 10 | 16 | 23 |
| Male | 9 | 15 | 22 |
| Female | 11 | 17 | 25 |
| English learner | 7 | 12 | 21 |
| English only | 10 | 16 | 22 |
| Reclassified fluent English proficient | 12 | 18 | 26 |
| Initial fluent English proficient | 13 | 18 | 25 |
| To be determined | 6 | 12 | 22 |
| English proficiency unknown | 7 | 12 | 19 |
| Economically disadvantaged | 9 | 16 | 24 |
| Not economically disadvantaged | 11 | 17 | 23 |
| American Indian or Alaska Native | 8 | 14 | 21 |
| Asian | 13 | 18 | 25 |
| Native Hawaiian or Other Pacific Islander | 8 | 14 | 21 |
| Filipino | 13 | 18 | 25 |
| Hispanic or Latino | 9 | 16 | 24 |
| Black or African American | 6 | 13 | 21 |
| White | 11 | 16 | 21 |
| Two or more races | 11 | 16 | 22 |
| Special education services | 6 | 12 | 20 |
| No special education services | 11 | 16 | 24 |
| Migrant | 10 | 17 | 25 |
| Nonmigrant | 10 | 16 | 23 |

Table 2.A.9 Time Spent (in Minutes) on the CAST for Grade Eight for Segment C Discrete Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 14 | 21 | 30 |
| Male | 13 | 20 | 28 |
| Female | 15 | 22 | 31 |
| English learner | 10 | 19 | 31 |
| English only | 14 | 20 | 27 |
| Reclassified fluent English proficient | 16 | 24 | 34 |
| Initial fluent English proficient | 15 | 21 | 29 |
| To be determined | 6 | 16 | 25 |
| English proficiency unknown | 10 | 16 | 24 |
| Economically disadvantaged | 13 | 22 | 32 |
| Not economically disadvantaged | 14 | 20 | 27 |
| American Indian or Alaska Native | 11 | 19 | 27 |
| Asian | 15 | 21 | 28 |
| Native Hawaiian or Other Pacific Islander | 12 | 19 | 29 |
| Filipino | 16 | 22 | 29 |
| Hispanic or Latino | 14 | 22 | 32 |
| Black or African American | 10 | 19 | 29 |
| White | 14 | 19 | 25 |
| Two or more races | 14 | 19 | 26 |
| Special education services | 9 | 17 | 27 |
| No special education services | 15 | 21 | 30 |
| Migrant | 15 | 24 | 35 |
| Nonmigrant | 14 | 21 | 30 |

Table 2.A.10 Time Spent (in Minutes) on the CAST for Grade Eight for Segment C Performance Task by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 5 | 8 | 12 |
| Male | 5 | 8 | 11 |
| Female | 6 | 9 | 13 |
| English learner | 4 | 7 | 11 |
| English only | 5 | 8 | 11 |
| Reclassified fluent English proficient | 6 | 9 | 13 |
| Initial fluent English proficient | 6 | 9 | 12 |
| To be determined | 4 | 6 | 11 |
| English proficiency unknown | 4 | 7 | 11 |
| Economically disadvantaged | 5 | 8 | 12 |
| Not economically disadvantaged | 6 | 8 | 11 |
| American Indian or Alaska Native | 4 | 7 | 10 |
| Asian | 7 | 9 | 13 |
| Native Hawaiian or Other Pacific Islander | 4 | 8 | 11 |
| Filipino | 6 | 9 | 12 |
| Hispanic or Latino | 5 | 8 | 13 |
| Black or African American | 3 | 7 | 11 |
| White | 5 | 8 | 11 |
| Two or more races | 5 | 8 | 11 |
| Special education services | 3 | 6 | 10 |
| No special education services | 6 | 8 | 12 |
| Migrant | 5 | 9 | 14 |
| Nonmigrant | 5 | 8 | 12 |

Table 2.A.11 Time Spent (in Minutes) on the CAST for Grade Eight for Total Test with Discrete Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 86 | 114 | 150 |
| Male | 81 | 108 | 143 |
| Female | 92 | 120 | 156 |
| English learner | 86 | 120 | 160 |
| English only | 81 | 106 | 137 |
| Reclassified fluent English proficient | 100 | 130 | 168 |
| Initial fluent English proficient | 88 | 113 | 145 |
| To be determined | 68 | 103 | 139 |
| English proficiency unknown | 69 | 94 | 124 |
| Economically disadvantaged | 91 | 122 | 161 |
| Not economically disadvantaged | 82 | 104 | 133 |
| American Indian or Alaska Native | 74 | 107 | 141 |
| Asian | 88 | 110 | 141 |
| Native Hawaiian or Other Pacific Islander | 81 | 108 | 142 |
| Filipino | 92 | 116 | 144 |
| Hispanic or Latino | 94 | 125 | 163 |
| Black or African American | 79 | 110 | 147 |
| White | 78 | 99 | 125 |
| Two or more races | 79 | 101 | 128 |
| Special education services | 69 | 103 | 144 |
| No special education services | 88 | 115 | 150 |
| Migrant | 103 | 135 | 172 |
| Nonmigrant | 86 | 114 | 149 |

Table 2.A.12 Time Spent (in Minutes) on the CAST for Grade Eight for Total Test with Performance Task Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 81 | 107 | 140 |
| Male | 76 | 101 | 134 |
| Female | 86 | 113 | 146 |
| English learner | 81 | 114 | 152 |
| English only | 76 | 99 | 128 |
| Reclassified fluent English proficient | 93 | 122 | 158 |
| Initial fluent English proficient | 82 | 105 | 136 |
| To be determined | 69 | 96 | 131 |
| English proficiency unknown | 62 | 84 | 118 |
| Economically disadvantaged | 85 | 115 | 151 |
| Not economically disadvantaged | 76 | 97 | 123 |
| American Indian or Alaska Native | 69 | 95 | 126 |
| Asian | 81 | 103 | 131 |
| Native Hawaiian or Other Pacific Islander | 78 | 105 | 134 |
| Filipino | 85 | 107 | 133 |
| Hispanic or Latino | 88 | 117 | 154 |
| Black or African American | 75 | 105 | 140 |
| White | 72 | 92 | 116 |
| Two or more races | 74 | 95 | 120 |
| Special education services | 65 | 97 | 135 |
| No special education services | 82 | 108 | 141 |
| Migrant | 95 | 127 | 162 |
| Nonmigrant | 80 | 107 | 140 |

Table 2.A.13 Time Spent (in Minutes) on the CAST for High School for Segment A by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 32 | 46 | 61 |
| Male | 30 | 44 | 60 |
| Female | 34 | 47 | 63 |
| English learner | 31 | 48 | 68 |
| English only | 29 | 42 | 56 |
| Reclassified fluent English proficient | 36 | 51 | 69 |
| Initial fluent English proficient | 34 | 47 | 62 |
| To be determined | 34 | 49 | 66 |
| English proficiency unknown | 30 | 43 | 59 |
| Economically disadvantaged | 33 | 48 | 66 |
| Not economically disadvantaged | 31 | 43 | 56 |
| American Indian or Alaska Native | 27 | 41 | 57 |
| Asian | 35 | 47 | 60 |
| Native Hawaiian or Other Pacific Islander | 28 | 43 | 59 |
| Filipino | 36 | 48 | 61 |
| Hispanic or Latino | 34 | 49 | 67 |
| Black or African American | 26 | 42 | 60 |
| White | 28 | 40 | 53 |
| Two or more races | 29 | 42 | 55 |
| Special education services | 25 | 41 | 61 |
| No special education services | 32 | 46 | 61 |
| Migrant | 38 | 55 | 74 |
| Nonmigrant | 32 | 46 | 61 |

Table 2.A.14 Time Spent (in Minutes) on the CAST for High School for Segment B by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 6 | 9 | 14 |
| Male | 5 | 9 | 13 |
| Female | 6 | 10 | 14 |
| English learner | 5 | 8 | 13 |
| English only | 5 | 9 | 13 |
| Reclassified fluent English proficient | 6 | 10 | 15 |
| Initial fluent English proficient | 7 | 10 | 15 |
| To be determined | 5 | 8 | 14 |
| English proficiency unknown | 5 | 8 | 12 |
| Economically disadvantaged | 5 | 9 | 14 |
| Not economically disadvantaged | 6 | 9 | 13 |
| American Indian or Alaska Native | 5 | 8 | 13 |
| Asian | 7 | 11 | 15 |
| Native Hawaiian or Other Pacific Islander | 5 | 8 | 13 |
| Filipino | 7 | 10 | 15 |
| Hispanic or Latino | 5 | 9 | 14 |
| Black or African American | 4 | 7 | 12 |
| White | 6 | 9 | 13 |
| Two or more races | 6 | 9 | 13 |
| Special education services | 4 | 7 | 12 |
| No special education services | 6 | 9 | 14 |
| Migrant | 6 | 10 | 16 |
| Nonmigrant | 6 | 9 | 14 |

Table 2.A.15 Time Spent (in Minutes) on the CAST for High School for Segment C Discrete Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 7 | 12 | 18 |
| Male | 6 | 12 | 17 |
| Female | 8 | 13 | 18 |
| English learner | 5 | 10 | 17 |
| English only | 7 | 12 | 17 |
| Reclassified fluent English proficient | 8 | 14 | 20 |
| Initial fluent English proficient | 9 | 13 | 19 |
| To be determined | 6 | 11 | 18 |
| English proficiency unknown | 6 | 10 | 16 |
| Economically disadvantaged | 7 | 12 | 19 |
| Not economically disadvantaged | 8 | 12 | 17 |
| American Indian or Alaska Native | 6 | 11 | 17 |
| Asian | 9 | 14 | 18 |
| Native Hawaiian or Other Pacific Islander | 5 | 11 | 17 |
| Filipino | 9 | 13 | 19 |
| Hispanic or Latino | 7 | 12 | 19 |
| Black or African American | 5 | 10 | 17 |
| White | 7 | 12 | 16 |
| Two or more races | 7 | 12 | 17 |
| Special education services | 5 | 9 | 16 |
| No special education services | 7 | 13 | 18 |
| Migrant | 7 | 14 | 21 |
| Nonmigrant | 7 | 12 | 18 |

Table 2.A.16 Time Spent (in Minutes) on the CAST for High School for Segment C Performance Task by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 2 | 5 | 8 |
| Male | 2 | 5 | 8 |
| Female | 3 | 5 | 8 |
| English learner | 2 | 4 | 7 |
| English only | 2 | 5 | 8 |
| Reclassified fluent English proficient | 3 | 5 | 9 |
| Initial fluent English proficient | 3 | 6 | 9 |
| To be determined | 2 | 4 | 7 |
| English proficiency unknown | 2 | 5 | 8 |
| Economically disadvantaged | 2 | 5 | 8 |
| Not economically disadvantaged | 3 | 5 | 8 |
| American Indian or Alaska Native | 2 | 4 | 7 |
| Asian | 4 | 6 | 9 |
| Native Hawaiian or Other Pacific Islander | 2 | 4 | 7 |
| Filipino | 4 | 6 | 8 |
| Hispanic or Latino | 2 | 5 | 8 |
| Black or African American | 2 | 3 | 7 |
| White | 3 | 5 | 8 |
| Two or more races | 3 | 5 | 8 |
| Special education services | 2 | 3 | 6 |
| No special education services | 3 | 5 | 8 |
| Migrant | 3 | 5 | 9 |
| Nonmigrant | 2 | 5 | 8 |

Table 2.A.17 Time Spent (in Minutes) on the CAST for High School for Total Test with Discrete Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 47 | 67 | 89 |
| Male | 44 | 65 | 87 |
| Female | 50 | 70 | 92 |
| English learner | 44 | 67 | 94 |
| English only | 43 | 63 | 83 |
| Reclassified fluent English proficient | 53 | 74 | 99 |
| Initial fluent English proficient | 52 | 71 | 92 |
| To be determined | 48 | 67 | 93 |
| English proficiency unknown | 44 | 58 | 86 |
| Economically disadvantaged | 47 | 70 | 95 |
| Not economically disadvantaged | 46 | 64 | 83 |
| American Indian or Alaska Native | 39 | 60 | 84 |
| Asian | 53 | 71 | 90 |
| Native Hawaiian or Other Pacific Islander | 40 | 61 | 88 |
| Filipino | 53 | 71 | 90 |
| Hispanic or Latino | 48 | 70 | 95 |
| Black or African American | 37 | 60 | 85 |
| White | 43 | 61 | 79 |
| Two or more races | 44 | 63 | 82 |
| Special education services | 36 | 59 | 86 |
| No special education services | 48 | 68 | 89 |
| Migrant | 56 | 81 | 107 |
| Nonmigrant | 47 | 67 | 89 |

Table 2.A.18 Time Spent (in Minutes) on the CAST for High School for Total Test with Performance Task Field Test Block by Student Group

|  |  |  |  |
| --- | --- | --- | --- |
| Student Group | 25th Percentile | 50th Percentile | 75th Percentile |
| All students | 44 | 63 | 85 |
| Male | 41 | 61 | 82 |
| Female | 47 | 66 | 87 |
| English learner | 42 | 64 | 89 |
| English only | 41 | 59 | 78 |
| Reclassified fluent English proficient | 50 | 71 | 94 |
| Initial fluent English proficient | 49 | 67 | 88 |
| To be determined | 47 | 65 | 90 |
| English proficiency unknown | 41 | 62 | 78 |
| Economically disadvantaged | 45 | 66 | 90 |
| Not economically disadvantaged | 43 | 60 | 79 |
| American Indian or Alaska Native | 39 | 59 | 79 |
| Asian | 50 | 67 | 85 |
| Native Hawaiian or Other Pacific Islander | 38 | 59 | 80 |
| Filipino | 50 | 67 | 86 |
| Hispanic or Latino | 46 | 67 | 91 |
| Black or African American | 36 | 57 | 81 |
| White | 40 | 57 | 74 |
| Two or more races | 41 | 58 | 76 |
| Special education services | 35 | 56 | 82 |
| No special education services | 45 | 64 | 85 |
| Migrant | 52 | 75 | 99 |
| Nonmigrant | 44 | 63 | 85 |

## Test Dimensionality

### Study Purpose

The standards in the California Next Generation Science Standards (CA NGSS) are considered “three dimensional” (3D) because of the interrelationships between the disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs). The California Science Test (CAST) is designed to reflect a commitment to the 3D approach in both the writing of the test items, all of which are aligned with at least two of the three dimensions; and in the assembly of test forms, which is directed by the CAST blueprint.

A number of questions need to be addressed for reporting reliable student scores that afford valid inferences about students’ mastery of the CA NGSS. For example:

* Is the test unidimensional for the integrated DCIs, SEPs, and CCCs; or is it multidimensional for any of these dimensions?
* Do the performance tasks (PTs) measure something different than the discrete items?
* Do the technology-enhanced items (TEIs) measure anything different from the traditional item types such as multiple-choice (MC) or constructed-response (CR) items?

These questions can be addressed by a test dimensionality study. The dimensionality study evaluates whether the CAST measures a single integrated science construct or several distinctive constructs. A determination will inform how the test items should be calibrated and how the scores should be reported to best measure students’ performance on the CAST. If the student response data from the assessment provides evidence of unidimensionality, or essential unidimensionality, all the items from different domains should be calibrated jointly, providing a total test score that can be reported based on all items the student took. If the student response data from the assessment provides evidence of multidimensionality, the items from different dimensions should be calibrated separately, and then the scores can be combined to provide a weighted composite score from these dimensions.

### Study Design

Traditional factor analyses are typically used to evaluate the underlying structure of multiple variables. Research has shown the equivalence of the classical factor analyses and the item response theory (IRT) models (Kamata & Bauer, 2008; Takane & Leeuw, 1987).

In this study, two different models within the multidimensional IRT (MIRT) framework were used to evaluate test dimensionality: a bifactor model and a correlated factor MIRT model. Model specifications are included in subsection 3.3.1 Model Specification.

Five hypothesized dimensional structures were of interest in this study:

1. Content domain
2. Item type
3. SEP
4. CCC
5. Task type (i.e., discrete items versus the PTs).

Table 3.1 provides the details on the levels in each of these five hypothesized dimensional structures.

Table 3.1 Hypothesized Dimensional Structures in the Study

|  |  |  |
| --- | --- | --- |
| Hypothesized Dimensional Structures | Number of Levels | Levels |
| Content domain | 3 | * Physical Sciences (PS) * Life Sciences (LS) * Earth and Space Sciences (ESS) |
| Item type | 3 | * CR items * MC items * TEIs |
| Science and engineering practice | 8 | * SEP 1: Asking questions (science), defining problems (engineering) * SEP 2: Developing and using models * SEP 3: Planning and carrying out investigations * SEP 4: Analyzing and interpreting data * SEP 5: Using mathematics and computational thinking * SEP 6: Constructing explanations (science), Designing solutions (engineering) * SEP 7: Engaging in argument from evidence * SEP 8: Obtaining, evaluating, and communicating information |
| Crosscutting concept | 7 | * CCC 1: Patterns * CCC 2: Cause and effect * CCC 3: Scale, proportion, and quantity * CCC 4: Systems and system models * CCC 5: Energy and matter * CCC 6: Structure and function * CCC 7: Stability and change |
| Task type | 2 | * Performance tasks * Non-performance tasks |

At the individual student level, only the total score and the content domain subscores were reported. Content domains were the most natural classifications of items, were of practical interest, and were best supported by the test design because they have sufficient numbers of items to support the reporting of these subscores.

Although only content domain scores were included in the reporting of individual student scores, additional MIRT models by other structures (i.e., SEP, CCC, task type, and item type) were run in this study to provide a complete picture on whether the test had any unexpected multidimensional structure that must be considered when calibrating and equating the test.

Students’ scores were reported based on the items the students received. Therefore, it was most informative to run the test dimensionality study at the form level, with all the blocks that a student took contributing to reporting the student’s scores, to inform individual student score reporting.

The CAST operational test used a block design, where each segment included multiple blocks and each student was randomly assigned a portion of the blocks. For each grade‑level assessment, there were two blocks in Segment A, so all students took the same two A blocks. However, there were multiple blocks from segments B and C. Each student was randomly assigned two PTs from two different domains in Segment B and randomly assigned either one discrete block or one PT from Segment C. This created multiple combinations of blocks in Segment A and Segment B that a student could receive. Instead of conducting the analyses on all possible combinations of PTs, three forms (i.e., block combinations) were carefully selected for evaluation for each grade.

Additionally, as described in Table 2.4, a third PT was added to the proposed blueprint. It will be helpful to evaluate the dimensionality of the test form when all content domains are represented by the PTs in Segment B.

Besides evaluating the test dimensionality under the current blueprint, where there are two operational PTs, this study also evaluated the test dimensionality when there were three operational PTs. This was accomplished by using a field-test PT from the third domain.

Note that in the proposed blueprint, the number of items in Segment A is reduced for grade five and grade eight. The reduction was done to maintain the proportion of items across the domains. The assumption was that the dimensional structure for a slightly shorter version of the test would be the same as for a longer version of the test.

To simulate the proposed blueprint with three PTs, each form included two operational PTs from Segment B and one field test PT from Segment C. The two operational PTs in the three forms represented different combinations of the three science domains (e.g., Form One has an LS PT and a PS PT from Segment B, and an ESS PT from Segment C; Form Two has LS and ESS PTs from Segment B and a PS PT from Segment C; and Form Three has PS and ESS PTs from Segment B and an LS PT from Segment C.) The PTs were chosen using the following guidelines:

1. All items performed reasonably well based on the item-analysis results.
2. The number of items from the three PTs were balanced.
3. For PTs that had items from two domains where one was designated as the primary domain, the ones that had more items from the primary domain were selected.

Table 3.2 shows the item block combination in each of the three forms for each grade. The FT PT blocks in this table refer to the field test PTs from Segment C. In each of these forms, the first four blocks were used to evaluate the test dimensionality under the current blueprint where the operational Segment B includes two PTs. All five blocks were used to evaluate the test dimensionality under the proposed blueprint where the operational Segment B includes three PTs.

Table 3.2 Forms Used in the Test Dimensionality Study

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level | Form One | Form Two | Form Three |
| Five | A1–A2–B1–B2–FT PT1 | A1–A2–B1–B5–FT PT6 | A1–A2–B4–B5–FT PT3 |
| Eight | A1–A2–B2–B4–FT PT5 | A1–A2–B1–B6–FT PT4 | A1–A2–B3–B6–FT PT3 |
| High School | A1–A2–B1–B2–FT PT1 | A1–A2–B1–B3–FT PT5 | A1–A2–B2–B3–FT PT3 |

All responses for CR items in the A and B blocks were scored because these items contribute to reporting individual students’ scores. For the field test CR items in the Segment C PTs (i.e., the FT PT blocks), only a sample of responses were scored to support item analyses and the building of the AI models.

The dimensionality study was run by form for the grade five, grade eight, and high school tests separately. For high school, the dimensionality study was run using a multigroup analysis with each grade as a group.

### Methods

#### Model Specification

A bifactor model and a correlated factor MIRT model were used to evaluate the test dimensionality for each hypothesized dimensional structure. Figure 3.1 and Figure 3.2 use path diagrams to show how the models were specified.

In the figures, “Gen” refers to the general factor, “SF” refers to a specific factor, and “E” refers to the error term for each item. These two diagrams illustrate the relationships between the items and the general and specific factors for a test with *K* items and three specific factors. In these two figures, item 1 to item 4 are associated with specific factor 1, item 5 to item 9 are associated with specific factor 2, and the last three items are associated with specific factor 3. The single-sided arrow indicates the source that contributes to the item variance.

In Figure 3.1, each item has three sources that contribute to the item variance: the general factor, the specific factor, and the error term.

Using the content domain structure as an example, CAST has three major content domains, and each item is associated with one of the three domains. In the bifactor model presented in Figure 3.1, each item has loadings on both the general factor and the content domain associated with an item. The correlations among the content domains are uncorrelated in this model because their correlations have been reflected by the general factor.

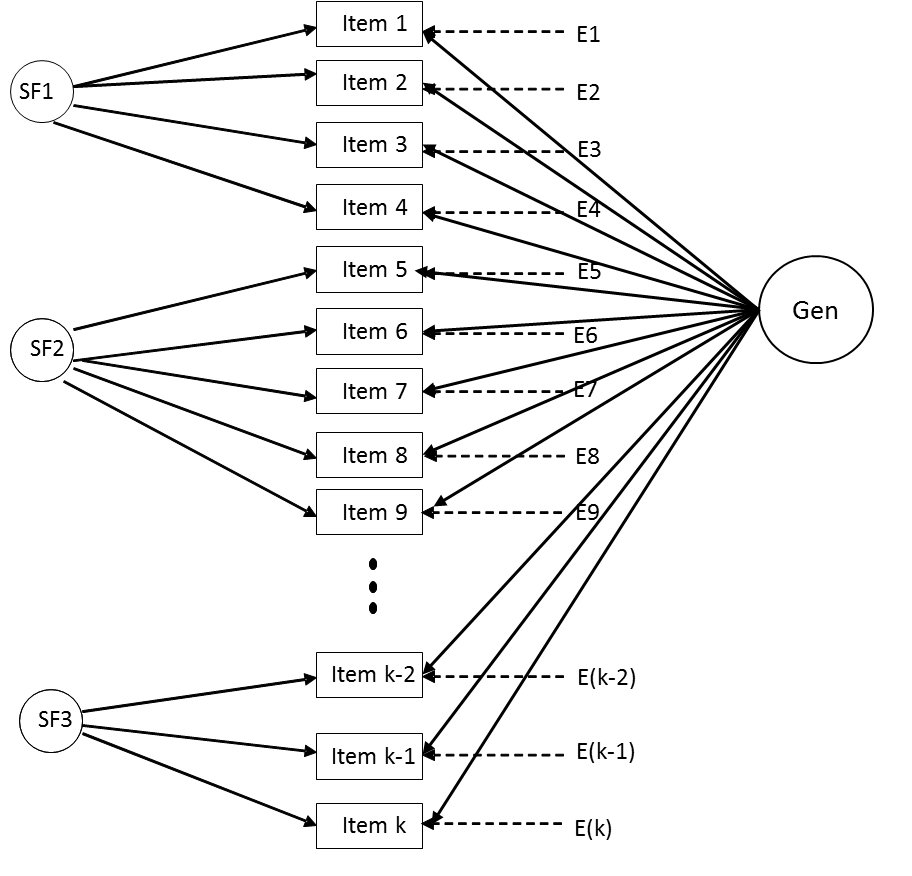


Figure 3.1. The Path Diagram for a Bifactor Model

In Figure 3.2, each item has two sources that contribute to the item variance: the specific factor and the error term. There are double-sided arrows in Figure 3.2 connecting each pair of the specific factors, indicating that the correlations between the specific factors are freely estimated in this model.

In the correlated factor MIRT model presented in Figure 3.2, however, each item has a loading on only the associated content domain. The correlations of the three content domain scores are freely estimated in this model.

The path diagram for a correlated factor MIRT model that is described in the previous paragraphs


Figure 3.2 The Path Diagram for a Correlated Factor MIRT Model

In the IRT framework, the probability of obtaining a response pattern on *K* items can be denoted as

 (3.1)

where,

**** is the probability for a student with ability vector *θ* to get a score point *y* on item *k*.

This probability  is related to a linear function of the latent variables *θ* through a link function *g(π)*, typically a probit or logit function. The difference in the bifactor model and the correlated factor MIRT model is in the link function. The link function (3.2) for a bifactor model is defined as

 (3.2)

where,

*ck* is the intercept parameter for item *k*,

*akg* is the loading on the general factor, and

*akm* is the loading on the dimension *m*.

The latent variables *θg* and *θm* are specified as uncorrelated; the link function for the correlated factor MIRT model is defined as

 (3.3)

where the instances of *θm* are correlated.

The bifactor model and the correlated factor MIRT model provide different parameter estimates but should provide similar information on the test dimensionality. Strong unidimensionality will be reflected in the MIRT model with high correlations among all dimensions. While in the bifactor model, high correlations among the dimensions will be absorbed by the general factor, and the items will have insignificant loadings on the specific factors. The bifactor model has an added benefit of showing the proportion of variance that can be explained by the specific factor as a direct way to evaluate the test dimensionality.

A multigroup analysis was run for the high school assessment. Each grade was included in the model as a separate group. The mean and variance of the ability estimates were fixed at 0 and 1 for the grade eleven population but were freely estimated for the grades ten and twelve populations. In both models, the intercept and the loadings were set to be equal for all three grades. In the correlated factor MIRT model, the covariances among the dimensions were freely estimated for all three grades.

Both the bifactor and the correlated factor MIRT model in this study were fitted with the commercial *FlexMIRT* 3.0 software.

#### Local Dependency Evaluation

Local item dependency (LID) can cause an assessment to show secondary factors in a test dimensionality evaluation. Before fitting the data with the bifactor and the correlated factor MIRT models, an evaluation of LID was conducted to remove any construct-irrelevant dependencies among items that might have caused secondary factors.

LID is a well-known assumption for a unidimensional IRT model where the success of one item depends on the success of another item. Yen (1993) listed the factors that could cause the dependency among items:

* External assistance or interference
* Speededness
* Fatigue
* Practice effects
* Item or response format
* Passage dependence
* Item chaining
* Explanation of a previous answer
* Scoring rubrics or raters
* Content, knowledge, and abilities

Some factors, such as item or response format, passage dependence, or content and knowledge, could cause some item pairs to show dependency. The test dimensionality caused by such factors is of interest in the study, so such item(s) should not be removed. However, other factors that are unrelated to the test design and construct but might cause the item pairs to show dependency should be avoided because the test multidimensionality caused by such factors is not of interest in this study. For example, if two items measure the exact same skill, students who know the answer to one item will likely know the answer to the other item. Such dependencies should be removed prior to fitting the MIRT models to evaluate the test dimensionality.

The primary focus of this evaluation was to identify the pairs of items that measured the same or highly similar content, which led to a high correlation among the item pairs after partialling out the general ability effect. The commonly used Q3 statistic (Yen, 1984) was used in this study to evaluate LID. Q3 calculates correlation among the residual of a pair of items, where the residual was defined as the difference between students’ scores and the expected item-score conditioning on a student’s ability level.

To control for the false discovery rate, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to flag the pairs that displayed high correlations among the residuals. These flagged pairs were then evaluated by examining the content. If two items in a pair were highly similar in terms of the knowledge being measured, one item—typically, the less discriminating item—could be removed from the subsequent bifactor and MIRT model fitting.

Note that a common cause of LID in a passage-based test is when common stimuli are shared between items. In the CAST test, all items in a PT were based on the same stimulus. However, these items were not removed simply because they shared the same stimuli. Instead, the commonalities shared with items in a CAST PT might indicate the PTs are measuring something different than discrete items, which would be addressed through the bifactor model by task type.

After evaluating the LID, none of the items were removed from the subsequent dimensionality studies. One of the reasons is that the majority of items included in the dimensionality study are operational items which were used in the 2017–18 field test administration. Items found to have issues in the field test were not included in the operational forms.

### Model Evaluation Criteria

A determination of test dimensionality is a subjective judgment that weighs the different sources of empirical evidence. The following evidence was considered when determining whether the CAST follows a multidimensional, essentially unidimensional, or unidimensional structure:

* Item loadings on the general factor and the group-specific factors were examined. High loadings on the general factor and low loadings on the group-specific factors for most of the items suggest that the unidimensional model is sufficient for the data.
* The indices proposed by Rodriguez, Reise, and Haviland (2016) evaluate the strength of the general factor and, therefore, indirectly assess practical unidimensionality. These indices, used in this study to evaluate test dimensionality, are:
* **Omega hierarchical (OmegaH) and Omega hierarchical subscale (OmegaHS):** OmegaH estimates the proportion of variance in total scores that can be attributed to a single general factor. OmegaHS reflects the reliability of a subscale score after controlling for the variance due to the general factor. High values of OmegaHS indicate that, after controlling for the variance due to the general factor, a large amount of the variance can still be explained by the group-specific variance, which could indicate multidimensionality.
* **Explained common variance (ECV):** ECV is the ratio of the variance explained by the general factor divided by the variance explained by the general and the group-specific factor. A high ECV value is evidence of an essentially unidimensional model (Sijtsma, 2009; Ten Berge and Socan, 2004).
* **Relative parameter bias (RPB):** An item parameter estimate could be biased when a test with multidimensional structure is forced to conform to a unidimensional model. That is, a unidimensional model is fit to data with a multidimensional structure. RPB reflects the amount of bias in the item loadings when the test with a bifactor structure is fitted using a unidimensional model and is defined as

 (3.4)

where,

 is the item loading on the general factor in a bifactor model, and

 is the item loading in a unidimensional model.

### Results

Results from the five hypothesized dimensional structures are presented in this subsection.

For each hypothesized dimensional structure, the factor loading matrices created from a bifactor model for Form One of the grade five assessment are presented in [appendix 3.A](#_Appendix_A:_Factor). The patterns of the factor loadings for the other forms in grade five and for all forms from grade eight and high school follow similar patterns and are not presented individually.

#### Test Dimensionality by Content Domain

The factor loading matrix for the bifactor model by content domain for grade five, Form One, is shown in Table 3.A.1 for the current blueprint with two PTs in Segment B, and Table 3.A.2 for the proposed blueprint with three PTs in Segment B. Items without loadings on the specific factors are noted with a hyphen (-). To make the factor loading matrix easy to read, the matrix is displayed by grouping the items that belong to the same dimension together. Therefore, the item numbers are not intended to serve as an item identifier (i.e., item 1 may refer to one item in Table 3.A.1 but might refer to a different item in Table 3.A.2.

As the data indicates, most items have higher loadings on the general factor and smaller loadings on the specific factors, which suggests that the assessment is essentially unidimensional. While the general factor is the main contributor to the total score variance, a few items have low loadings on both the general and the specific factor (e.g., items 2, 22, 36, and 47 in Table 3.A.1). Possible causes of the low loadings are that these items have low discriminating power or that they measure something different than all other items in general.

##### Test Dimensionality Indices by Content Domain

Table 3.3 provides the other indices used to evaluate the test dimensionality for content domains for the forms with two PTs: OmegaH, OmegaHS, ECV, and RPB.

Table 3.3 Evaluation Indices for the Bifactor Model by Content Domain (Two PTs)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: ESS | OmegaHS: LS | OmegaHS: PS | ECV | RPB |
| Grade Five, Form One | 0.94 | 0.00 | 0.00 | 0.01 | 0.94 | 0.011 |
| Grade Five, Form Two | 0.94 | 0.01 | 0.04 | 0.01 | 0.92 | 0.000 |
| Grade Five, Form Three | 0.93 | 0.00 | 0.04 | 0.00 | 0.93 | 0.000 |
| Grade Eight, Form One | 0.92 | 0.00 | 0.01 | 0.02 | 0.88 | -0.001 |
| Grade Eight, Form Two | 0.92 | 0.00 | 0.01 | 0.03 | 0.89 | 0.002 |
| Grade Eight, Form Three | 0.92 | 0.00 | 0.03 | 0.02 | 0.87 | 0.005 |
| High School, Form One | 0.91 | 0.02 | 0.03 | 0.02 | 0.89 | 0.016 |
| High School, Form Two | 0.91 | 0.03 | 0.02 | 0.01 | 0.90 | 0.008 |
| High School, Form Three | 0.92 | 0.03 | 0.02 | 0.02 | 0.90 | 0.000 |

Table 3.4 provides similar information for when the forms include three PTs (i.e., under the proposed blueprint).

Table 3.4 Evaluation Indices for the Bifactor Model by Content Domain (Three PTs)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: ESS | OmegaHS: LS | OmegaHS: PS | ECV | RPB |
| Grade Five, Form One | 0.95 | 0.00 | 0.00 | 0.01 | 0.93 | -0.002 |
| Grade Five, Form Two | 0.94 | 0.01 | 0.04 | 0.01 | 0.91 | 0.003 |
| Grade Five, Form Three | 0.95 | 0.00 | 0.00 | 0.01 | 0.94 | -0.002 |
| Grade Eight, Form One | 0.93 | 0.00 | 0.01 | 0.01 | 0.88 | -0.007 |
| Grade Eight, Form Two | 0.93 | 0.00 | 0.01 | 0.04 | 0.89 | 0.001 |
| Grade Eight, Form Three | 0.94 | 0.00 | 0.01 | 0.02 | 0.89 | 0.006 |
| High School, Form One | 0.93 | 0.01 | 0.02 | 0.01 | 0.91 | 0.008 |
| High School, Form Two | 0.91 | 0.03 | 0.02 | 0.09 | 0.86 | 0.006 |
| High School, Form Three | 0.93 | 0.03 | 0.03 | 0.01 | 0.89 | -0.006 |

Using Form One from grade five as an example, OmegaH is 0.94 for the form with the two PTs, suggesting that 94 percent of the variance in the total score could be attributed to the single general factor. For the form with three PTs, the OmegaH is 0.95. The values of OmegaHS for three content domains are close to zero for this form either with two PTs or three PTs, suggesting that, after accounting for the general factor, the group-specific factors only accounted for a very small proportion of the total score variance.

The ECV index shows the percent of common variance explained by the general factor and is a more direct way to measure the strength of the general factor in comparison to the content-domain specific factors. The ECV for the form with two PTs is 0.94, which means that 94 percent of the common variance was explained by the general factor and only six percent by the three content-domain factors. For the form with three PTs, the ECV is 0.93.

The RPB is 0.011 for the form with two PTs and -0.002 for the form with three PTs, indicating that if the data has a multidimensional structure but is fitted with a unidimensional model, the average bias in the factor loadings is small and negligible.

Overall, the results are consistent across forms and grades, indicating essentially unidimensional data structure based on the content-domain specification.

##### Factor Correlations by Content Domain

Table 3.5 shows the ranges of the correlations of the group-specific factors (i.e., content domain scores) across forms and grades when the data was fitted with the correlated factor MIRT model under the current blueprint with two PTs. The high correlations among the three content domain scores indicate the unidimensionality of the test.

Table 3.5 Correlations Among the Latent Content Domain Scores (Two PTs)

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | ESS | LS | PS |
| ESS | 1.00 | 0.88–0.95 | 0.90–0.97 |
| LS | 0.88–0.95 | 1.00 | 0.91–0.95 |
| PS | 0.90–0.97 | 0.91–0.95 | 1.00 |

Table 3.6 shows marginally higher correlations among the three content domains for the forms with three PTs.

Table 3.6 Correlations Among the Latent Content Domain Scores (Three PTs)

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | ESS | LS | PS |
| ESS | 1.00 | 0.90–0.97 | 0.91–0.96 |
| LS | 0.90–0.97 | 1.00 | 0.92–0.96 |
| PS | 0.91–0.96 | 0.92–0.96 | 1.00 |

#### Test Dimensionality by SEP

The factor loading matrices for the bifactor model by SEP for grade five, Form One, is shown in Table 3.A.3 and Table 3.A.4 for the form with two PTs and three PTs, respectively. Items without loadings on the specific factors are noted with a hyphen (-). The names of the SEPs are listed in Table 3.1.

In the test dimensionality analysis, only SEPs with five or more items were included as specific factors, and items associated with SEPs with fewer than five items had loadings on the general factor only.

While the CA NGSS includes eight SEPs, only four SEPs had five or more items in this form. As a result, only four SEPs are included as group-specific factors in Table 3.A.3. There were 14 items in this table with loadings on the general factor only and not on the specific factors. (There were fewer than five items per SEP for the other four SEPs not listed in the table.)

Similar to the results from the content domain structure, the moderate-to-large loadings on the general factor and the small loadings on the SEP-specific factors suggest that the general factor is the main contributor to the total test score variance.

##### Test Dimensionality Indices by SEP

Table 3.7 and Table 3.8 provide the other indices used to evaluate the test dimensionality by SEP across grade-level forms: OmegaH, OmegaHS, ECV, and RPB for the forms with two PTs and three PTs, respectively. An SEP that was not measured on a form or had fewer than five items is noted with a hyphen (-). Table 3.7 provides this information for forms with two PTs.

Table 3.7 Evaluation Indices for the Bifactor Model by SEP (Two PTs)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: SEP 2 | OmegaHS: SEP 3 | OmegaHS: SEP 4 | OmegaHS: SEP 5 | OmegaHS: SEP 6 | OmegaHS: SEP 7 | ECV | RPB |
| Grade Five, Form One | 0.94 | 0.03 | 0.04 | - | - | 0.03 | 0.03 | 0.91 | 0.010 |
| Grade Five, Form Two | 0.94 | 0.02 | 0.02 | 0.02 | - | 0.01 | 0.03 | 0.92 | -0.003 |
| Grade Five, Form Three | 0.93 | 0.00 | 0.07 | 0.02 | - | 0.03 | 0.04 | 0.87 | -0.001 |
| Grade Eight, Form One | 0.91 | 0.04 | - | 0.03 | - | 0.04 | - | 0.88 | 0.007 |
| Grade Eight, Form Two | 0.91 | 0.01 | - | 0.04 | - | 0.04 | - | 0.90 | 0.000 |
| Grade Eight, Form Three | 0.92 | 0.05 | - | 0.03 | - | 0.01 | - | 0.89 | 0.004 |
| High School, Form One | 0.91 | 0.03 | - | 0.04 | 0.00 | 0.00 | - | 0.92 | 0.000 |
| High School, Form Two | 0.91 | 0.01 | - | 0.06 | - | 0.04 | 0.07 | 0.90 | 0.000 |
| High School, Form Three | 0.92 | 0.05 | - | - | 0.00 | 0.04 | 0.08 | 0.89 | 0.001 |

Table 3.8 provides similar information for forms with three PTs. It shows OmegaH with modest improvement when the number of PTs increase from two to three.

Table 3.8 Evaluation Indices for the Bifactor Model by SEP (Three PTs)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: SEP 2 | OmegaHS: SEP 3 | OmegaHS: SEP 4 | OmegaHS: SEP 5 | OmegaHS: SEP 6 | OmegaHS: SEP 7 | OmegaHS: SEP 8 | ECV | RPB |
| Grade Five, Form One | 0.95 | 0.03 | 0.04 | 0.00 | - | 0.02 | 0.03 | 0.08 | 0.88 | -0.008 |
| Grade Five, Form Two | 0.94 | 0.02 | 0.03 | 0.02 | - | 0.01 | 0.04 | - | 0.90 | -0.001 |
| Grade Five, Form Three | 0.94 | 0.01 | 0.07 | 0.01 | - | 0.02 | 0.02 | - | 0.88 | -0.002 |
| Grade Eight, Form One | 0.93 | 0.01 | 0.05 | 0.03 | - | 0.04 | - | - | 0.85 | -0.006 |
| Grade Eight, Form Two | 0.93 | 0.03 | 0.04 | 0.04 | - | 0.02 | - | - | 0.89 | 0.006 |
| Grade Eight, Form Three | 0.94 | 0.03 | - | 0.03 | - | 0.00 | - | - | 0.90 | 0.004 |
| High School, Form One | 0.93 | 0.02 | - | 0.03 | 0.02 | 0.03 | - | - | 0.92 | 0.000 |
| High School, Form Two | 0.91 | 0.01 | - | 0.10 | 0.20 | 0.04 | 0.08 | - | 0.85 | -0.011 |
| High School, Form Three | 0.94 | 0.05 | - | 0.06 | 0.00 | 0.01 | 0.10 | - | 0.87 | -0.018 |

The results from the two-PT forms and the three-PT forms are very similar. The OmegaH index for the total test is high, ranging from 0.91 to 0.94 for the two-PT forms, and 0.91 to 0.95 for three-PT forms. The OmegaHS values are close to zero, which suggests that, after accounting for the general factor, the SEP-specific factors only accounted for a very small proportion of the total score variance.

The ECV values are high, ranging from 0.87 to 0.92 for the two-PT forms, and 0.85 to 0.92 for the three-PT forms, suggesting that more than 85 percent of the common variance is explained by the general factor and only less than 15 percent of the common variance is spread across the SEP-specific factors. The RPB is negligible for all nine forms, indicating that even if there is a bifactor structure in the data, items can be calibrated properly using a unidimensional model.

##### Factor Correlations by SEP

Table 3.9 and Table 3.10 show the range of the correlations between the group-specific factors (i.e., SEPs) across forms and grades for forms with two PTs and three PTs, respectively, when the data was fitted with the correlated-factor MIRT model. A hyphen (-) notes where no correlation can be estimated for this pair of SEPs either because both SEPs in the pair were not measured or there were not at least five items in all nine forms.

For most forms, the correlations are above 0.85. The high correlations among the SEPs indicate that there are no clearly distinctive dimensions by SEPs for the CAST. Table 3.9 provides this information for forms with two PTs.

Table 3.9 Correlations Among the Latent SEP Scores (Two PTs)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SEP | SEP 2 | SEP 3 | SEP 4 | SEP 5 | SEP 6 | SEP 7 |
| SEP 2 | 1.00 | 0.92–0.93 | 0.87–0.94 | 0.85–0.93 | 0.86–0.95 | 0.85–0.93 |
| SEP 3 | 0.92–0.93 | 1.00 | 0.92–0.93 | - | 0.92–0.93 | 0.90–0.91 |
| SEP 4 | 0.87–0.94 | 0.92–0.93 | 1.00 | 0.87–0.90 | 0.84–0.94 | 0.89–0.92 |
| SEP 5 | 0.85–0.93 | - | 0.87–0.9 | 1.00 | 0.88–0.92 | 0.90–0.92 |
| SEP 6 | 0.86–0.95 | 0.92–0.93 | 0.84–0.94 | 0.88–0.92 | 1.00 | 0.86–0.93 |
| SEP 7 | 0.85–0.93 | 0.90–0.91 | 0.89–0.92 | 0.90–0.92 | 0.86–0.93 | 1.00 |

Table 3.10 provides this information for forms with three PTs.

Table 3.10 Correlations Among the Latent SEP Scores (Three PTs)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SEP | SEP 2 | SEP 3 | SEP 4 | SEP 5 | SEP 6 | SEP 7 | SEP 8 |
| SEP 2 | 1.00 | 0.92–0.93 | 0.87–0.95 | 0.80–0.94 | 0.86–0.95 | 0.85–0.93 | 0.93–0.93 |
| SEP 3 | 0.92–0.93 | 1.00 | 0.91–0.93 | - | 0.92–0.93 | 0.91–0.92 | 0.90–0.90 |
| SEP 4 | 0.87–0.95 | 0.91–0.93 | 1.00 | 0.75–0.92 | 0.86–0.95 | 0.86–0.94 | 0.91–0.91 |
| SEP 5 | 0.80–0.94 | - | 0.75–0.92 | 1.00 | 0.81–0.93 | 0.77–0.91 | - |
| SEP 6 | 0.86–0.95 | 0.92–0.93 | 0.86–0.95 | 0.81–0.93 | 1.00 | 0.87–0.95 | 0.92–0.92 |
| SEP 7 | 0.85–0.93 | 0.91–0.92 | 0.86–0.94 | 0.77–0.91 | 0.87–0.95 | 1.00 | 0.92–0.92 |
| SEP 8 | 0.93–0.93 | 0.9–0.9 | 0.91–0.91 | - | 0.92–0.92 | 0.92–0.92 | 1.00 |

#### Test Dimensionality by CCC

The factor loading matrices from the bifactor model by CCC for grade five, Form One, are shown in Table 3.A.5 and Table 3.A.6 . for two-PT forms and three-PT forms, respectively. Items without loadings on the specific factors are noted with a hyphen (-).

The names of the CCCs are listed in Table 3.1. Similar to the dimensionality analysis performed on the SEPs, the analysis includes only the CCCs with at least five items. While the CA NGSS includes seven CCCs, only five of them have five or more items in the two-PT form. The four items with loadings on the general factor and none on the CCC-specific factors are the ones associated with those remaining two CCCs because of the small number of items.

Most items have higher loadings on the general factor than on the CCC-specific factors, although this relationship is reversed for a few items (e.g., item 31 in Table 3.A.5 and items 32 and 34 in Table 3.A.6). The loadings on the CCC-specific factors for the items belonging to the same CCC appear to show different patterns, with some large loadings, some close to zero, and some negative. This suggests that some items might be measuring something different from other items that cannot be attributed to the same CCC.

##### Test Dimensionality Indices by CCC

Table 3.11 and Table 3.12 provide the other indices used to evaluate the test dimensionality for CCCs across grade-level forms (OmegaH, OmegaHS, ECV, and RPB). Table 3.11 provides this information for the two-PT form.

Table 3.11 Evaluation Indices for the Bifactor Model by CCC (Two PTs)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: CCC-1 | OmegaHS: CCC-2 | OmegaHS: CCC-3 | OmegaHS: CCC-4 | OmegaHS: CCC-5 | OmegaHS: CCC-6 | OmegaHS: CCC-7 | ECV | RPB |
| Grade Five, Form One | 0.94 | 0.02 | 0.00 | 0.07 | 0.01 | 0.02 | - | - | 0.92 | 0.007 |
| Grade Five, Form Two | 0.94 | 0.01 | 0.01 | 0.06 | 0.03 | - | - | - | 0.90 | 0.001 |
| Grade Five, Form Three | 0.94 | 0.01 | 0.01 | 0.01 | 0.02 | - | - | - | 0.93 | 0.000 |
| Grade Eight, Form One | 0.91 | 0.04 | 0.00 | - | - | 0.06 | - | - | 0.88 | 0.002 |
| Grade Eight, Form Two | 0.91 | 0.03 | 0.03 | 0.00 | - | - | - | - | 0.91 | -0.001 |
| Grade Eight, Form Three | 0.92 | 0.00 | 0.03 | 0.00 | - | 0.05 | 0.08 | - | 0.83 | -0.003 |
| High School, Form One | 0.91 | 0.03 | 0.05 | 0.04 | - | - | - | 0.01 | 0.88 | 0.003 |
| High School, Form Two | 0.92 | 0.05 | 0.00 | 0.01 | - | 0.00 | - | 0.01 | 0.90 | -0.012 |
| High School, Form Three | 0.92 | 0.03 | 0.08 | - | - | 0.00 | - | 0.01 | 0.88 | -0.009 |

Table 3.12 provides this information for the three-PT form.

Table 3.12 Evaluation Indices for the Bifactor Model by CCC (Three PTs)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: CCC-1 | OmegaHS: CCC-2 | OmegaHS: CCC-3 | OmegaHS: CCC-4 | OmegaHS: CCC-5 | OmegaHS: CCC-6 | OmegaHS: CCC-7 | ECV | RPB |
| Grade Five, Form One | 0.95 | 0.02 | 0.00 | 0.06 | 0.03 | 0.02 | - | - | 0.87 | -0.004 |
| Grade Five, Form Two | 0.94 | 0.00 | 0.01 | 0.06 | 0.03 | - | - | - | 0.91 | 0.001 |
| Grade Five, Form Three | 0.95 | 0.00 | 0.00 | 0.01 | 0.02 | - | - | - | 0.93 | -0.003 |
| Grade Eight, Form One | 0.92 | 0.04 | 0.00 | - | 0.11 | 0.06 | - | - | 0.85 | 0.000 |
| Grade Eight, Form Two | 0.93 | 0.04 | 0.01 | 0.02 | - | - | - | - | 0.92 | 0.004 |
| Grade Eight, Form Three | 0.94 | 0.01 | 0.00 | 0.00 | - | 0.05 | 0.06 | - | 0.85 | 0.000 |
| High School, Form One | 0.93 | 0.02 | 0.04 | 0.03 | - | 0.04 | - | 0.01 | 0.92 | 0.013 |
| High School, Form Two | 0.91 | 0.05 | 0.09 | 0.01 | 0.20 | 0.00 | - | 0.02 | 0.85 | -0.008 |
| High School, Form Three | 0.94 | 0.00 | 0.04 | - | - | 0.12 | - | 0.03 | 0.87 | -0.007 |

The OmegaH values range from 0.91 to 0.94 for two-PT forms, and 0.91 to 0.95 for three-PT forms, indicating a large proportion of the total score variance can be attributed to the general factor. The values for OmegaHS are small, indicating that, after accounting for the general factor, these CCC-specific factors contribute little to the total score variance.

The ECV values range from 0.83 to 0.93 for the two-PT forms and 0.85 to 0.93 for the three-PT forms, indicating that the general factor accounts for the majority of the common variance. The small RPB indicates that items can be calibrated properly using a unidimensional model.

##### Factor Correlations by CCC

Table 3.13 and Table 3.14 show the ranges of each pair of the latent CCC scores under the correlated factor MIRT model for forms with two PTs and three PTs, respectively. A hyphen (-) notes where no correlation can be estimated for this pair of CCCs because either both CCCs in the pair were not measured or there were not at least five items in all nine forms.

Most correlations are higher than 0.85. The moderate-to-high correlations suggest the test is essentially unidimensional with the CCC structure.

Table 3.13 provides this information for forms with two PTs.

Table 3.13 Correlations Among Latent CCC Scores (Two PTs)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CCC | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 | CCC 6 | CCC 7 |
| CCC 1 | 1.00 | 0.85–0.95 | 0.85–0.94 | 0.93–0.93 | 0.82–0.92 | 0.91–0.91 | 0.9–0.94 |
| CCC 2 | 0.85–0.95 | 1.00 | 0.74–0.94 | 0.95–0.95 | 0.78–0.91 | 0.91–0.91 | 0.84–0.9 |
| CCC 3 | 0.85–0.94 | 0.74–0.94 | 1.00 | 0.91–0.91 | 0.82–0.91 | 0.89–0.89 | 0.85–0.91 |
| CCC 4 | 0.93–0.93 | 0.95–0.95 | 0.91–0.91 | 1.00 | 0.92–0.92 | - | - |
| CCC 5 | 0.82–0.92 | 0.78–0.91 | 0.82–0.91 | 0.92–0.92 | 1.00 | 0.93–0.93 | 0.83–0.90 |
| CCC 6 | 0.91–0.91 | 0.91–0.91 | 0.89–0.89 | - | 0.93–0.93 | 1.00 | - |
| CCC 7 | 0.90–0.94 | 0.84–0.90 | 0.85–0.91 | - | 0.83–0.90 | - | 1.00 |

Table 3.14 provides this information for forms with three PTs.

Table 3.14 Correlations Among Latent CCC Scores (Three PTs)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CCC | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 | CCC 6 | CCC 7 |
| CCC 1 | 1.00 | 0.88–0.96 | 0.83–0.94 | 0.85–0.94 | 0.79–0.92 | 0.94–0.94 | 0.87–0.94 |
| CCC 2 | 0.88–0.96 | 1.00 | 0.76–0.94 | 0.83–0.95 | 0.77–0.94 | 0.95–0.95 | 0.85–0.90 |
| CCC 3 | 0.83–0.94 | 0.76–0.94 | 1.00 | 0.80–0.91 | 0.80–0.92 | 0.93–0.93 | 0.83–0.91 |
| CCC 4 | 0.85–0.94 | 0.83–0.95 | 0.80–0.91 | 1.00 | 0.83–0.9 | - | 0.83–0.88 |
| CCC 5 | 0.79–0.92 | 0.77–0.94 | 0.80–0.92 | 0.83–0.90 | 1.00 | 0.94–0.94 | 0.80–0.90 |
| CCC 6 | 0.94–0.94 | 0.95–0.95 | 0.93–0.93 | - | 0.94–0.94 | 1.00 | - |
| CCC 7 | 0.87–0.94 | 0.85–0.90 | 0.83–0.91 | 0.83–0.88 | 0.80–0.90 | - | 1.00 |

#### Test Dimensionality by Item Type

Three items types (MC, CR, and TEIs) were included. The MC items include both the single-selection and multiple-selection items. The items that required extended text input were considered CR items. The rest of the items are considered the TEIs (e.g., grid, inline choice, numeric entry items).

##### Test Dimensionality Indices by Item Type

Table 3.15 and Table 3.A.8 show the factor loading matrix from the bifactor model by item type and for grade five, Form 1. Most items have high loadings on the general factor and low loadings on the item-type-specific factors. A few items (e.g., items 9, 18, and 37) have low loadings on both the general factor and the item-type-specific factor, indicating they might be measuring something different from both the general and the item-type-specific factors.

Table 3.15 and Table 3.16 provide the other indices used to evaluate the test dimensionality across forms and grades for forms with two-PT forms and three-PT forms, respectively. The OmegaH values range from 0.91 to 0.94 for the two-PT forms and 0.91 to 0.95 for the three-PT forms, indicating a large proportion of the total score variance can be attributed to the general factor. The ECV values range from 0.87 to 0.93 for the two-PT forms, and 0.88 to 0.93 for the three-PT forms, indicating that the general factor accounts for the majority of the common variance. Compared to the OmegaHS values for the MC and TEI dimensions, which are mostly close to zero, these values for the CR dimension are slightly larger. The slightly larger OmegaHS values for the CR dimension suggest that there might be a slight multidimensionality structure for the CR dimension; however, the small RPB indicates it is still appropriate to calibrate the data with a unidimensional model. Table 3.15 provides this information for forms with two PTs.

Table 3.15 Evaluation Indices for the Bifactor Model by Item Type (Two PTs)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: CR | OmegaHS: MC | OmegaHS: TEI | ECV | RPB |
| Grade Five, Form One | 0.94 | 0.06 | 0.00 | 0.01 | 0.93 | -0.001 |
| Grade Five, Form Two | 0.94 | 0.09 | 0.01 | 0.02 | 0.91 | 0.000 |
| Grade Five, Form Three | 0.94 | 0.09 | 0.00 | 0.00 | 0.92 | -0.001 |
| Grade Eight, Form One | 0.91 | 0.07 | 0.06 | 0.00 | 0.91 | 0.025 |
| Grade Eight, Form Two | 0.91 | 0.08 | 0.05 | 0.00 | 0.90 | 0.005 |
| Grade Eight, Form Three | 0.92 | 0.08 | 0.03 | 0.01 | 0.87 | 0.008 |
| High School, Form One | 0.92 | 0.08 | 0.00 | 0.00 | 0.91 | -0.008 |
| High School, Form Two | 0.92 | 0.04 | 0.00 | 0.01 | 0.87 | 0.000 |
| High School, Form Three | 0.92 | 0.08 | 0.00 | 0.01 | 0.91 | -0.007 |

Table 3.16 provides this information for forms with three PTs.

Table 3.16 Evaluation Indices for the Bifactor Model by Item Type (Three PTs)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: CR | OmegaHS: MC | OmegaHS: TEI | ECV | RPB |
| Grade Five, Form One | 0.95 | 0.05 | 0.00 | 0.00 | 0.93 | 0.000 |
| Grade Five, Form Two | 0.94 | 0.08 | 0.00 | 0.01 | 0.90 | -0.003 |
| Grade Five, Form Three | 0.94 | 0.10 | 0.00 | 0.01 | 0.93 | -0.001 |
| Grade Eight, Form One | 0.92 | 0.08 | 0.03 | 0.00 | 0.91 | 0.008 |
| Grade Eight, Form Two | 0.93 | 0.05 | 0.02 | 0.00 | 0.91 | 0.003 |

Table 3.16 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: CR | OmegaHS: MC | OmegaHS: TEI | ECV | RPB |
| Grade Eight, Form Three | 0.93 | 0.08 | 0.03 | 0.01 | 0.89 | 0.008 |
| High School, Form One | 0.93 | 0.08 | 0.00 | 0.00 | 0.92 | -0.004 |
| High School, Form Two | 0.91 | 0.07 | 0.00 | 0.03 | 0.88 | 0.003 |
| High School, Form Three | 0.94 | 0.07 | 0.00 | 0.01 | 0.91 | -0.007 |

Table 3.17 and Table 3.18 show the correlations among the latent item-type scores. The correlations are all higher than 0.88, suggesting an essential unidimensionality by the item-type structure. Table 3.17 provides this information for forms with two PTs.

Table 3.17 Correlations Among the Latent Item-Type Scores (Two PTs)

|  |  |  |  |
| --- | --- | --- | --- |
| Item Type | MC | CR | TEI |
| MC | 1.00 | 0.88–0.92 | 0.89–0.94 |
| CR | 0.88–0.92 | 1.00 | 0.92–0.97 |
| TEI | 0.89–0.94 | 0.92–0.97 | 1.00 |

Table 3.18 provides this information for forms with three PTs.

Table 3.18 Correlations Among the Latent Item-Type Scores (Three PTs)

|  |  |  |  |
| --- | --- | --- | --- |
| Item Type | MC | CR | TEI |
| MC | 1.00 | 0.88–0.92 | 0.89–0.94 |
| CR | 0.88–0.92 | 1.00 | 0.92–0.98 |
| TEI | 0.89–0.94 | 0.92–0.98 | 1.00 |

The correlations between MC and CR are slightly smaller than the correlations between MC and TEIs, mostly because both MC and TEIs are both selected-response items while CR items require students to produce content as part of their response. This result is consistent with the findings from Table 3.15 and Table 3.16, which show the CR items might be measuring something different than the MC and TEIs.

#### Test Dimensionality by Task Type

To evaluate whether the PTs and the non-PTs show distinctive dimensionality, the bifactor model and the correlated factor MIRT model were fitted to the data by the task type. Table 3.A.9 and Table 3.A.10 show the factor loading matrix for the bifactor model by non-PT items and PT items. Similar to the results from the other hypothesized dimensional structures, the results for most items reveal higher loadings on the general factor than the task-type-specific factor. The loadings on the task-type-specific factors are low in general, with exceptions to a few items. Overall, there is no clear evidence of multidimensionality by the task type.

Table 3.19 and Table 3.20 provide the other indices used to evaluate the test dimensionality across forms and grades. Table 3.19 does so for forms with two PTs.

Table 3.19 Evaluation Indices for the Bifactor Model by Task Type (Two PTs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: Non-PTs | OmegaHS: PTs | ECV | RPB |
| Grade Five, Form One | 0.93 | 0.01 | 0.05 | 0.92 | 0.016 |
| Grade Five, Form Two | 0.94 | 0.00 | 0.04 | 0.93 | 0.002 |
| Grade Five, Form Three | 0.93 | 0.02 | 0.01 | 0.91 | 0.005 |
| Grade Eight, Form One | 0.92 | 0.00 | 0.06 | 0.87 | -0.007 |
| Grade Eight, Form Two | 0.92 | 0.00 | 0.04 | 0.89 | -0.003 |
| Grade Eight, Form Three | 0.92 | 0.01 | 0.05 | 0.85 | 0.006 |
| High School, Form One | 0.91 | 0.01 | 0.04 | 0.91 | 0.009 |
| High School, Form Two | 0.91 | 0.02 | 0.04 | 0.91 | 0.004 |
| High School, Form Three | 0.88 | 0.08 | 0.01 | 0.89 | 0.022 |

Table 3.20 provides this information for forms with three PTs.

Table 3.20 Evaluation Indices for the Bifactor Model by Task Type (Three PTs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade and Form | OmegaH | OmegaHS: Non-PTs | OmegaHS: PTs | ECV | RPB |
| Grade Five, Form One | 0.94 | 0.00 | 0.05 | 0.93 | 0.003 |
| Grade Five, Form Two | 0.94 | 0.01 | 0.02 | 0.92 | 0.005 |
| Grade Five, Form Three | 0.94 | 0.01 | 0.04 | 0.93 | 0.002 |
| Grade Eight, Form One | 0.92 | 0.00 | 0.04 | 0.88 | -0.004 |
| Grade Eight, Form Two | 0.93 | 0.00 | 0.04 | 0.91 | 0.003 |
| Grade Eight, Form Three | 0.94 | 0.00 | 0.02 | 0.89 | 0.007 |
| High School, Form One | 0.92 | 0.02 | 0.02 | 0.93 | 0.022 |
| High School, Form Two | 0.90 | 0.04 | 0.03 | 0.89 | 0.021 |
| High School, Form Three | 0.92 | 0.06 | 0.02 | 0.91 | 0.012 |

The results from the two-PT forms are very similar to those from the three-PT forms. The OmegaH values range from 0.88 to 0.94 for the two-PT forms and 0.90 to 0.94 for the three-PT forms, indicating that a large proportion of the total score variance can be attributed to the general factor. The values of the OmegaHS for the PTs and non-PTs are small. The ECV values range from 0.85 to 0.93 for the two-PT forms and 0.88 to 0.93 for the three-PT forms, indicating that the general factor accounts for most of the common variance. The small RPB indicates it is appropriate to calibrate the data with a unidimensional model. The correlations between the latent PT and non-PT scores under the correlated factor MIRT model range from 0.90 to ‍0.95 for the two-PT forms and 0.90 to 0.96 for the three-PT forms. These high correlations also indicate an essential unidimensionality by task type.

### Implications on Calibration and Score Reporting

Test dimensionality has implications regarding how items should be calibrated and how scores should be reported. For example, if a test shows strong multidimensionality by content domain, it would be appropriate to calibrate the items from different content domains separately and establish a separate scale for each of them. The total score reported could then be a composite score of the different content domains.

For the CAST, the results from all five hypothesized dimensional structures consistently indicate unidimensionality for this test. When a test is unidimensional by content domain, all items in different content domains can be calibrated jointly using a unidimensional model. As a result, all items from different domains are put on the same IRT scale. Because student performance is largely undifferentiated across the content domains, it is appropriate to use a single integrated score based on all items to reflect students’ performance on the test. Reporting at the domain level is still possible, although it might not provide much unique information from the total score for most students.

The results from the two-PT forms are very close to the three-PT forms, indicating no change in the test dimensionality when a third PT is added to Segment B. This suggests if the proposed blueprint is adopted, there is no impact on the test dimensionality and the score reporting can follow the current approach, which is to report an overall score and domain-specific levels.

### References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc*. *B, 57*, 289–300.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–53.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–50.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika*, *74*(1), 145–154.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Ten Berge, J. M., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–25.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–45.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

### Appendix 3.A: Factor Loading Matrix

Note: The hyphens in the tables in appendix 3.A denote that data was not available because the item did not belong to that domain.

Table 3.A.1 Factor Loading Matrix by Content Domain—Grade Five, Form One, Two PTs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | Earth and Space Sciences | Life Sciences | Physical Sciences |
| Item 1 | 0.59 | 0.24 | - | - |
| Item 2 | 0.19 | 0.15 | - | - |
| Item 3 | 0.37 | 0.10 | - | - |
| Item 4 | 0.55 | 0.06 | - | - |
| Item 5 | 0.48 | 0.04 | - | - |
| Item 6 | 0.48 | 0.03 | - | - |
| Item 7 | 0.45 | -0.15 | - | - |
| Item 8 | 0.74 | -0.05 | - | - |
| Item 9 | 0.51 | -0.04 | - | - |
| Item 10 | 0.41 | -0.00 | - | - |
| Item 11 | 0.64 | -0.00 | - | - |
| Item 12 | 0.81 | - | 0.26 | - |
| Item 13 | 0.55 | - | 0.24 | - |
| Item 14 | 0.68 | - | 0.16 | - |
| Item 15 | 0.34 | - | 0.07 | - |
| Item 16 | 0.59 | - | 0.07 | - |
| Item 17 | 0.69 | - | 0.03 | - |
| Item 18 | 0.47 | - | 0.03 | - |
| Item 19 | 0.73 | - | 0.02 | - |
| Item 20 | 0.75 | - | 0.01 | - |
| Item 21 | 0.69 | - | 0.00 | - |
| Item 22 | 0.18 | - | -0.29 | - |
| Item 23 | 0.54 | - | -0.17 | - |
| Item 24 | 0.47 | - | -0.14 | - |
| Item 25 | 0.51 | - | -0.12 | - |
| Item 26 | 0.32 | - | -0.11 | - |
| Item 27 | 0.55 | - | -0.09 | - |
| Item 28 | 0.41 | - | - | 0.31 |
| Item 29 | 0.33 | - | - | 0.29 |

Table 3.A.1 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | Earth and Space Sciences | Life Sciences | Physical Sciences |
| Item 30 | 0.61 | - | - | 0.25 |
| Item 31 | 0.44 | - | - | 0.22 |
| Item 32 | 0.44 | - | - | 0.17 |
| Item 33 | 0.63 | - | - | 0.16 |
| Item 34 | 0.36 | - | - | 0.09 |
| Item 35 | 0.41 | - | - | 0.04 |
| Item 36 | 0.02 | - | - | 0.03 |
| Item 37 | 0.45 | - | - | 0.01 |
| Item 38 | 0.55 | - | - | -0.14 |
| Item 39 | 0.38 | - | - | -0.08 |
| Item 40 | 0.41 | - | - | -0.08 |
| Item 41 | 0.55 | - | - | -0.08 |
| Item 42 | 0.64 | - | - | -0.07 |
| Item 43 | 0.55 | - | - | -0.06 |
| Item 44 | 0.61 | - | - | -0.03 |
| Item 45 | 0.43 | - | - | -0.01 |
| Item 46 | 0.71 | - | - | -0.01 |
| Item 47 | 0.19 | - | - | -0.01 |

Table 3.A.2 Factor Loading Matrix by Content Domain—Grade Five, Form One, Three PTs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | Earth and Space Sciences | Life Sciences | Physical Sciences |
| Item 1 | 0.25 | 0.25 | - | - |
| Item 2 | 0.65 | 0.24 | - | - |
| Item 3 | 0.62 | 0.21 | - | - |
| Item 4 | 0.45 | 0.17 | - | - |
| Item 5 | 0.64 | 0.08 | - | - |
| Item 6 | 0.57 | 0.05 | - | - |
| Item 7 | 0.50 | 0.04 | - | - |
| Item 8 | 0.74 | 0.02 | - | - |
| Item 9 | 0.59 | 0.02 | - | - |
| Item 10 | 0.48 | 0.01 | - | - |
| Item 11 | 0.35 | -0.19 | - | - |
| Item 12 | 0.59 | -0.15 | - | - |
| Item 13 | 0.19 | -0.11 | - | - |
| Item 14 | 0.37 | -0.10 | - | - |
| Item 15 | 0.55 | -0.09 | - | - |
| Item 16 | 0.41 | -0.05 | - | - |
| Item 17 | 0.29 | -0.04 | - | - |
| Item 18 | 0.48 | -0.00 | - | - |
| Item 19 | 0.81 | - | 0.24 | - |
| Item 20 | 0.55 | - | 0.23 | - |
| Item 21 | 0.68 | - | 0.16 | - |
| Item 22 | 0.59 | - | 0.08 | - |
| Item 23 | 0.34 | - | 0.07 | - |
| Item 24 | 0.69 | - | 0.03 | - |
| Item 25 | 0.72 | - | 0.03 | - |
| Item 26 | 0.47 | - | 0.03 | - |
| Item 27 | 0.68 | - | 0.02 | - |
| Item 28 | 0.75 | - | 0.02 | - |
| Item 29 | 0.18 | - | -0.30 | - |
| Item 30 | 0.54 | - | -0.17 | - |
| Item 31 | 0.47 | - | -0.14 | - |
| Item 32 | 0.50 | - | -0.13 | - |
| Item 33 | 0.32 | - | -0.12 | - |

Table 3.A.2 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | Earth and Space Sciences | Life Sciences | Physical Sciences |
| Item 34 | 0.55 | - | -0.07 | - |
| Item 35 | 0.41 | - | - | 0.31 |
| Item 36 | 0.33 | - | - | 0.28 |
| Item 37 | 0.61 | - | - | 0.25 |
| Item 38 | 0.44 | - | - | 0.22 |
| Item 39 | 0.44 | - | - | 0.18 |
| Item 40 | 0.63 | - | - | 0.16 |
| Item 41 | 0.36 | - | - | 0.09 |
| Item 42 | 0.41 | - | - | 0.04 |
| Item 43 | 0.03 | - | - | 0.03 |
| Item 44 | 0.45 | - | - | 0.01 |
| Item 45 | 0.55 | - | - | -0.14 |
| Item 46 | 0.42 | - | - | -0.09 |
| Item 47 | 0.38 | - | - | -0.08 |
| Item 48 | 0.55 | - | - | -0.08 |
| Item 49 | 0.63 | - | - | -0.07 |
| Item 50 | 0.55 | - | - | -0.07 |
| Item 51 | 0.61 | - | - | -0.03 |
| Item 52 | 0.19 | - | - | -0.02 |
| Item 53 | 0.43 | - | - | -0.01 |
| Item 54 | 0.70 | - | - | -0.00 |

Table 3.A.3 Factor Loading Matrix by Science and Engineering Practices (SEP)—Grade Five, Form One, Two PTs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item | General | SEP 2 | SEP 3 | SEP 6 | SEP 7 |
| Item 1 | 0.16 | 0.30 | - | - | - |
| Item 2 | 0.69 | 0.28 | - | - | - |
| Item 3 | 0.53 | 0.21 | - | - | - |
| Item 4 | 0.46 | 0.16 | - | - | - |
| Item 5 | 0.50 | 0.15 | - | - | - |
| Item 6 | 0.73 | 0.12 | - | - | - |
| Item 7 | 0.18 | 0.12 | - | - | - |
| Item 8 | 0.45 | 0.05 | - | - | - |
| Item 9 | 0.48 | 0.05 | - | - | - |
| Item 10 | 0.02 | 0.04 | - | - | - |
| Item 11 | 0.75 | 0.01 | - | - | - |
| Item 12 | 0.55 | -0.12 | - | - | - |
| Item 13 | 0.43 | -0.06 | - | - | - |
| Item 14 | 0.38 | -0.04 | - | - | - |
| Item 15 | 0.42 | - | 0.07 | - | - |
| Item 16 | 0.63 | - | 0.04 | - | - |
| Item 17 | 0.19 | - | 0.01 | - | - |
| Item 18 | 0.45 | - | 0.01 | - | - |
| Item 19 | 0.55 | - | -0.77 | - | - |
| Item 20 | 0.41 | - | -0.02 | - | - |
| Item 21 | 0.51 | - | - | 0.25 | - |
| Item 22 | 0.55 | - | - | 0.23 | - |
| Item 23 | 0.54 | - | - | 0.15 | - |
| Item 24 | 0.55 | - | - | 0.13 | - |
| Item 25 | 0.69 | - | - | 0.08 | - |
| Item 26 | 0.37 | - | - | -0.09 | - |
| Item 27 | 0.55 | - | - | -0.05 | - |
| Item 28 | 0.68 | - | - | - | 0.31 |
| Item 29 | 0.59 | - | - | - | 0.23 |
| Item 30 | 0.82 | - | - | - | 0.16 |
| Item 31 | 0.61 | - | - | - | 0.14 |
| Item 32 | 0.47 | - | - | - | 0.08 |
| Item 33 | 0.34 | - | - | - | 0.05 |
| Item 34 | 0.48 | - | - | - | -0.05 |
| Item 35 | 0.32 | - | - | - | -0.05 |
| Item 36 | 0.41 | - | - | - | - |
| Item 37 | 0.45 | - | - | - | - |
| Item 38 | 0.74 | - | - | - | - |
| Item 39 | 0.63 | - | - | - | - |

Table 3.A.3 *(continuation)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item | General | SEP 2 | SEP 3 | SEP 6 | SEP 7 |
| Item 40 | 0.36 | - | - | - | - |
| Item 41 | 0.59 | - | - | - | - |
| Item 42 | 0.69 | - | - | - | - |
| Item 43 | 0.55 | - | - | - | - |
| Item 44 | 0.64 | - | - | - | - |
| Item 45 | 0.61 | - | - | - | - |
| Item 46 | 0.34 | - | - | - | - |
| Item 47 | 0.42 | - | - | - | - |

Table 3.A.4 Factor Loading Matrix by SEP—Grade Five, Form One, Three PTs

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Item | General | SEP 2 | SEP 3 | SEP 4 | SEP 6 | SEP 7 | SEP 8 |
| Item 1 | 0.68 | 0.29 | - | - | - | - | - |
| Item 2 | 0.17 | 0.28 | - | - | - | - | - |
| Item 3 | 0.53 | 0.20 | - | - | - | - | - |
| Item 4 | 0.50 | 0.16 | - | - | - | - | - |
| Item 5 | 0.46 | 0.15 | - | - | - | - | - |
| Item 6 | 0.73 | 0.13 | - | - | - | - | - |
| Item 7 | 0.19 | 0.11 | - | - | - | - | - |
| Item 8 | 0.48 | 0.06 | - | - | - | - | - |
| Item 9 | 0.44 | 0.05 | - | - | - | - | - |
| Item 10 | 0.03 | 0.04 | - | - | - | - | - |
| Item 11 | 0.75 | 0.00 | - | - | - | - | - |
| Item 12 | 0.55 | -0.13 | - | - | - | - | - |
| Item 13 | 0.43 | -0.05 | - | - | - | - | - |
| Item 14 | 0.38 | -0.04 | - | - | - | - | - |
| Item 15 | 0.42 | - | 0.07 | - | - | - | - |
| Item 16 | 0.63 | - | 0.04 | - | - | - | - |
| Item 17 | 0.19 | - | 0.01 | - | - | - | - |
| Item 18 | 0.46 | - | 0.01 | - | - | - | - |
| Item 19 | 0.55 | - | -0.77 | - | - | - | - |
| Item 20 | 0.42 | - | -0.02 | - | - | - | - |
| Item 21 | 0.66 | - | - | 0.20 | - | - | - |
| Item 22 | 0.60 | - | - | 0.19 | - | - | - |
| Item 23 | 0.65 | - | - | 0.09 | - | - | - |
| Item 24 | 0.41 | - | - | 0.04 | - | - | - |
| Item 25 | 0.69 | - | - | -0.15 | - | - | - |
| Item 26 | 0.55 | - | - | -0.15 | - | - | - |
| Item 27 | 0.55 | - | - | - | 0.23 | - | - |
| Item 28 | 0.51 | - | - | - | 0.21 | - | - |
| Item 29 | 0.55 | - | - | - | 0.17 | - | - |
| Item 30 | 0.55 | - | - | - | 0.13 | - | - |
| Item 31 | 0.62 | - | - | - | 0.08 | - | - |
| Item 32 | 0.70 | - | - | - | 0.07 | - | - |
| Item 33 | 0.58 | - | - | - | 0.00 | - | - |
| Item 34 | 0.38 | - | - | - | -0.12 | - | - |
| Item 35 | 0.55 | - | - | - | -0.07 | - | - |
| Item 36 | 0.68 | - | - | - | - | 0.30 | - |
| Item 37 | 0.59 | - | - | - | - | 0.24 | - |
| Item 38 | 0.83 | - | - | - | - | 0.15 | - |
| Item 39 | 0.61 | - | - | - | - | 0.14 | - |

Table 3.A.4 *(continuation)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Item | General | SEP 2 | SEP 3 | SEP 4 | SEP 6 | SEP 7 | SEP 8 |
| Item 40 | 0.47 | - | - | - | - | 0.07 | - |
| Item 41 | 0.35 | - | - | - | - | 0.05 | - |
| Item 42 | 0.32 | - | - | - | - | -0.06 | - |
| Item 43 | 0.48 | - | - | - | - | -0.05 | - |
| Item 44 | 0.44 | - | - | - | - | - | - |
| Item 45 | 0.74 | - | - | - | - | - | -0.03 |
| Item 46 | 0.63 | - | - | - | - | - | - |
| Item 47 | 0.36 | - | - | - | - | - | - |
| Item 48 | 0.59 | - | - | - | - | - | 0.04 |
| Item 49 | 0.61 | - | - | - | - | - | - |
| Item 50 | 0.34 | - | - | - | - | - | - |
| Item 51 | 0.42 | - | - | - | - | - | - |
| Item 52 | 0.36 | - | - | - | - | - | 0.82 |
| Item 53 | 0.29 | - | - | - | - | - | -0.00 |
| Item 54 | 0.23 | - | - | - | - | - | -0.02 |

Table 3.A.5 Factor Loading Matrix by Crosscutting Concept (CCC)—Grade Five, Form One, Two PTs

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | General | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 |
| Item 1 | 0.55 | 0.29 | - | - | - | - |
| Item 2 | 0.56 | 0.21 | - | - | - | - |
| Item 3 | 0.37 | 0.07 | - | - | - | - |
| Item 4 | 0.41 | 0.00 | - | - | - | - |
| Item 5 | 0.64 | -0.02 | - | - | - | - |
| Item 6 | 0.46 | -0.02 | - | - | - | - |
| Item 7 | 0.43 | -0.00 | - | - | - | - |
| Item 8 | 0.81 | - | 0.33 | - | - | - |
| Item 9 | 0.45 | - | 0.19 | - | - | - |
| Item 10 | 0.55 | - | 0.17 | - | - | - |
| Item 11 | 0.69 | - | 0.16 | - | - | - |
| Item 12 | 0.63 | - | 0.10 | - | - | - |
| Item 13 | 0.61 | - | 0.07 | - | - | - |
| Item 14 | 0.48 | - | 0.03 | - | - | - |
| Item 15 | 0.60 | - | -0.17 | - | - | - |
| Item 16 | 0.20 | - | -0.12 | - | - | - |
| Item 17 | 0.45 | - | -0.09 | - | - | - |
| Item 18 | 0.63 | - | -0.06 | - | - | - |
| Item 19 | 0.46 | - | -0.05 | - | - | - |
| Item 20 | 0.70 | - | -0.02 | - | - | - |
| Item 21 | 0.41 | - | - | 0.39 | - | - |
| Item 22 | 0.62 | - | - | 0.33 | - | - |
| Item 23 | 0.33 | - | - | 0.25 | - | - |
| Item 24 | 0.36 | - | - | 0.09 | - | - |
| Item 25 | 0.02 | - | - | 0.02 | - | - |
| Item 26 | 0.38 | - | - | -0.04 | - | - |
| Item 27 | 0.69 | - | - | -0.00 | - | - |
| Item 28 | 0.19 | - | - | -0.00 | - | - |
| Item 29 | 0.73 | - | - | - | 0.33 | - |
| Item 30 | 0.51 | - | - | - | 0.07 | - |
| Item 31 | 0.17 | - | - | - | 0.07 | - |
| Item 32 | 0.48 | - | - | - | 0.06 | - |
| Item 33 | 0.74 | - | - | - | 0.05 | - |
| Item 34 | 0.35 | - | - | - | 0.04 | - |
| Item 35 | 0.60 | - | - | - | 0.03 | - |
| Item 36 | 0.75 | - | - | - | -0.03 | - |
| Item 37 | 0.54 | - | - | - | -0.01 | - |
| Item 38 | 0.70 | - | - | - | - | 0.46 |
| Item 39 | 0.47 | - | - | - | - | 0.12 |

Table 3.A.5 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | General | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 |
| Item 40 | 0.32 | - | - | - | - | 0.06 |
| Item 41 | 0.45 | - | - | - | - | -0.05 |
| Item 42 | 0.55 | - | - | - | - | -0.04 |
| Item 43 | 0.55 | - | - | - | - | -0.01 |
| Item 44 | 0.42 | - | - | - | - | - |
| Item 45 | 0.41 | - | - | - | - | - |
| Item 46 | 0.51 | - | - | - | - | - |
| Item 47 | 0.55 | - | - | - | - | - |

Table 3.A.6 Factor Loading Matrix by Crosscutting Concept (CCC)—Grade Five, Form One, Three PTs

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | General | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 |
| Item 1 | 0.35 | 0.77 | - | - | - | - |
| Item 2 | 0.59 | 0.08 | - | - | - | - |
| Item 3 | 0.37 | 0.06 | - | - | - | - |
| Item 4 | 0.46 | 0.06 | - | - | - | - |
| Item 5 | 0.54 | 0.03 | - | - | - | - |
| Item 6 | 0.55 | 0.03 | - | - | - | - |
| Item 7 | 0.41 | 0.02 | - | - | - | - |
| Item 8 | 0.29 | 0.00 | - | - | - | - |
| Item 9 | 0.64 | -0.02 | - | - | - | - |
| Item 10 | 0.43 | -0.01 | - | - | - | - |
| Item 11 | 0.66 | -0.01 | - | - | - | - |
| Item 12 | 0.23 | -0.01 | - | - | - | - |
| Item 13 | 0.81 | - | 0.29 | - | - | - |
| Item 14 | 0.68 | - | 0.18 | - | - | - |
| Item 15 | 0.45 | - | 0.18 | - | - | - |
| Item 16 | 0.55 | - | 0.18 | - | - | - |
| Item 17 | 0.62 | - | 0.14 | - | - | - |
| Item 18 | 0.63 | - | 0.11 | - | - | - |
| Item 19 | 0.61 | - | 0.08 | - | - | - |
| Item 20 | 0.48 | - | 0.03 | - | - | - |
| Item 21 | 0.60 | - | -0.18 | - | - | - |
| Item 22 | 0.20 | - | -0.15 | - | - | - |
| Item 23 | 0.44 | - | -0.06 | - | - | - |
| Item 24 | 0.46 | - | -0.05 | - | - | - |
| Item 25 | 0.58 | - | -0.05 | - | - | - |
| Item 26 | 0.63 | - | -0.04 | - | - | - |
| Item 27 | 0.70 | - | -0.04 | - | - | - |
| Item 28 | 0.41 | - | - | 0.39 | - | - |
| Item 29 | 0.62 | - | - | 0.33 | - | - |
| Item 30 | 0.33 | - | - | 0.25 | - | - |
| Item 31 | 0.36 | - | - | 0.09 | - | - |
| Item 32 | 0.03 | - | - | 0.01 | - | - |
| Item 33 | 0.38 | - | - | -0.04 | - | - |
| Item 34 | 0.19 | - | - | -0.01 | - | - |
| Item 35 | 0.69 | - | - | -0.00 | - | - |
| Item 36 | 0.74 | - | - | - | 0.07 | - |
| Item 37 | 0.59 | - | - | - | 0.02 | - |
| Item 38 | 0.48 | - | - | - | 0.01 | - |
| Item 39 | 0.75 | - | - | - | 0.01 | - |

Table 3.A.6 *(continuation)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | General | CCC 1 | CCC 2 | CCC 3 | CCC 4 | CCC 5 |
| Item 40 | 0.18 | - | - | - | -0.83 | - |
| Item 41 | 0.54 | - | - | - | -0.09 | - |
| Item 42 | 0.73 | - | - | - | -0.04 | - |
| Item 43 | 0.35 | - | - | - | -0.01 | - |
| Item 44 | 0.51 | - | - | - | -0.01 | - |
| Item 45 | 0.69 | - | - | - | - | 0.47 |
| Item 46 | 0.47 | - | - | - | - | 0.12 |
| Item 47 | 0.32 | - | - | - | - | 0.06 |
| Item 48 | 0.44 | - | - | - | - | -0.04 |
| Item 49 | 0.55 | - | - | - | - | -0.04 |
| Item 50 | 0.55 | - | - | - | - | -0.01 |
| Item 51 | 0.41 | - | - | - | - | - |
| Item 52 | 0.42 | - | - | - | - | - |
| Item 53 | 0.51 | - | - | - | - | - |
| Item 54 | 0.55 | - | - | - | - | - |

Table 3.A.7 Factor Loading Matrix by Item Type—Grade Five, Form One, Two PTs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | CR | MC | TEI |
| Item 1 | 0.54 | 0.32 | - | - |
| Item 2 | 0.46 | 0.31 | - | - |
| Item 3 | 0.75 | 0.23 | - | - |
| Item 4 | 0.38 | 0.14 | - | - |
| Item 5 | 0.55 | -0.04 | - | - |
| Item 6 | 0.71 | -0.01 | - | - |
| Item 7 | 0.17 | - | 0.28 | - |
| Item 8 | 0.19 | - | 0.20 | - |
| Item 9 | 0.42 | - | 0.17 | - |
| Item 10 | 0.45 | - | 0.17 | - |
| Item 11 | 0.45 | - | 0.15 | - |
| Item 12 | 0.59 | - | 0.13 | - |
| Item 13 | 0.63 | - | 0.10 | - |
| Item 14 | 0.46 | - | 0.08 | - |
| Item 15 | 0.19 | - | 0.07 | - |
| Item 16 | 0.32 | - | 0.07 | - |
| Item 17 | 0.34 | - | 0.06 | - |
| Item 18 | 0.69 | - | 0.01 | - |
| Item 19 | 0.81 | - | -0.20 | - |
| Item 20 | 0.55 | - | -0.17 | - |
| Item 21 | 0.69 | - | -0.16 | - |
| Item 22 | 0.61 | - | -0.13 | - |
| Item 23 | 0.51 | - | -0.13 | - |
| Item 24 | 0.59 | - | -0.11 | - |
| Item 25 | 0.41 | - | -0.10 | - |
| Item 26 | 0.69 | - | -0.02 | - |
| Item 27 | 0.47 | - | -0.02 | - |
| Item 28 | 0.41 | - | - | 0.42 |
| Item 29 | 0.61 | - | - | 0.29 |
| Item 30 | 0.53 | - | - | 0.18 |
| Item 31 | 0.50 | - | - | 0.10 |
| Item 32 | 0.37 | - | - | 0.09 |
| Item 33 | 0.36 | - | - | 0.09 |
| Item 34 | 0.55 | - | - | 0.09 |
| Item 35 | 0.45 | - | - | 0.05 |
| Item 36 | 0.48 | - | - | 0.01 |
| Item 37 | 0.41 | - | - | 0.01 |
| Item 38 | 0.03 | - | - | 0.01 |
| Item 39 | 0.55 | - | - | -0.13 |

Table 3.A.7 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | CR | MC | TEI |
| Item 40 | 0.55 | - | - | -0.08 |
| Item 41 | 0.64 | - | - | -0.07 |
| Item 42 | 0.43 | - | - | -0.06 |
| Item 43 | 0.74 | - | - | -0.05 |
| Item 44 | 0.35 | - | - | -0.05 |
| Item 45 | 0.73 | - | - | -0.03 |
| Item 46 | 0.48 | - | - | -0.03 |
| Item 47 | 0.64 | - | - | -0.01 |

Table 3.A.8 Factor Loading Matrix by Item Type—Grade Five, Form One, Three PTs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | CR | MC | TEI |
| Item 1 | 0.46 | 0.31 | - | - |
| Item 2 | 0.54 | 0.31 | - | - |
| Item 3 | 0.75 | 0.23 | - | - |
| Item 4 | 0.38 | 0.14 | - | - |
| Item 5 | 0.22 | 0.05 | - | - |
| Item 6 | 0.55 | -0.04 | - | - |
| Item 7 | 0.70 | -0.01 | - | - |
| Item 8 | 0.17 | - | 0.28 | - |
| Item 9 | 0.44 | - | 0.20 | - |
| Item 10 | 0.19 | - | 0.20 | - |
| Item 11 | 0.44 | - | 0.19 | - |
| Item 12 | 0.41 | - | 0.18 | - |
| Item 13 | 0.59 | - | 0.15 | - |
| Item 14 | 0.34 | - | 0.15 | - |
| Item 15 | 0.63 | - | 0.13 | - |
| Item 16 | 0.46 | - | 0.10 | - |
| Item 17 | 0.19 | - | 0.09 | - |
| Item 18 | 0.32 | - | 0.07 | - |
| Item 19 | 0.34 | - | 0.06 | - |
| Item 20 | 0.69 | - | 0.04 | - |
| Item 21 | 0.57 | - | 0.03 | - |
| Item 22 | 0.82 | - | -0.16 | - |
| Item 23 | 0.55 | - | -0.16 | - |
| Item 24 | 0.65 | - | -0.13 | - |
| Item 25 | 0.69 | - | -0.11 | - |
| Item 26 | 0.61 | - | -0.11 | - |
| Item 27 | 0.51 | - | -0.11 | - |
| Item 28 | 0.42 | - | -0.10 | - |
| Item 29 | 0.59 | - | -0.08 | - |
| Item 30 | 0.59 | - | -0.03 | - |
| Item 31 | 0.69 | - | -0.02 | - |
| Item 32 | 0.47 | - | -0.01 | - |
| Item 33 | 0.41 | - | - | 0.42 |
| Item 34 | 0.62 | - | - | 0.27 |
| Item 35 | 0.53 | - | - | 0.17 |
| Item 36 | 0.50 | - | - | 0.11 |
| Item 37 | 0.55 | - | - | 0.10 |
| Item 38 | 0.37 | - | - | 0.09 |
| Item 39 | 0.36 | - | - | 0.09 |

Table 3.A.8 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | General | CR | MC | TEI |
| Item 40 | 0.45 | - | - | 0.04 |
| Item 41 | 0.41 | - | - | 0.02 |
| Item 42 | 0.29 | - | - | 0.02 |
| Item 43 | 0.55 | - | - | -0.15 |
| Item 44 | 0.62 | - | - | -0.11 |
| Item 45 | 0.55 | - | - | -0.10 |
| Item 46 | 0.63 | - | - | -0.07 |
| Item 47 | 0.74 | - | - | -0.06 |
| Item 48 | 0.43 | - | - | -0.06 |
| Item 49 | 0.35 | - | - | -0.05 |
| Item 50 | 0.48 | - | - | -0.03 |
| Item 51 | 0.72 | - | - | -0.02 |
| Item 52 | 0.64 | - | - | -0.02 |
| Item 53 | 0.03 | - | - | -0.01 |
| Item 54 | 0.48 | - | - | -0.00 |

Table 3.A.9 Factor Loading Matrix by Task Type—Grade Five, Form One, Two PTs

|  |  |  |  |
| --- | --- | --- | --- |
| Item | General | non-PT | PT |
| Item 1 | 0.52 | 0.34 | - |
| Item 2 | 0.78 | 0.32 | - |
| Item 3 | 0.43 | 0.21 | - |
| Item 4 | 0.54 | 0.19 | - |
| Item 5 | 0.50 | 0.18 | - |
| Item 6 | 0.62 | 0.17 | - |
| Item 7 | 0.53 | 0.17 | - |
| Item 8 | 0.67 | 0.16 | - |
| Item 9 | 0.60 | 0.15 | - |
| Item 10 | 0.37 | 0.14 | - |
| Item 11 | 0.73 | 0.14 | - |
| Item 12 | 0.40 | 0.14 | - |
| Item 13 | 0.68 | 0.12 | - |
| Item 14 | 0.33 | 0.10 | - |
| Item 15 | 0.59 | 0.10 | - |
| Item 16 | 0.63 | 0.10 | - |
| Item 17 | 0.72 | 0.07 | - |
| Item 18 | 0.48 | 0.07 | - |
| Item 19 | 0.41 | 0.06 | - |
| Item 20 | 0.43 | 0.05 | - |
| Item 21 | 0.48 | 0.04 | - |
| Item 22 | 0.45 | 0.04 | - |
| Item 23 | 0.58 | -0.21 | - |
| Item 24 | 0.47 | -0.18 | - |
| Item 25 | 0.47 | -0.16 | - |
| Item 26 | 0.43 | -0.10 | - |
| Item 27 | 0.38 | -0.10 | - |
| Item 28 | 0.65 | -0.10 | - |
| Item 29 | 0.20 | -0.09 | - |
| Item 30 | 0.56 | -0.07 | - |
| Item 31 | 0.60 | -0.06 | - |
| Item 32 | 0.70 | -0.04 | - |
| Item 33 | 0.36 | -0.03 | - |
| Item 34 | 0.46 | -0.03 | - |
| Item 35 | 0.41 | - | 0.40 |
| Item 36 | 0.61 | - | 0.29 |
| Item 37 | 0.33 | - | 0.21 |
| Item 38 | 0.53 | - | 0.20 |
| Item 39 | 0.17 | - | 0.16 |

Table 3.A.9 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Item | General | non-PT | PT |
| Item 40 | 0.50 | - | 0.12 |
| Item 41 | 0.31 | - | 0.12 |
| Item 42 | 0.18 | - | 0.07 |
| Item 43 | 0.02 | - | 0.04 |
| Item 44 | 0.70 | - | 0.03 |
| Item 45 | 0.47 | - | 0.01 |
| Item 46 | 0.55 | - | -0.10 |
| Item 47 | 0.75 | - | -0.02 |

Table 3.A.10 Factor Loading Matrix by Task Type—Grade Five, Form One, Three PTs

|  |  |  |  |
| --- | --- | --- | --- |
| Item | General | non-PT | PT |
| Item 1 | 0.55 | 0.30 | - |
| Item 2 | 0.81 | 0.21 | - |
| Item 3 | 0.45 | 0.16 | - |
| Item 4 | 0.55 | 0.13 | - |
| Item 5 | 0.55 | 0.12 | - |
| Item 6 | 0.51 | 0.12 | - |
| Item 7 | 0.38 | 0.11 | - |
| Item 8 | 0.42 | 0.10 | - |
| Item 9 | 0.63 | 0.09 | - |
| Item 10 | 0.61 | 0.09 | - |
| Item 11 | 0.69 | 0.07 | - |
| Item 12 | 0.34 | 0.06 | - |
| Item 13 | 0.74 | 0.05 | - |
| Item 14 | 0.69 | 0.04 | - |
| Item 15 | 0.64 | 0.04 | - |
| Item 16 | 0.59 | 0.03 | - |
| Item 17 | 0.48 | 0.03 | - |
| Item 18 | 0.41 | 0.01 | - |
| Item 19 | 0.48 | 0.00 | - |
| Item 20 | 0.56 | -0.28 | - |
| Item 21 | 0.45 | -0.24 | - |
| Item 22 | 0.44 | -0.20 | - |
| Item 23 | 0.63 | -0.17 | - |
| Item 24 | 0.41 | -0.14 | - |
| Item 25 | 0.55 | -0.13 | - |
| Item 26 | 0.37 | -0.12 | - |
| Item 27 | 0.59 | -0.12 | - |
| Item 28 | 0.70 | -0.12 | - |
| Item 29 | 0.19 | -0.10 | - |
| Item 30 | 0.46 | -0.07 | - |
| Item 31 | 0.36 | -0.06 | - |
| Item 32 | 0.43 | -0.00 | - |
| Item 33 | 0.45 | -0.00 | - |
| Item 34 | 0.72 | -0.00 | - |
| Item 35 | 0.40 | - | 0.36 |
| Item 36 | 0.61 | - | 0.29 |
| Item 37 | 0.52 | - | 0.24 |
| Item 38 | 0.16 | - | 0.21 |
| Item 39 | 0.34 | - | 0.20 |

Table 3.A.10 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Item | General | non-PT | PT |
| Item 40 | 0.33 | - | 0.18 |
| Item 41 | 0.31 | - | 0.16 |
| Item 42 | 0.49 | - | 0.14 |
| Item 43 | 0.59 | - | 0.13 |
| Item 44 | 0.18 | - | 0.10 |
| Item 45 | 0.57 | - | 0.08 |
| Item 46 | 0.29 | - | 0.07 |
| Item 47 | 0.70 | - | 0.06 |
| Item 48 | 0.47 | - | 0.05 |
| Item 49 | 0.66 | - | 0.04 |
| Item 50 | 0.03 | - | 0.03 |
| Item 51 | 0.55 | - | -0.07 |
| Item 52 | 0.62 | - | -0.03 |
| Item 53 | 0.76 | - | -0.02 |
| Item 54 | 0.23 | - | -0.01 |

## The Practicality of a Multistage Adaptive Test Design

### Study Purpose

Adaptive tests can provide more precise estimates of student ability, with improvement most notable at extreme ability levels (van der Linden, 2005). They do so by tailoring the difficulty of the test to the performance level of the student. As such, most of the items administered to each student are challenging but not discouragingly so.

Of the available adaptive test designs, multistage testing (MST) provides the best fit to the content requirements, measurement needs, and administrative circumstances of the California Science Test (CAST). The particular MST proposed for CAST is a two-stage design, where the first stage is a routing block and the second stage could include either two (i.e., easy and hard) or three (i.e., easy, medium, and hard) item blocks.

An MST panel that includes a router and three difficulty levels is illustrated in Figure 4.1, where S1 denotes the Stage 1 routing block and S2, the Stage 2 block. S1 and S2 together are termed a “panel.”

s<t1

t1≤s< t2

s≥t2

S1

S2 Medium

S2 Hard

S2 Easy

Figure 4.1 Typical MST Panel

All students begin by taking S1, a router test comprised primarily of moderately difficult items. Once the router test is complete, the algorithm decides which of the available second-stage item blocks is most appropriate given the student’s demonstrated performance. Students who struggled on the router are sent to second-stage blocks comprised primarily of easier items, while those who excel on the router are sent to blocks of more difficult items. Operational MST implementation may be supported by two or more panels. Each panel includes a group of blocks from different stages and can be used as the parallel form as in linear tests.

For an MST panel with three difficulty levels, there are two thresholds, *t1* and *t2*, to route students to the appropriate second-stage blocks. These decisions are made for each student in such a way that test precision is maximized for each as described in the study design (i.e., set thresholds at or close to the points at which the information curves for each S2 difficulty block cross). Students with a router score that is less than *t1* are assigned an easy block; students with router scores equal to or higher than the second threshold, *t2*, are assigned the hard block; all other students are routed the medium difficulty block. For an MST panel with two difficulty levels, a single threshold is used to assign students to easy and hard blocks.

The purpose of this study is to evaluate whether an adaptive Segment A is likely to offer substantial advantages in comparison to conventional linear forms with respect to measurement precision or efficiency. It must be noted that there is no absolute threshold separating improvements judged as substantial from those that are not. Rather, the judgment is whether the improvement is great enough to offset the complexity and risk inherent in all adaptive testing.

### Methods

The study requires first that MST panels be assembled from the available item pools and in accordance with all content and other substantive requirements. Once assembled, panels are then evaluated with respect to a variety of criteria and their performance compared to that of conventional linear forms. The following sections described both the assembly and evaluation procedures.

#### Assemble MST Panels with Two and Three Levels of Difficulty

The discrete items in segments A and C that passed the 2018–19 data review and the California Department of Education (CDE) item review form the pool to assemble the MST panels. Segment A of the linear CAST forms consists of 32 to 34 items totaling 42 to 44 score points. These basic requirements were essentially divided in half for the MST, with the router including 17 items (totaling 22 score points) while the second-stage blocks contained 16 items each (totaling 21 score points).

The assembled MST panels must also conform to additional substantive requirements, with these outlined in the left columns of Table 4.1 for grade five, Table 4.2 for grade eight, and Table 4.3 for high school. Requirements were imposed by specifying the minimum and maximum numbers of items that a test form (or item block) should contain. These requirements were developed by content experts and serve to both dictate what the test measures and to ensure that what is measured remains consistent across forms.

Full form content requirements were essentially divided in half to yield specifications for the router and second stages of the MST. These requirements are listed in the center and right columns of Table 4.1 through Table 4.3. The result is to guarantee that regardless of what route a student takes through the MST, a content-conforming test is administered.

Table 4.1 through Table 4.3 use the following abbreviations: Physical Sciences (PS), Life Sciences (LS), Earth and Space Sciences (ESS), Engineering, Technology, and Applications of Science (ETS), and Science and Engineering Practices (SEP).

Table 4.1 provides this information for grade five.

Table 4.1 Content Specification and Conformation for Grade Five

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Min Linear | Max Linear | Selected Linear | Min MST Router | Max MST Router | Selected MST Router | Min MST Second Stage | Max MST Second Stage | Easy MST Second Stage | Med MST Second Stage | Hard MST Second Stage |
| PS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| PS1 | 1 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| PS2 | 1 | 4 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| PS3 | 1 | 4 | 4 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| PS4 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| LS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| LS1 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| LS2 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| LS3 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| LS4 | 2 | 4 | 4 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| ESS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| ESS1 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| ESS2 | 1 | 5 | 5 | 1 | 3 | 3 | 0 | 2 | 2 | 2 | 2 |
| ESS3 | 1 | 3 | 3 | 1 | 2 | 1 | 0 | 1 | 2 | 2 | 2 |
| ETS Total | 2 | 4 | 3 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 |
| SEP 1 | 1 | 4 | 3 | 1 | 2 | 2 | 0 | 2 | 1 | 1 | 1 |
| SEP 2 | 1 | 7 | 4 | 1 | 2 | 2 | 0 | 2 | 2 | 2 | 2 |
| SEP 3 | 1 | 7 | 6 | 1 | 4 | 3 | 0 | 3 | 3 | 3 | 3 |
| SEP 4 | 2 | 4 | 4 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| SEP 5 | 1 | 2 | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 |
| SEP 6 | 2 | 8 | 4 | 1 | 4 | 2 | 1 | 4 | 2 | 2 | 2 |
| SEP 7 | 1 | 8 | 8 | 1 | 4 | 4 | 0 | 4 | 4 | 4 | 4 |
| SEP 8 | 1 | 3 | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 |
| Polytomous | 10 | 10 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |

Table 4.2 provides similar information for grade eight.

Table 4.2 Content Specification and Conformation for Grade Eight

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Min Linear | Max Linear | Selected Linear | Min MST Router | Max MST Router | Selected MST Router | Min MST Second Stage | Max MST Second Stage | Easy MST Second Stage | Med MST Second Stage | Hard MST Second Stage |
| PS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| PS1 | 1 | 5 | 3 | 1 | 3 | 2 | 0 | 2 | 1 | 1 | 1 |
| PS2 | 1 | 4 | 3 | 1 | 2 | 1 | 0 | 2 | 2 | 2 | 2 |
| PS3 | 1 | 4 | 3 | 1 | 2 | 1 | 0 | 2 | 2 | 2 | 2 |
| PS4 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| LS Total | 8 | 10 | 9 | 4 | 5 | 5 | 4 | 5 | 4 | 4 | 4 |
| LS1 | 1 | 6 | 2 | 1 | 3 | 1 | 0 | 3 | 1 | 1 | 1 |
| LS2 | 1 | 4 | 4 | 1 | 2 | 2 | 0 | 2 | 2 | 2 | 2 |
| LS3 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| LS4 | 1 | 5 | 2 | 1 | 3 | 1 | 0 | 2 | 1 | 1 | 1 |
| ESS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| ESS1 | 1 | 3 | 3 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 1 |
| ESS2 | 1 | 5 | 5 | 1 | 3 | 2 | 0 | 2 | 3 | 3 | 3 |
| ESS3 | 1 | 4 | 2 | 1 | 2 | 1 | 0 | 2 | 1 | 1 | 1 |
| ETS Total | 2 | 4 | 4 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| ETS1 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| ETS2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| ETS3 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| SEP 1 | 1 | 3 | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 |
| SEP 2 | 1 | 16 | 6 | 1 | 8 | 3 | 0 | 8 | 3 | 3 | 3 |
| SEP 3 | 1 | 5 | 1 | 1 | 3 | 1 | 0 | 2 | 0 | 0 | 0 |
| SEP 4 | 1 | 9 | 8 | 1 | 5 | 4 | 0 | 4 | 4 | 4 | 4 |
| SEP 5 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| SEP 6 | 1 | 12 | 8 | 1 | 6 | 5 | 0 | 6 | 3 | 3 | 3 |
| SEP 7 | 1 | 8 | 5 | 1 | 4 | 1 | 0 | 4 | 4 | 4 | 4 |
| SEP 8 | 1 | 4 | 1 | 1 | 2 | 1 | 0 | 2 | 0 | 0 | 0 |
| Polytomous | 10 | 10 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |

Table 4.3 provides similar information for high school.

Table 4.3 Content Specification and Conformation for High School

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Min Linear | Max Linear | Selected Linear | Min MST Router | Max MST Router | Selected MST Router | Min MST Second Stage | Max MST Second Stage | Easy MST Second Stage | Med MST Second Stage | Hard MST Second Stage |
| PS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| PS1 | 1 | 5 | 2 | 1 | 3 | 1 | 0 | 2 | 1 | 1 | 1 |
| PS2 | 1 | 4 | 3 | 1 | 3 | 1 | 0 | 2 | 2 | 2 | 2 |
| PS3 | 1 | 4 | 2 | 1 | 2 | 1 | 0 | 2 | 1 | 1 | 1 |
| PS4 | 1 | 2 | 3 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 1 |
| LS Total | 8 | 10 | 10 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| LS1 | 1 | 6 | 2 | 1 | 3 | 1 | 0 | 3 | 1 | 1 | 2 |
| LS2 | 1 | 4 | 3 | 1 | 4 | 1 | 0 | 2 | 3 | 2 | 1 |
| LS3 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| LS4 | 1 | 5 | 3 | 1 | 3 | 2 | 0 | 2 | 1 | 1 | 1 |
| ESS Total | 8 | 10 | 9 | 4 | 5 | 5 | 4 | 5 | 5 | 4 | 4 |
| ESS1 | 1 | 3 | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 |
| ESS2 | 1 | 5 | 3 | 1 | 3 | 1 | 0 | 2 | 3 | 2 | 2 |
| ESS3 | 1 | 4 | 4 | 1 | 3 | 3 | 0 | 2 | 1 | 1 | 1 |
| ETS Total | 3 | 4 | 4 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 |
| ETS1 | 1 | 2 | 2 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| ETS2 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| ETS3 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| SEP 1 | 1 | 3 | 3 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 1 |
| SEP 2 | 1 | 16 | 4 | 1 | 3 | 2 | 0 | 8 | 3 | 2 | 2 |
| SEP 3 | 1 | 5 | 4 | 1 | 3 | 2 | 0 | 2 | 2 | 2 | 1 |
| SEP 4 | 1 | 9 | 5 | 1 | 3 | 3 | 0 | 4 | 2 | 2 | 2 |
| SEP 5 | 1 | 2 | 3 | 1 | 3 | 2 | 0 | 1 | 1 | 1 | 2 |
| SEP 6 | 1 | 12 | 5 | 1 | 3 | 2 | 0 | 6 | 3 | 3 | 3 |
| SEP 7 | 1 | 8 | 3 | 1 | 2 | 1 | 0 | 4 | 2 | 2 | 2 |
| SEP 8 | 1 | 4 | 6 | 1 | 3 | 3 | 0 | 2 | 2 | 3 | 3 |
| Polytomous | 10 | 10 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |

Statistical requirements were also imposed on each item block. This was done by setting targets for the item response theory (IRT) information that each block was to contribute to measurement. The information functions for the item blocks that comprise an MST can provide a concise visual impression of how the test will perform. The heights of the information functions are proportional to test precision, while the separation between higher-stage blocks indicates the breadth of the proficiency range across which measurement is good. The target information functions were roughly bell-shaped and centered at midrange proficiency values for the router and the medium-difficulty second-stage blocks. The easy and hard second-stage block targets were centered at proficiency values roughly .6 standard deviations below and above mean levels, respectively. A sample of information targets for blocks is shown in Figure 4.2, next.

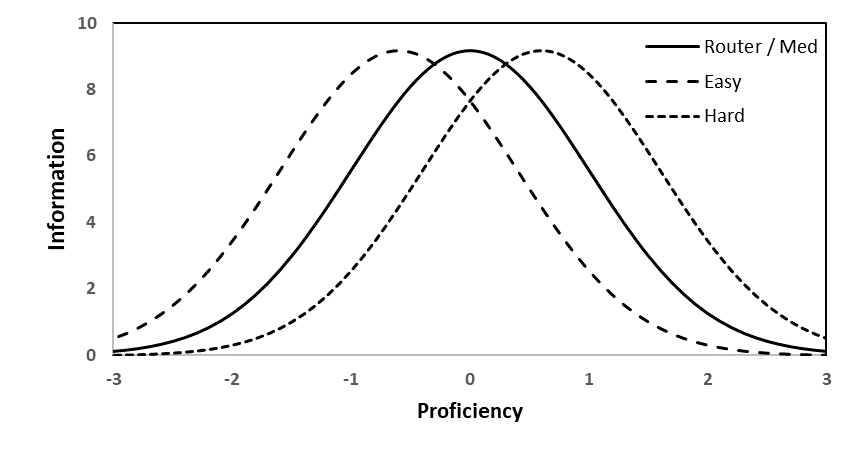


Figure 4.2 Sample Block Information Targets

Blocks were assembled using an automated test assembly method based on an optimization method called mixed-integer programming (refer to van der Linden [2005] for a discussion of the method).

#### Evaluate Panel Performance

Simulations are commonly used to predict or project how an adaptive test is likely to perform when administered operationally. Simulations can be used as the basis for strong assumptions because they are based on the IRT parameters that characterize students and items.

Although MST performance could be evaluated by simulations, many performance characteristics also can be evaluated by what are essentially analytic methods. In the MST Practicality Study, the recursive algorithm suggested by Lord and Wingersky (1984) was used to calculate the observed score distributions for fixed items blocks directly. This was then the basis to compute multiple evaluation criteria and is described in detail in the next subsection.

The recursive algorithm is based on the IRT probabilities of students at specific proficiency levels answering each item correctly. These probabilities can be presented as *P1(t), P2(t),… Pn(t)* across the *n* test items, where *t* is a particular proficiency value. Suppose the test consists of only the first item, and the probabilities of each possible score (0 and 1) are given by *Q1(t)* = 1-*P1(t)* and *P1(t).* Adding the second item to this test now allows three possible scores: 0, 1, and 2, where the probabilities of each are: *Q1(t)Q2(t), P1(t)Q2(t) + Q1(t)P2(t),* and *P1(t)P2(t)*. Each item is added to the test in turn, adjusting the accumulating probabilities under the condition that the new item is answered correctly or incorrectly. Once the final item is added, the probabilities become the distribution of total score for the test. Although this example assumes that items are scored dichotomously—correct or incorrect—the algorithm generalizes readily to items scored in more than two categories.

The expected score distributions produced by the recursion are specific to or conditional upon each particular proficiency level, *t*. To compute the score distribution for the entire student population, these conditional distributions must be summed across proficiency levels and then weighted by the population distribution of proficiency. Computing score distributions analytically is both computationally simpler and more precise than simulating large numbers of test responses. It will be shown that most of the performance characteristics of MSTs can be projected directly from conditional (proficiency-specific) and unconditional (population-specific) score distributions.

##### Measurement Precision

The performance of MST panels was evaluated against linear forms in terms of measurement precision as expressed by conditional standard errors of measurement (CSEM). The linear forms in this study were defined by the combination of the router and the medium difficulty blocks.

CAST scoring uses equated observed scores (point totals) rather than maximum likelihood estimates. Measure precision (i.e., CSEM) can be projected or estimated using the conditional recursion-based score distributions. The CSEM is the standard deviation (SD) of the distribution of estimated proficiency at the given true proficiency value.

To calculate the CSEM, a raw-score (i.e., point total) distribution is first computed for the router using the recursive algorithm at each grid of proficiency values. Each point total maps uniquely to a proficiency value through the test characteristic curve.

A second recursion is used to produce raw-score distributions for each test path (e.g., router plus the easy block). These are the distributions of total-path raw scores, or the sums of router and second-stage scores. Again, each total-test raw score maps uniquely to a proficiency value.

The distribution of total-test proficiency estimates is computed by summing the distributions of each path, with weights given by the probabilities of students taking each path. This results in the distribution of estimated proficiency given a true proficiency value.

The CSEM at the given true proficiency value is then computed as the SD of the distribution of estimated proficiency. For example,

  (4.1)

where,

*h* corresponds to a particular proficiency value, and

*K* corresponds to the number of proficiency values in the grid.

The variance for the combined test over all students is obtained, regardless of which difficulty tier the students were routed to, by the law of total probability. That is, it is the sum of (1) the weighted average of the three variances (for easy, medium, and hard combined tests) and (2) the variance of the weighted means, with weights in each case equal to the classification probabilities for each difficulty tier. The CSEM is then the square root of these conditional variances.

Figure 4.3 shows an example of a CSEM function.

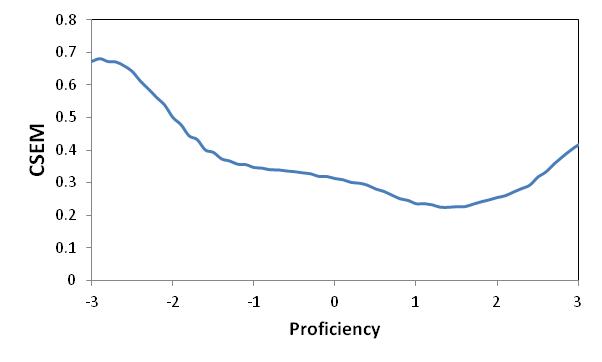


Figure 4.3 Example Conditional Standard Errors of Measurement

##### Conditional Bias

The conditional bias is the difference between the expected value of the proficiency estimates and the true proficiency value. The expected value of the proficiency estimates can be calculated as the weighted average of the conditional path expected value (refer to equation 4.1 for the formula), weighting by the routing rate to each of the easy, medium, and difficult paths.

##### Relative Efficiency

The relative efficiency of a test score *y* with respect to a test score *x* is defined as the ratio of their information functions. For example, if test score *y* comes from a linear form and test score *x* comes from an MST form, then the relative efficiency of a linear form compared to an MST design is the ratio of their information functions given the same ability.

##### Test Reliability

Marginal test reliabilities were also computed for both MST and linear forms. Reliability is one of the standard indices of the measurement strength of a test. It can be computed from the student ability distribution and the CSEMs discussed in the previous section. Since the latter depends in part on the decision thresholds, reliability, too, varies somewhat according to the choice of those thresholds. All things considered, thresholds are chosen to maximize reliability. Test reliability for an MST design is compared against the test reliability for the linear form that is comprised of a router block and a medium-difficulty, second-stage block.

Specifically, the marginal reliability is

 (4.2)

where,

 (where  is the density from the student ability distribution for ), and

, where  is the square of the CSEMs discussed in the previous section.

### Item Pools

Summary statistics for difficulty parameters from each grade’s item pool can be found in Table 4.4 through Table 4.6. The pool includes all items that were used in segments A and C in the 2018–19 operational administration and are eligible for use in future forms. The total numbers of items for grade five, grade eight, and high school are 209, 201, and 195 respectively, after excluding outlying items with extreme difficulty parameters (i.e., items with IRT b-parameters greater than 4 or lower than -4). The mean difficulty parameters for grade eight and high school are positive, indicating the pools are difficult in general. This is especially true for the high school test, where the mean difficulty for the pool is 0.78. The mean difficulty parameter for the entire pool for grade five is slightly lower than zero, indicating the pool is easy in general.

Note that the pool described in this section includes all the items available for the MST assembly at the time of this study. However, not all items have an equal probability of being selected due to the content constraints. Therefore, even though the pool may appear to have difficult items, those items might not be eligible for use in assembly and the results from the assembly might still suggest a lack of difficult items to support building the difficult blocks. The development of a more robust item pool is an ongoing goal for the CAST.

Table 4.4 provides this information for grade five.

Table 4.4 Summary of Difficulty Parameters for Grade Five

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Content | Total Number of Items | Mean b | Min b | Max b |
| Entire Pool | 209 | -0.11 | -2.73 | 3.63 |
| CR | 14 | -0.14 | -1.32 | 0.97 |
| MC | 95 | -0.01 | -1.67 | 3.63 |
| TEI | 100 | -0.20 | -2.73 | 3.14 |
| ESS | 64 | 0.09 | -1.35 | 3.63 |
| LS | 68 | -0.41 | -2.32 | 3.48 |
| PS | 77 | -0.01 | -2.73 | 3.14 |

Table 4.5 provides this information for grade eight.

Table 4.5 Summary of Difficulty Parameters for Grade Eight

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Content | Total Number of Items | Mean b | Min b | Max b |
| Entire Pool | 201 | 0.46 | -1.18 | 3.44 |
| CR | 15 | 0.14 | -0.94 | 0.84 |
| MC | 89 | 0.49 | -1.18 | 3.44 |
| TEI | 97 | 0.49 | -1.10 | 3.02 |
| ESS | 63 | 0.59 | -1.08 | 3.11 |
| LS | 61 | 0.43 | -1.18 | 2.41 |
| PS | 77 | 0.38 | -1.10 | 3.44 |

Table 4.6 provides this information for high school.

Table 4.6 Summary of Difficulty Parameters for High School

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Content | Total Number of Items | Mean b | Min b | Max b |
| Entire Pool | 195 | 0.72 | -1.33 | 3.91 |
| CR | 14 | 1.03 | 0.03 | 1.47 |
| MC | 100 | 0.67 | -1.21 | 3.91 |
| TEI | 81 | 0.73 | -1.33 | 3.87 |
| ESS | 67 | 0.47 | -1.33 | 3.24 |
| LS | 63 | 0.66 | -1.21 | 3.91 |
| PS | 65 | 1.05 | -0.86 | 3.70 |

Figure 4.4 through Figure 4.6 present the scatter plots of the discrimination and the difficulty parameters.

Negative correlations between the discrimination and the difficulty parameters were observed for all three grades, indicating that difficult items tend to have low discriminating power.

Figure 4.4 for grade five uses the data from Table 4.A.1

Figure 4.4 Scatter Plot of Discrimination and Difficulty Parameters for Grade Five

Figure 4.5 uses the data from Table 4.A.2 .

Figure 4.5 Scatter Plot of Discrimination and Difficulty Parameters for Grade Eight

Figure 4.6 uses the data from Table 4.A.3.

Figure 4.6 Scatter Plot of Discrimination and Difficulty Parameters for High School

### Results

#### Information Functions

Based on the calibrated item pools of grade five, grade eight, and high school, two MST panels—one for the MST 1-2 design and one for the MST 1-3 design—were assembled using an automated test assembly (ATA) tool. An ATA tool is a computer program utilizing linear programming methods that can automatically assemble tests that satisfy both content and statistical requirements.

##### Grade Five

MST block assembly is characterized by conformance to the substantive (with the only exception of one item over for ESS3 for the second-stage blocks) and the statistical requirements. The former is shown in the center and right columns of Table 4.1, where the numbers of items selected into each block type are compared to the numbers of items required. All grade five blocks conformed to all requirements, helping to ensure that measurement is substantively equivalent regardless of the specific test administered to any student.

Conformance to block information targets was more equivocal, largely reflecting disconnects between the targets and the available item resources. In many cases, it proved difficult or impossible to closely meet information targets, either because items at the necessary difficulty levels were unavailable or because content requirements rendered them unusable.

The information functions for the grade five MST 1-2 and MST 1-3 design panels are shown in Figure 4.7 and Figure 4.8, respectively. The data used to create Figure 4.7 can be found in Table 4.A.4 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.7 Information Curves of the Assembled MST 1-2 Design Panel for Grade Five

The data used to create Figure 4.8 can be found in Table 4.A.5 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.8 Information Curves of the Assembled MST 1-3 Design Panel for Grade Five

The block information functions demonstrate that in grade five, the principal problems centered on the hard second-stage block. Since the bulk of the information in the current assembly pool was on the easier end of the performance spectrum, it was not possible to create a block that was difficult enough and still met content requirements. Because the hard second-stage block was scarcely distinguished from the medium block, it is not surprising that little performance difference was found between two- and three-level MSTs.

##### Grade Eight

Table 4.2 again shows that all MST blocks conformed to the substantive requirements (with the exception of one item over for ESS2 for the second-stage blocks).

The information functions for the grade eight MST 1-2 and MST 1-3 design panels are shown in Figure 4.9 and Figure 4.10, respectively. The data used to create Figure 4.9 can be found in Table 4.A.6 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.9  Information Curves of the Assembled MST 1-2 Design Panel for Grade Eight

The data used to create Figure 4.10 can be found in Table 4.A.7 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.10 Information Curves of the Assembled MST 1-3 Design Panel for Grade Eight

The information functions show that the grade eight item pool allowed clear distinctions between the easy, medium, and hard second-stage blocks. The fit of the selected blocks to the information targets was much better than in grade five, demonstrating that the current grade eight item pool is much better suited to MST administration. However, the easy, medium, and hard second-stage information functions all intersect at the same point, meaning that the middle path is superfluous and a two-level design will once again perform equivalently to a three-level design.

##### High School

Block assembly was somewhat more difficult for high school in comparison to the other grade levels. Table 4.3 shows that the high school second-stage blocks do not all conform identically with the assembly requirements, as was the case with grades five and eight. However, this is not a major concern because the total form administered to every student does conform; it simply does so in a slightly different way depending on the path taken.

The information functions for the high school MST 1-2 and MST 1-3 design panels are shown in Figure 4.11 and Figure 4.12, respectively. The data used to create Figure 4.11 can be found in Table 4.A.8 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.11 Information Curves of the Assembled MST 1-2 Design Panel for High School

The data used to create Figure 4.12 can be found in Table 4.A.9 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.12 Information Curves of the Assembled MST 1-3 Design Panel for High School

With regard to information requirements, Figure 4.11 and Figure 4.12 show that the second-stage item blocks are clearly separated in the level of difficulty. In fact, high school is unique in that the performance of the three-level test is not identical to that of the two-level test.

#### Conditional Standard Error of Measurement (CSEM)

Figure 4.13 through Figure 4.15 show the CSEMs for grade five, grade eight, and high school, respectively. Data used to create Figure 4.13 can be found in Table 4.A.10 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.13 Conditional Standard Error of Measurement for Grade Five

Data used to create Figure 4.14 can be found in Table 4.A.11 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.14 Conditional Standard Error of Measurement for Grade Eight

Data used to create Figure 4.15 can be found in Table 4.A.12 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.15 Conditional Standard Error of Measurement for High School

Three designs—the MST 1-2, MST 1-3, and linear form—were compared in each of these figures. The pattern was consistent across grades, and the CSEMs of the MST 1‑2 design and MST 1-3 design were very similar—so similar, in fact, that the indicators for MST 1-2 and MST 1-3 cannot be distinguished on all graphs, which makes the simpler MST 1-2 design the clear choice.

For all three grades, the MST is more precise than the linear form, at both the high and low proficiency extremes. And for all three grades, linear forms were comparable to the MST panels at the center of the student proficiency scale.

#### Conditional Bias

Figure 4.16 through Figure 4.18 show the conditional bias for grade five, grade eight, and high school respectively. Data used to create Figure 4.16 can be found in Table 4.A.10 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.16 Conditional Bias for Grade Five

Data used to create Figure 4.17 can be found in Table 4.A.11 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.17 Conditional Bias for Grade Eight

Data used to create Figure 4.18 can be found in Table 4.A.12 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.18 Conditional Bias for High School

The conditional bias for MST 1-2 and MST 1-3 designs are so similar that they cannot be distinguished on all graphs.

For grade eight and high school, the linear form has larger conditional bias than the MST throughout most of the proficiency range. For grade five, the bias for the linear form is larger than that for the MST panel in the low proficiency range and is somewhat close for the remaining parts of the proficiency range.

#### Relative Efficiency

Figure 4.19 through Figure 4.21 show the relative efficiency for grade five, grade eight, and high school, respectively. Data used to create Figure 4.19 can be found in Table 4.A.16 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.19 Relative Efficiency for Grade Five

Data used to create Figure 4.20 can be found in Table 4.A.17 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.20 Relative Efficiency for Grade Eight

Data used to create Figure 4.21 can be found in Table 4.A.18 in [appendix 4.A](#_Appendix_2.A:_Table).

Figure 4.21 Relative Efficiency for High School

Relative efficiency is defined as the ratio of the information from the linear form to that from the MST panel. If the relative efficiency is greater than one, the linear form is more efficient. Otherwise, the MST panel is more efficient. For example, if the ratio is 0.5, it means the linear form would have to be double its test length to be as efficient as the MST form.

For grade five, relative efficiency is never greater than one, indicating MST outperformed the linear form in terms of measurement efficiency at all ability levels. For grade eight and high school, the relative efficiency is slightly larger than one near the center of ability scale, indicating the linear form had more measurement efficiency at the average abilities; and smaller than one at extreme ability levels, indicating that the MST panel is more efficient at the extreme ability levels.

#### Reliability

Table 4.7 shows the marginal test reliability for each design and each grade. The reliabilities of the MST 1-3 and 1-2 designs are almost identical. The reliabilities of the MST designs make modest improvements over of the linear forms.

Table 4.7 Marginal Test Reliabilities

|  |  |  |  |
| --- | --- | --- | --- |
| Grade Level | MST 1-3 Design | MST 1-2 Design | Linear Form |
| Grade Five | 0.89 | 0.89 | 0.88 |
| Grade Eight | 0.91 | 0.91 | 0.89 |
| High School | 0.91 | 0.91 | 0.89 |

### Implications for Test Blueprint Change

For the two grades and the grade span, MST performance exceeds that of linear forms for both the lower and higher end of the ability continuum. The improvement is relatively modest except at far extremes of the proficiency range.

The relative strengths and weaknesses of the assembly pools are largely reflected in the degree and location of the measurement improvement. The grade five pool is strongest at lower levels of student performance, and that is where the MST is strongest as well. The grade eight pool is well balanced in terms of item difficulty and MST performance particularly at both lower and upper proficiency ranges. The high school pool also shows improvement at both the lower and upper ranges.

Given the relatively modest improvement in measurement precision and since the full benefit of the MST panel can be realized once the pool is expanded to support more differentiable easy and hard blocks, it is recommended to hold off on the implementation of the MST design and reevaluate once the pool expands beyond the current level.

### References

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement,* *8,* 453–61.

van der Linden, W. (2005). *Linear models for optimal test design*. New York, NY: Springer.

### Appendix 4.A: Table Data for Graphs

#### Data for Scatter Plots of Discrimination and Difficulty Parameters

Table 4.A.1 Data for Scatter Plot of Discrimination and Difficulty Parameters—Grade Five

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| -0.32 | -2.73 |
| 0.51 | -2.32 |
| 0.53 | -1.96 |
| 1.01 | -1.85 |
| 0.15 | -1.82 |
| 0.80 | -1.75 |
| 0.47 | -1.74 |
| 0.74 | -1.67 |
| 1.22 | -1.66 |
| 0.94 | -1.63 |
| 1.01 | -1.61 |
| 0.38 | -1.58 |
| 0.53 | -1.52 |
| 0.52 | -1.52 |
| 0.33 | -1.47 |
| 0.61 | -1.46 |
| 0.74 | -1.40 |
| 0.52 | -1.39 |
| 0.71 | -1.35 |
| 0.86 | -1.34 |
| 0.43 | -1.32 |
| 0.34 | -1.31 |
| 1.21 | -1.22 |
| 1.36 | -1.16 |
| 0.42 | -1.16 |
| 0.71 | -1.13 |
| 0.92 | -1.13 |
| 0.68 | -1.03 |
| 0.79 | -1.01 |
| 0.64 | -1.00 |
| 0.28 | -0.99 |
| 0.64 | -0.98 |
| 0.52 | -0.94 |

Table 4.A.1 *(continuation one)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.66 | -0.94 |
| 0.45 | -0.91 |
| 0.81 | -0.90 |
| 0.67 | -0.90 |
| 0.48 | -0.88 |
| 0.48 | -0.86 |
| 0.82 | -0.82 |
| 0.96 | -0.80 |
| 0.75 | -0.80 |
| 0.68 | -0.78 |
| 0.70 | -0.77 |
| 0.96 | -0.74 |
| 0.84 | -0.73 |
| 0.38 | -0.73 |
| 0.89 | -0.70 |
| 0.74 | -0.68 |
| 1.02 | -0.65 |
| 0.62 | -0.62 |
| 0.91 | -0.60 |
| 1.06 | -0.59 |
| 0.76 | -0.59 |
| 0.84 | -0.59 |
| 0.44 | -0.58 |
| 0.81 | -0.58 |
| 1.00 | -0.57 |
| 0.65 | -0.57 |
| 0.44 | -0.57 |
| 0.64 | -0.56 |
| 0.68 | -0.55 |
| 0.59 | -0.54 |
| 0.61 | -0.53 |
| 0.41 | -0.53 |
| 0.86 | -0.50 |
| 0.37 | -0.50 |
| 0.98 | -0.50 |
| 0.90 | -0.49 |
| 0.80 | -0.47 |
| 0.67 | -0.46 |

Table 4.A.1 *(continuation two)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.81 | -0.45 |
| 0.76 | -0.45 |
| 0.44 | -0.44 |
| 0.52 | -0.39 |
| 1.11 | -0.39 |
| 0.95 | -0.37 |
| 0.30 | -0.36 |
| 1.16 | -0.35 |
| 0.36 | -0.35 |
| 0.62 | -0.35 |
| 0.81 | -0.34 |
| 0.84 | -0.32 |
| 0.55 | -0.32 |
| 0.36 | -0.32 |
| 1.06 | -0.32 |
| 0.69 | -0.31 |
| 0.71 | -0.30 |
| 0.49 | -0.28 |
| 0.69 | -0.28 |
| 0.54 | -0.27 |
| 0.76 | -0.27 |
| 0.99 | -0.26 |
| 0.49 | -0.25 |
| 0.98 | -0.24 |
| 0.64 | -0.23 |
| 0.78 | -0.23 |
| 0.55 | -0.23 |
| 0.39 | -0.23 |
| 0.99 | -0.22 |
| 0.71 | -0.21 |
| 0.46 | -0.21 |
| 0.94 | -0.21 |
| 0.74 | -0.20 |
| 0.63 | -0.19 |
| 0.53 | -0.19 |
| 0.72 | -0.18 |
| 0.75 | -0.18 |
| 0.60 | -0.16 |

Table 4.A.1 *(continuation three)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.64 | -0.16 |
| 0.79 | -0.15 |
| 0.72 | -0.14 |
| 0.52 | -0.14 |
| 0.53 | -0.13 |
| 0.45 | -0.12 |
| 0.70 | -0.12 |
| 0.36 | -0.10 |
| 0.69 | -0.09 |
| 0.39 | -0.08 |
| 0.86 | -0.07 |
| 0.61 | -0.07 |
| 0.59 | -0.06 |
| 0.60 | -0.04 |
| 0.72 | -0.04 |
| 1.02 | -0.04 |
| 0.51 | -0.03 |
| 0.62 | -0.03 |
| 1.01 | 0.02 |
| 1.25 | 0.03 |
| 0.65 | 0.03 |
| 0.54 | 0.06 |
| 0.38 | 0.06 |
| 0.63 | 0.07 |
| 0.49 | 0.08 |
| 0.67 | 0.08 |
| 0.45 | 0.11 |
| 0.42 | 0.11 |
| 0.61 | 0.13 |
| 0.45 | 0.15 |
| 0.79 | 0.15 |
| 0.49 | 0.16 |
| 0.32 | 0.17 |
| 0.67 | 0.18 |
| 0.56 | 0.19 |
| 0.34 | 0.19 |
| 0.49 | 0.21 |
| 0.81 | 0.22 |

Table 4.A.1 *(continuation four)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.42 | 0.25 |
| 0.71 | 0.26 |
| 0.46 | 0.26 |
| 0.53 | 0.26 |
| 0.57 | 0.27 |
| 0.99 | 0.29 |
| 0.56 | 0.30 |
| 0.26 | 0.30 |
| 0.65 | 0.36 |
| 0.47 | 0.38 |
| 0.34 | 0.40 |
| 0.48 | 0.41 |
| 0.45 | 0.41 |
| 0.69 | 0.46 |
| 0.43 | 0.46 |
| 1.24 | 0.49 |
| 0.44 | 0.53 |
| 0.55 | 0.56 |
| 0.51 | 0.56 |
| 0.45 | 0.58 |
| 0.29 | 0.62 |
| 0.49 | 0.62 |
| 0.62 | 0.69 |
| 0.85 | 0.71 |
| 0.36 | 0.74 |
| 0.45 | 0.78 |
| 0.61 | 0.80 |
| 0.69 | 0.83 |
| 0.69 | 0.84 |
| 0.65 | 0.85 |
| 0.89 | 0.85 |
| 0.37 | 0.85 |
| 0.56 | 0.86 |
| 0.57 | 0.88 |
| 0.55 | 0.89 |
| 0.32 | 0.92 |
| 0.65 | 0.92 |
| 0.52 | 0.97 |

Table 4.A.1 *(continuation five)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.39 | 1.00 |
| 0.28 | 1.14 |
| 0.44 | 1.17 |
| 0.52 | 1.17 |
| 0.42 | 1.34 |
| 0.25 | 1.36 |
| 0.46 | 1.40 |
| 0.36 | 1.40 |
| 0.31 | 1.41 |
| 0.28 | 1.43 |
| 0.58 | 1.49 |
| 0.16 | 1.58 |
| 0.19 | 1.60 |
| 0.66 | 1.62 |
| 0.32 | 1.75 |
| 0.18 | 1.79 |
| 0.76 | 2.01 |
| 0.48 | 2.09 |
| 0.55 | 2.49 |
| 0.31 | 3.14 |
| 0.12 | 3.48 |
| 0.22 | 3.63 |

Table 4.A.2 Data for Scatter Plot of Discrimination and Difficulty Parameters—Grade Eight

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.98 | -1.18 |
| 0.55 | -1.10 |
| 0.71 | -1.10 |
| 0.91 | -1.08 |
| 0.39 | -0.95 |
| 0.51 | -0.94 |
| 1.23 | -0.92 |
| 0.48 | -0.90 |
| 0.84 | -0.88 |
| 1.42 | -0.86 |
| 0.61 | -0.86 |
| 1.04 | -0.84 |
| 0.38 | -0.80 |
| 0.79 | -0.78 |
| 0.64 | -0.75 |
| 0.94 | -0.75 |
| 1.05 | -0.71 |
| 0.60 | -0.69 |
| 0.69 | -0.69 |
| 0.61 | -0.67 |
| 0.55 | -0.65 |
| 0.76 | -0.65 |
| 0.28 | -0.63 |
| 0.72 | -0.57 |
| 0.76 | -0.57 |
| 0.98 | -0.56 |
| 0.76 | -0.56 |
| 0.96 | -0.56 |
| 0.56 | -0.55 |
| 0.84 | -0.55 |
| 1.27 | -0.47 |
| 0.93 | -0.46 |
| 0.41 | -0.46 |
| 0.86 | -0.45 |
| 0.68 | -0.43 |
| 0.57 | -0.42 |
| 0.36 | -0.41 |

Table 4.A.2 *(continuation one)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.58 | -0.40 |
| 0.66 | -0.40 |
| 0.48 | -0.40 |
| 0.52 | -0.39 |
| 0.76 | -0.39 |
| 0.59 | -0.38 |
| 0.66 | -0.37 |
| 0.54 | -0.36 |
| 1.04 | -0.34 |
| 0.63 | -0.33 |
| 0.77 | -0.33 |
| 0.58 | -0.33 |
| 0.27 | -0.29 |
| 0.55 | -0.28 |
| 0.85 | -0.25 |
| 0.47 | -0.23 |
| 0.68 | -0.21 |
| 0.77 | -0.20 |
| 0.81 | -0.19 |
| 0.84 | -0.15 |
| 1.01 | -0.12 |
| 0.89 | -0.12 |
| 0.62 | -0.11 |
| 0.80 | -0.11 |
| 0.35 | -0.10 |
| 0.61 | -0.10 |
| 0.55 | -0.07 |
| 0.58 | -0.07 |
| 0.86 | -0.06 |
| 1.05 | -0.06 |
| 0.67 | -0.03 |
| 0.59 | -0.02 |
| 0.73 | -0.02 |
| 0.62 | 0.02 |
| 0.55 | 0.03 |
| 0.96 | 0.04 |
| 0.38 | 0.04 |
| 1.11 | 0.05 |

Table 4.A.2 *(continuation two)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.47 | 0.06 |
| 0.37 | 0.07 |
| 0.44 | 0.10 |
| 0.58 | 0.13 |
| 0.72 | 0.14 |
| 0.68 | 0.15 |
| 0.92 | 0.15 |
| 0.75 | 0.15 |
| 0.90 | 0.17 |
| 0.86 | 0.18 |
| 0.59 | 0.19 |
| 0.58 | 0.20 |
| 0.56 | 0.24 |
| 0.95 | 0.25 |
| 0.46 | 0.25 |
| 0.61 | 0.27 |
| 0.78 | 0.28 |
| 0.81 | 0.29 |
| 0.82 | 0.30 |
| 0.48 | 0.30 |
| 0.61 | 0.31 |
| 0.79 | 0.34 |
| 0.58 | 0.34 |
| 0.85 | 0.34 |
| 0.47 | 0.35 |
| 0.08 | 0.35 |
| 0.58 | 0.37 |
| 0.62 | 0.39 |
| 0.47 | 0.42 |
| 0.55 | 0.42 |
| 0.96 | 0.44 |
| 1.02 | 0.44 |
| 0.35 | 0.45 |
| 0.42 | 0.45 |
| 0.74 | 0.46 |
| 0.99 | 0.46 |
| 0.56 | 0.47 |
| 0.88 | 0.50 |

Table 4.A.2 *(continuation three)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.41 | 0.51 |
| 0.34 | 0.52 |
| 0.45 | 0.52 |
| 0.45 | 0.53 |
| 0.70 | 0.54 |
| 0.91 | 0.55 |
| 0.55 | 0.56 |
| 0.38 | 0.56 |
| 0.66 | 0.56 |
| 0.95 | 0.57 |
| 0.26 | 0.57 |
| 0.90 | 0.57 |
| 0.62 | 0.57 |
| 0.48 | 0.58 |
| 0.38 | 0.59 |
| 0.56 | 0.59 |
| 0.69 | 0.70 |
| 0.39 | 0.73 |
| 0.55 | 0.74 |
| 0.45 | 0.76 |
| 0.47 | 0.81 |
| 0.47 | 0.81 |
| 0.59 | 0.83 |
| 0.37 | 0.84 |
| 0.76 | 0.84 |
| 0.48 | 0.85 |
| 0.28 | 0.87 |
| 0.59 | 0.89 |
| 0.74 | 0.89 |
| 0.57 | 0.90 |
| 0.81 | 0.92 |
| 0.58 | 0.94 |
| 0.22 | 0.95 |
| 0.09 | 0.97 |
| 0.79 | 0.97 |
| 0.39 | 0.97 |
| 0.15 | 0.98 |
| 0.45 | 0.99 |

Table 4.A.2 *(continuation four)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.25 | 1.03 |
| 0.30 | 1.04 |
| 0.57 | 1.08 |
| 0.29 | 1.08 |
| 0.65 | 1.09 |
| 0.51 | 1.10 |
| 0.66 | 1.10 |
| 0.57 | 1.10 |
| 0.42 | 1.14 |
| 0.28 | 1.20 |
| 0.69 | 1.22 |
| 0.17 | 1.22 |
| 0.26 | 1.24 |
| 0.78 | 1.24 |
| 0.40 | 1.28 |
| 0.27 | 1.29 |
| 0.59 | 1.32 |
| 0.94 | 1.34 |
| 0.47 | 1.37 |
| 0.19 | 1.47 |
| 0.89 | 1.49 |
| 0.59 | 1.55 |
| 0.35 | 1.56 |
| 0.54 | 1.59 |
| 0.15 | 1.59 |
| 1.17 | 1.60 |
| 0.57 | 1.66 |
| 0.37 | 1.69 |
| 0.45 | 1.69 |
| 0.31 | 1.71 |
| 0.46 | 1.73 |
| 0.36 | 1.74 |
| 0.55 | 1.76 |
| 0.56 | 1.84 |
| 0.65 | 1.93 |
| 0.32 | 1.99 |
| 0.34 | 2.03 |
| 1.21 | 2.03 |

Table 4.A.2 *(continuation five)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.23 | 2.08 |
| 0.11 | 2.17 |
| 0.44 | 2.19 |
| 0.42 | 2.27 |
| 0.22 | 2.41 |
| 0.64 | 2.57 |
| 0.44 | 2.65 |
| 0.08 | 2.77 |
| 0.33 | 3.02 |
| 0.29 | 3.06 |
| 0.35 | 3.11 |
| 0.17 | 3.44 |

Table 4.A.3 Data for Scatter Plot of Discrimination and Difficulty Parameters—High School

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.91 | -1.33 |
| 1.01 | -1.25 |
| 0.93 | -1.21 |
| 0.37 | -0.86 |
| 0.30 | -0.86 |
| 0.54 | -0.80 |
| 0.75 | -0.77 |
| 0.24 | -0.77 |
| 0.83 | -0.72 |
| 1.03 | -0.70 |
| 0.95 | -0.69 |
| 1.14 | -0.61 |
| 0.59 | -0.61 |
| 0.56 | -0.60 |
| 1.40 | -0.60 |
| 0.51 | -0.58 |
| 0.76 | -0.58 |
| 0.84 | -0.57 |
| 0.70 | -0.57 |
| 0.52 | -0.53 |
| 1.09 | -0.50 |
| 0.66 | -0.49 |
| 0.95 | -0.46 |
| 1.19 | -0.43 |
| 0.40 | -0.42 |
| 0.68 | -0.41 |
| 0.52 | -0.41 |
| 0.62 | -0.30 |
| 0.49 | -0.30 |
| 0.51 | -0.28 |
| 0.58 | -0.28 |
| 0.57 | -0.28 |
| 0.63 | -0.24 |
| 1.30 | -0.24 |
| 0.54 | -0.23 |
| 0.46 | -0.19 |
| 0.70 | -0.18 |

Table 4.A.3 *(continuation one)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.79 | -0.18 |
| 0.86 | -0.17 |
| 0.66 | -0.17 |
| 0.44 | -0.16 |
| 0.43 | -0.13 |
| 0.63 | -0.13 |
| 0.72 | -0.12 |
| 0.46 | -0.11 |
| 0.81 | -0.10 |
| 0.93 | -0.05 |
| 0.93 | -0.05 |
| 0.73 | -0.05 |
| 0.52 | -0.04 |
| 0.44 | -0.02 |
| 0.66 | -0.01 |
| 0.69 | 0.00 |
| 0.89 | 0.00 |
| 0.67 | 0.03 |
| 0.79 | 0.04 |
| 0.40 | 0.06 |
| 0.62 | 0.07 |
| 1.36 | 0.08 |
| 0.62 | 0.09 |
| 0.67 | 0.11 |
| 0.61 | 0.12 |
| 0.47 | 0.14 |
| 0.60 | 0.15 |
| 0.31 | 0.19 |
| 0.61 | 0.19 |
| 0.19 | 0.20 |
| 0.22 | 0.21 |
| 0.31 | 0.23 |
| 0.91 | 0.23 |
| 0.40 | 0.23 |
| 0.75 | 0.24 |
| 1.25 | 0.24 |
| 0.31 | 0.25 |
| 0.65 | 0.29 |

Table 4.A.3 (*continuation two)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.39 | 0.35 |
| 0.36 | 0.36 |
| 0.46 | 0.36 |
| 0.41 | 0.39 |
| 0.48 | 0.39 |
| 0.66 | 0.39 |
| 0.54 | 0.40 |
| 0.97 | 0.41 |
| 0.52 | 0.41 |
| 0.98 | 0.42 |
| 0.71 | 0.43 |
| 0.73 | 0.44 |
| 0.69 | 0.44 |
| 0.40 | 0.44 |
| 0.34 | 0.45 |
| 0.31 | 0.47 |
| 0.21 | 0.47 |
| 0.49 | 0.50 |
| 0.58 | 0.52 |
| 0.65 | 0.55 |
| 0.32 | 0.58 |
| 0.41 | 0.58 |
| 0.92 | 0.58 |
| 0.64 | 0.59 |
| 0.34 | 0.60 |
| 0.39 | 0.61 |
| 0.59 | 0.67 |
| 0.66 | 0.74 |
| 1.14 | 0.74 |
| 0.28 | 0.76 |
| 0.64 | 0.76 |
| 0.80 | 0.79 |
| 0.39 | 0.81 |
| 0.31 | 0.82 |
| 0.27 | 0.83 |
| 0.72 | 0.85 |
| 0.48 | 0.86 |
| 0.82 | 0.91 |

Table 4.A.3 *(continuation three)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.40 | 0.92 |
| 0.37 | 0.96 |
| 0.57 | 0.97 |
| 0.42 | 0.98 |
| 0.65 | 1.00 |
| 0.49 | 1.00 |
| 0.77 | 1.00 |
| 0.66 | 1.01 |
| 0.69 | 1.02 |
| 0.35 | 1.03 |
| 0.40 | 1.04 |
| 0.84 | 1.05 |
| 0.99 | 1.06 |
| 0.65 | 1.06 |
| 0.70 | 1.07 |
| 0.39 | 1.09 |
| 0.74 | 1.10 |
| 0.75 | 1.12 |
| 0.54 | 1.14 |
| 1.02 | 1.15 |
| 0.54 | 1.15 |
| 0.39 | 1.16 |
| 0.72 | 1.17 |
| 0.36 | 1.18 |
| 0.31 | 1.19 |
| 0.75 | 1.21 |
| 0.67 | 1.22 |
| 0.26 | 1.26 |
| 0.38 | 1.27 |
| 0.46 | 1.27 |
| 0.71 | 1.27 |
| 0.99 | 1.30 |
| 0.36 | 1.32 |
| 0.49 | 1.32 |
| 0.47 | 1.35 |
| 0.49 | 1.36 |
| 0.73 | 1.36 |
| 0.54 | 1.38 |

Table 4.A.3 *(continuation four)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.38 | 1.42 |
| 0.18 | 1.43 |
| 0.45 | 1.44 |
| 0.63 | 1.46 |
| 0.60 | 1.47 |
| 0.52 | 1.48 |
| 0.22 | 1.49 |
| 0.59 | 1.50 |
| 0.76 | 1.52 |
| 0.37 | 1.53 |
| 0.54 | 1.53 |
| 0.27 | 1.55 |
| 0.22 | 1.60 |
| 0.36 | 1.72 |
| 0.17 | 1.73 |
| 0.40 | 1.74 |
| 0.68 | 1.79 |
| 0.18 | 1.82 |
| 0.57 | 1.84 |
| 0.56 | 1.85 |
| 0.21 | 1.86 |
| 0.15 | 1.93 |
| 0.14 | 1.96 |
| 1.15 | 2.05 |
| 0.38 | 2.08 |
| 0.25 | 2.18 |
| 0.97 | 2.26 |
| 0.30 | 2.35 |
| 0.32 | 2.39 |
| 0.09 | 2.42 |
| 0.43 | 2.42 |
| 0.43 | 2.47 |
| 0.88 | 2.49 |
| 0.41 | 2.56 |
| 0.32 | 2.62 |
| 0.25 | 2.63 |
| 0.45 | 2.64 |
| 0.16 | 2.69 |

Table 4.A.3 *(continuation five)*

|  |  |
| --- | --- |
| Discrimination Parameter | Difficulty Parameter |
| 0.54 | 2.93 |
| 0.15 | 3.24 |
| 0.44 | 3.25 |
| 0.48 | 3.70 |
| 0.27 | 3.87 |
| 0.16 | 3.91 |

#### Data for Information Curves

Table 4.A.4 Data for Information Curves of the Assembled MST 1-2 Design Panel—Grade Five

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| -2.5 | 1.25 | 2.38 | 1.35 |
| -2.4 | 1.39 | 2.65 | 1.49 |
| -2.3 | 1.56 | 2.95 | 1.65 |
| -2.2 | 1.75 | 3.28 | 1.83 |
| -2.1 | 1.96 | 3.64 | 2.03 |
| -2.0 | 2.20 | 4.03 | 2.26 |
| -1.9 | 2.46 | 4.44 | 2.51 |
| -1.8 | 2.74 | 4.87 | 2.79 |
| -1.7 | 3.06 | 5.32 | 3.10 |
| -1.6 | 3.40 | 5.77 | 3.44 |
| -1.5 | 3.78 | 6.22 | 3.81 |
| -1.4 | 4.18 | 6.65 | 4.21 |
| -1.3 | 4.61 | 7.06 | 4.63 |
| -1.2 | 5.06 | 7.43 | 5.09 |
| -1.1 | 5.53 | 7.77 | 5.57 |
| -1.0 | 6.01 | 8.05 | 6.06 |
| -0.9 | 6.48 | 8.28 | 6.55 |
| -0.8 | 6.94 | 8.47 | 7.04 |
| -0.7 | 7.38 | 8.59 | 7.50 |
| -0.6 | 7.79 | 8.66 | 7.93 |
| -0.5 | 8.14 | 8.67 | 8.31 |
| -0.4 | 8.43 | 8.62 | 8.61 |
| -0.3 | 8.65 | 8.50 | 8.84 |
| -0.2 | 8.78 | 8.31 | 8.98 |
| -0.1 | 8.83 | 8.06 | 9.03 |
| 0.0 | 8.80 | 7.74 | 8.98 |
| 0.1 | 8.68 | 7.38 | 8.85 |
| 0.2 | 8.48 | 6.97 | 8.63 |
| 0.3 | 8.22 | 6.54 | 8.34 |
| 0.4 | 7.90 | 6.08 | 7.98 |
| 0.5 | 7.54 | 5.63 | 7.58 |
| 0.6 | 7.14 | 5.17 | 7.13 |
| 0.7 | 6.73 | 4.73 | 6.65 |
| 0.8 | 6.31 | 4.30 | 6.16 |
| 0.9 | 5.89 | 3.90 | 5.67 |
| 1.0 | 5.47 | 3.52 | 5.19 |

Table 4.A.4 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| 1.1 | 5.06 | 3.17 | 4.72 |
| 1.2 | 4.67 | 2.85 | 4.28 |
| 1.3 | 4.30 | 2.56 | 3.86 |
| 1.4 | 3.95 | 2.29 | 3.47 |
| 1.5 | 3.62 | 2.05 | 3.12 |
| 1.6 | 3.31 | 1.83 | 2.80 |
| 1.7 | 3.02 | 1.64 | 2.51 |
| 1.8 | 2.75 | 1.46 | 2.25 |
| 1.9 | 2.50 | 1.31 | 2.02 |
| 2.0 | 2.28 | 1.17 | 1.81 |
| 2.1 | 2.07 | 1.05 | 1.62 |
| 2.2 | 1.87 | 0.94 | 1.46 |
| 2.3 | 1.70 | 0.84 | 1.31 |
| 2.4 | 1.54 | 0.75 | 1.18 |
| 2.5 | 1.39 | 0.67 | 1.06 |

Table 4.A.5 Data for Information Curves of the Assembled MST 1-3 Design Panel—Grade Five

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| -2.5 | 1.25 | 2.38 | 1.57 | 1.35 |
| -2.4 | 1.39 | 2.65 | 1.73 | 1.49 |
| -2.3 | 1.56 | 2.95 | 1.90 | 1.65 |
| -2.2 | 1.75 | 3.28 | 2.09 | 1.83 |
| -2.1 | 1.96 | 3.64 | 2.30 | 2.03 |
| -2.0 | 2.20 | 4.03 | 2.54 | 2.26 |
| -1.9 | 2.46 | 4.44 | 2.79 | 2.51 |
| -1.8 | 2.74 | 4.87 | 3.06 | 2.79 |
| -1.7 | 3.06 | 5.32 | 3.36 | 3.10 |
| -1.6 | 3.40 | 5.77 | 3.69 | 3.44 |
| -1.5 | 3.78 | 6.22 | 4.03 | 3.81 |
| -1.4 | 4.18 | 6.65 | 4.39 | 4.21 |
| -1.3 | 4.61 | 7.06 | 4.78 | 4.63 |
| -1.2 | 5.06 | 7.43 | 5.17 | 5.09 |
| -1.1 | 5.53 | 7.77 | 5.58 | 5.57 |
| -1.0 | 6.01 | 8.05 | 5.99 | 6.06 |
| -0.9 | 6.48 | 8.28 | 6.39 | 6.55 |
| -0.8 | 6.94 | 8.47 | 6.78 | 7.04 |
| -0.7 | 7.38 | 8.59 | 7.14 | 7.50 |
| -0.6 | 7.79 | 8.66 | 7.46 | 7.93 |
| -0.5 | 8.14 | 8.67 | 7.74 | 8.31 |
| -0.4 | 8.43 | 8.62 | 7.95 | 8.61 |
| -0.3 | 8.65 | 8.50 | 8.09 | 8.84 |
| -0.2 | 8.78 | 8.31 | 8.15 | 8.98 |
| -0.1 | 8.83 | 8.06 | 8.13 | 9.03 |
| 0.0 | 8.80 | 7.74 | 8.04 | 8.98 |
| 0.1 | 8.68 | 7.38 | 7.87 | 8.85 |
| 0.2 | 8.48 | 6.97 | 7.63 | 8.63 |
| 0.3 | 8.22 | 6.54 | 7.33 | 8.34 |
| 0.4 | 7.90 | 6.08 | 6.98 | 7.98 |
| 0.5 | 7.54 | 5.63 | 6.60 | 7.58 |
| 0.6 | 7.14 | 5.17 | 6.18 | 7.13 |
| 0.7 | 6.73 | 4.73 | 5.75 | 6.65 |
| 0.8 | 6.31 | 4.30 | 5.30 | 6.16 |
| 0.9 | 5.89 | 3.90 | 4.86 | 5.67 |
| 1.0 | 5.47 | 3.52 | 4.43 | 5.19 |
| 1.1 | 5.06 | 3.17 | 4.02 | 4.72 |
| 1.2 | 4.67 | 2.85 | 3.63 | 4.28 |

Table 4.A.5 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| 1.3 | 4.30 | 2.56 | 3.27 | 3.86 |
| 1.4 | 3.95 | 2.29 | 2.93 | 3.47 |
| 1.5 | 3.62 | 2.05 | 2.63 | 3.12 |
| 1.6 | 3.31 | 1.83 | 2.36 | 2.80 |
| 1.7 | 3.02 | 1.64 | 2.11 | 2.51 |
| 1.8 | 2.75 | 1.46 | 1.89 | 2.25 |
| 1.9 | 2.50 | 1.31 | 1.70 | 2.02 |
| 2.0 | 2.28 | 1.17 | 1.53 | 1.81 |
| 2.1 | 2.07 | 1.05 | 1.38 | 1.62 |
| 2.2 | 1.87 | 0.94 | 1.24 | 1.46 |
| 2.3 | 1.70 | 0.84 | 1.12 | 1.31 |
| 2.4 | 1.54 | 0.75 | 1.01 | 1.18 |
| 2.5 | 1.39 | 0.67 | 0.92 | 1.06 |

Table 4.A.6 Data for Information Curves of the Assembled MST 1-2 Design Panel—Grade Eight

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| -2.5 | 1.44 | 1.86 | 0.79 |
| -2.4 | 1.59 | 2.09 | 0.88 |
| -2.3 | 1.77 | 2.34 | 0.97 |
| -2.2 | 1.95 | 2.62 | 1.08 |
| -2.1 | 2.16 | 2.93 | 1.21 |
| -2.0 | 2.39 | 3.27 | 1.34 |
| -1.9 | 2.64 | 3.64 | 1.49 |
| -1.8 | 2.91 | 4.05 | 1.66 |
| -1.7 | 3.21 | 4.48 | 1.85 |
| -1.6 | 3.52 | 4.93 | 2.05 |
| -1.5 | 3.86 | 5.41 | 2.28 |
| -1.4 | 4.22 | 5.90 | 2.54 |
| -1.3 | 4.60 | 6.39 | 2.82 |
| -1.2 | 5.00 | 6.87 | 3.12 |
| -1.1 | 5.41 | 7.33 | 3.47 |
| -1.0 | 5.83 | 7.77 | 3.84 |
| -0.9 | 6.25 | 8.15 | 4.25 |
| -0.8 | 6.67 | 8.49 | 4.69 |
| -0.7 | 7.08 | 8.77 | 5.16 |
| -0.6 | 7.46 | 8.98 | 5.66 |
| -0.5 | 7.82 | 9.12 | 6.18 |
| -0.4 | 8.13 | 9.19 | 6.71 |
| -0.3 | 8.39 | 9.19 | 7.22 |
| -0.2 | 8.59 | 9.11 | 7.71 |
| -0.1 | 8.71 | 8.95 | 8.15 |
| 0.0 | 8.76 | 8.71 | 8.53 |
| 0.1 | 8.73 | 8.40 | 8.82 |
| 0.2 | 8.61 | 8.01 | 9.01 |
| 0.3 | 8.42 | 7.56 | 9.10 |
| 0.4 | 8.16 | 7.06 | 9.08 |
| 0.5 | 7.83 | 6.53 | 8.97 |
| 0.6 | 7.46 | 5.98 | 8.77 |
| 0.7 | 7.05 | 5.43 | 8.49 |
| 0.8 | 6.61 | 4.91 | 8.16 |
| 0.9 | 6.16 | 4.40 | 7.79 |
| 1.0 | 5.71 | 3.93 | 7.38 |
| 1.1 | 5.27 | 3.51 | 6.95 |
| 1.2 | 4.84 | 3.12 | 6.50 |

Table 4.A.6 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| 1.3 | 4.43 | 2.77 | 6.06 |
| 1.4 | 4.04 | 2.46 | 5.61 |
| 1.5 | 3.67 | 2.19 | 5.18 |
| 1.6 | 3.33 | 1.94 | 4.76 |
| 1.7 | 3.02 | 1.73 | 4.35 |
| 1.8 | 2.73 | 1.54 | 3.97 |
| 1.9 | 2.46 | 1.38 | 3.60 |
| 2.0 | 2.22 | 1.23 | 3.26 |
| 2.1 | 2.00 | 1.10 | 2.95 |
| 2.2 | 1.80 | 0.99 | 2.66 |
| 2.3 | 1.62 | 0.89 | 2.39 |
| 2.4 | 1.46 | 0.80 | 2.15 |
| 2.5 | 1.31 | 0.72 | 1.92 |

Table 4.A.7 Data for Information Curves of the Assembled MST 1-3 Design Panel—Grade Eight

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| -2.5 | 1.44 | 1.86 | 1.19 | 0.79 |
| -2.4 | 1.59 | 2.09 | 1.33 | 0.88 |
| -2.3 | 1.77 | 2.34 | 1.50 | 0.97 |
| -2.2 | 1.95 | 2.62 | 1.69 | 1.08 |
| -2.1 | 2.16 | 2.93 | 1.90 | 1.21 |
| -2.0 | 2.39 | 3.27 | 2.13 | 1.34 |
| -1.9 | 2.64 | 3.64 | 2.39 | 1.49 |
| -1.8 | 2.91 | 4.05 | 2.68 | 1.66 |
| -1.7 | 3.21 | 4.48 | 2.99 | 1.85 |
| -1.6 | 3.52 | 4.93 | 3.33 | 2.05 |
| -1.5 | 3.86 | 5.41 | 3.69 | 2.28 |
| -1.4 | 4.22 | 5.90 | 4.08 | 2.54 |
| -1.3 | 4.60 | 6.39 | 4.47 | 2.82 |
| -1.2 | 5.00 | 6.87 | 4.88 | 3.12 |
| -1.1 | 5.41 | 7.33 | 5.29 | 3.47 |
| -1.0 | 5.83 | 7.77 | 5.69 | 3.84 |
| -0.9 | 6.25 | 8.15 | 6.09 | 4.25 |
| -0.8 | 6.67 | 8.49 | 6.48 | 4.69 |
| -0.7 | 7.08 | 8.77 | 6.86 | 5.16 |
| -0.6 | 7.46 | 8.98 | 7.22 | 5.66 |
| -0.5 | 7.82 | 9.12 | 7.56 | 6.18 |
| -0.4 | 8.13 | 9.19 | 7.88 | 6.71 |
| -0.3 | 8.39 | 9.19 | 8.16 | 7.22 |
| -0.2 | 8.59 | 9.11 | 8.39 | 7.71 |
| -0.1 | 8.71 | 8.95 | 8.57 | 8.15 |
| 0.0 | 8.76 | 8.71 | 8.66 | 8.53 |
| 0.1 | 8.73 | 8.40 | 8.68 | 8.82 |
| 0.2 | 8.61 | 8.01 | 8.60 | 9.01 |
| 0.3 | 8.42 | 7.56 | 8.43 | 9.10 |
| 0.4 | 8.16 | 7.06 | 8.17 | 9.08 |
| 0.5 | 7.83 | 6.53 | 7.84 | 8.97 |
| 0.6 | 7.46 | 5.98 | 7.45 | 8.77 |
| 0.7 | 7.05 | 5.43 | 7.02 | 8.49 |
| 0.8 | 6.61 | 4.91 | 6.57 | 8.16 |
| 0.9 | 6.16 | 4.40 | 6.10 | 7.79 |
| 1.0 | 5.71 | 3.93 | 5.63 | 7.38 |
| 1.1 | 5.27 | 3.51 | 5.18 | 6.95 |
| 1.2 | 4.84 | 3.12 | 4.74 | 6.50 |

Table 4.A.7 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| 1.3 | 4.43 | 2.77 | 4.32 | 6.06 |
| 1.4 | 4.04 | 2.46 | 3.93 | 5.61 |
| 1.5 | 3.67 | 2.19 | 3.56 | 5.18 |
| 1.6 | 3.33 | 1.94 | 3.22 | 4.76 |
| 1.7 | 3.02 | 1.73 | 2.91 | 4.35 |
| 1.8 | 2.73 | 1.54 | 2.62 | 3.97 |
| 1.9 | 2.46 | 1.38 | 2.36 | 3.60 |
| 2.0 | 2.22 | 1.23 | 2.12 | 3.26 |
| 2.1 | 2.00 | 1.10 | 1.91 | 2.95 |
| 2.2 | 1.80 | 0.99 | 1.71 | 2.66 |
| 2.3 | 1.62 | 0.89 | 1.54 | 2.39 |
| 2.4 | 1.46 | 0.80 | 1.38 | 2.15 |
| 2.5 | 1.31 | 0.72 | 1.24 | 1.92 |

Table 4.A.8 Data for Information Curves of the Assembled MST 1-2 Design Panel—High School

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| -2.5 | 1.26 | 1.83 | 0.58 |
| -2.4 | 1.42 | 2.05 | 0.65 |
| -2.3 | 1.59 | 2.29 | 0.72 |
| -2.2 | 1.78 | 2.55 | 0.81 |
| -2.1 | 2.00 | 2.83 | 0.90 |
| -2.0 | 2.25 | 3.13 | 1.00 |
| -1.9 | 2.53 | 3.45 | 1.12 |
| -1.8 | 2.84 | 3.79 | 1.24 |
| -1.7 | 3.18 | 4.14 | 1.38 |
| -1.6 | 3.57 | 4.50 | 1.54 |
| -1.5 | 3.99 | 4.87 | 1.71 |
| -1.4 | 4.46 | 5.24 | 1.89 |
| -1.3 | 4.96 | 5.60 | 2.10 |
| -1.2 | 5.50 | 5.95 | 2.32 |
| -1.1 | 6.06 | 6.28 | 2.56 |
| -1.0 | 6.63 | 6.57 | 2.82 |
| -0.9 | 7.20 | 6.84 | 3.09 |
| -0.8 | 7.75 | 7.05 | 3.38 |
| -0.7 | 8.25 | 7.22 | 3.69 |
| -0.6 | 8.69 | 7.34 | 4.01 |
| -0.5 | 9.05 | 7.40 | 4.35 |
| -0.4 | 9.33 | 7.40 | 4.69 |
| -0.3 | 9.51 | 7.36 | 5.04 |
| -0.2 | 9.59 | 7.27 | 5.38 |
| -0.1 | 9.58 | 7.13 | 5.72 |
| 0.0 | 9.48 | 6.96 | 6.06 |
| 0.1 | 9.29 | 6.76 | 6.37 |
| 0.2 | 9.04 | 6.54 | 6.66 |
| 0.3 | 8.72 | 6.29 | 6.92 |
| 0.4 | 8.35 | 6.04 | 7.14 |
| 0.5 | 7.94 | 5.78 | 7.32 |
| 0.6 | 7.50 | 5.52 | 7.44 |
| 0.7 | 7.05 | 5.26 | 7.51 |
| 0.8 | 6.59 | 5.00 | 7.52 |
| 0.9 | 6.13 | 4.74 | 7.48 |
| 1.0 | 5.68 | 4.49 | 7.37 |
| 1.1 | 5.25 | 4.24 | 7.22 |
| 1.2 | 4.85 | 4.00 | 7.01 |

Table 4.A.8 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | Router | Easy | Hard |
| 1.3 | 4.46 | 3.76 | 6.77 |
| 1.4 | 4.11 | 3.53 | 6.49 |
| 1.5 | 3.77 | 3.31 | 6.19 |
| 1.6 | 3.47 | 3.09 | 5.86 |
| 1.7 | 3.19 | 2.88 | 5.53 |
| 1.8 | 2.93 | 2.68 | 5.19 |
| 1.9 | 2.70 | 2.49 | 4.85 |
| 2.0 | 2.49 | 2.30 | 4.51 |
| 2.1 | 2.29 | 2.13 | 4.18 |
| 2.2 | 2.11 | 1.96 | 3.86 |
| 2.3 | 1.95 | 1.80 | 3.56 |
| 2.4 | 1.80 | 1.65 | 3.27 |
| 2.5 | 1.66 | 1.51 | 3.00 |

Table 4.A.9 Data for Information Curves of the Assembled MST 1-3 Design Panel—High School

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| -2.5 | 1.26 | 1.83 | 0.92 | 0.58 |
| -2.4 | 1.42 | 2.05 | 1.04 | 0.65 |
| -2.3 | 1.59 | 2.29 | 1.17 | 0.72 |
| -2.2 | 1.78 | 2.55 | 1.32 | 0.81 |
| -2.1 | 2.00 | 2.83 | 1.48 | 0.90 |
| -2.0 | 2.25 | 3.13 | 1.66 | 1.00 |
| -1.9 | 2.53 | 3.45 | 1.86 | 1.12 |
| -1.8 | 2.84 | 3.79 | 2.08 | 1.24 |
| -1.7 | 3.18 | 4.14 | 2.32 | 1.38 |
| -1.6 | 3.57 | 4.50 | 2.57 | 1.54 |
| -1.5 | 3.99 | 4.87 | 2.85 | 1.71 |
| -1.4 | 4.46 | 5.24 | 3.14 | 1.89 |
| -1.3 | 4.96 | 5.60 | 3.44 | 2.10 |
| -1.2 | 5.50 | 5.95 | 3.76 | 2.32 |
| -1.1 | 6.06 | 6.28 | 4.08 | 2.56 |
| -1.0 | 6.63 | 6.57 | 4.41 | 2.82 |
| -0.9 | 7.20 | 6.84 | 4.73 | 3.09 |
| -0.8 | 7.75 | 7.05 | 5.05 | 3.38 |
| -0.7 | 8.25 | 7.22 | 5.36 | 3.69 |
| -0.6 | 8.69 | 7.34 | 5.66 | 4.01 |
| -0.5 | 9.05 | 7.40 | 5.94 | 4.35 |
| -0.4 | 9.33 | 7.40 | 6.21 | 4.69 |
| -0.3 | 9.51 | 7.36 | 6.44 | 5.04 |
| -0.2 | 9.59 | 7.27 | 6.66 | 5.38 |
| -0.1 | 9.58 | 7.13 | 6.85 | 5.72 |
| 0.0 | 9.48 | 6.96 | 7.00 | 6.06 |
| 0.1 | 9.29 | 6.76 | 7.13 | 6.37 |
| 0.2 | 9.04 | 6.54 | 7.23 | 6.66 |
| 0.3 | 8.72 | 6.29 | 7.28 | 6.92 |
| 0.4 | 8.35 | 6.04 | 7.30 | 7.14 |
| 0.5 | 7.94 | 5.78 | 7.28 | 7.32 |
| 0.6 | 7.50 | 5.52 | 7.22 | 7.44 |
| 0.7 | 7.05 | 5.26 | 7.12 | 7.51 |
| 0.8 | 6.59 | 5.00 | 6.98 | 7.52 |
| 0.9 | 6.13 | 4.74 | 6.79 | 7.48 |
| 1.0 | 5.68 | 4.49 | 6.58 | 7.37 |
| 1.1 | 5.25 | 4.24 | 6.33 | 7.22 |
| 1.2 | 4.85 | 4.00 | 6.05 | 7.01 |

Table 4.A.9 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | Router | Easy | Medium | Hard |
| 1.3 | 4.46 | 3.76 | 5.76 | 6.77 |
| 1.4 | 4.11 | 3.53 | 5.45 | 6.49 |
| 1.5 | 3.77 | 3.31 | 5.14 | 6.19 |
| 1.6 | 3.47 | 3.09 | 4.82 | 5.86 |
| 1.7 | 3.19 | 2.88 | 4.50 | 5.53 |
| 1.8 | 2.93 | 2.68 | 4.19 | 5.19 |
| 1.9 | 2.70 | 2.49 | 3.88 | 4.85 |
| 2.0 | 2.49 | 2.30 | 3.59 | 4.51 |
| 2.1 | 2.29 | 2.13 | 3.31 | 4.18 |
| 2.2 | 2.11 | 1.96 | 3.04 | 3.86 |
| 2.3 | 1.95 | 1.80 | 2.78 | 3.56 |
| 2.4 | 1.80 | 1.65 | 2.54 | 3.27 |
| 2.5 | 1.66 | 1.51 | 2.32 | 3.00 |

#### Data for Conditional Standard Errors of Measurement (CSEM) Graphs

Table 4.A.10 Data for CSEMs—Grade Five

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | 0.64 | 0.64 | 0.61 |
| -2.4 | 0.62 | 0.62 | 0.61 |
| -2.3 | 0.60 | 0.60 | 0.60 |
| -2.2 | 0.57 | 0.57 | 0.59 |
| -2.1 | 0.54 | 0.54 | 0.57 |
| -2.0 | 0.51 | 0.51 | 0.55 |
| -1.9 | 0.48 | 0.48 | 0.53 |
| -1.8 | 0.45 | 0.45 | 0.51 |
| -1.7 | 0.42 | 0.42 | 0.48 |
| -1.6 | 0.39 | 0.39 | 0.46 |
| -1.5 | 0.37 | 0.37 | 0.43 |
| -1.4 | 0.35 | 0.35 | 0.41 |
| -1.3 | 0.33 | 0.33 | 0.38 |
| -1.2 | 0.32 | 0.32 | 0.36 |
| -1.1 | 0.31 | 0.31 | 0.34 |
| -1.0 | 0.30 | 0.30 | 0.33 |
| -0.9 | 0.29 | 0.29 | 0.31 |
| -0.8 | 0.28 | 0.28 | 0.30 |
| -0.7 | 0.27 | 0.27 | 0.29 |
| -0.6 | 0.27 | 0.27 | 0.28 |
| -0.5 | 0.26 | 0.26 | 0.28 |
| -0.4 | 0.26 | 0.26 | 0.27 |
| -0.3 | 0.26 | 0.26 | 0.27 |
| -0.2 | 0.26 | 0.26 | 0.27 |
| -0.1 | 0.26 | 0.26 | 0.27 |
| 0.0 | 0.26 | 0.26 | 0.27 |
| 0.1 | 0.26 | 0.26 | 0.27 |
| 0.2 | 0.26 | 0.26 | 0.28 |
| 0.3 | 0.27 | 0.27 | 0.28 |
| 0.4 | 0.28 | 0.28 | 0.29 |
| 0.5 | 0.28 | 0.28 | 0.30 |
| 0.6 | 0.29 | 0.29 | 0.31 |
| 0.7 | 0.31 | 0.31 | 0.32 |
| 0.8 | 0.32 | 0.32 | 0.34 |
| 0.9 | 0.33 | 0.33 | 0.36 |
| 1.0 | 0.35 | 0.35 | 0.37 |

Table 4.A.10 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.1 | 0.37 | 0.37 | 0.40 |
| 1.2 | 0.40 | 0.40 | 0.42 |
| 1.3 | 0.42 | 0.42 | 0.45 |
| 1.4 | 0.45 | 0.45 | 0.48 |
| 1.5 | 0.48 | 0.48 | 0.51 |
| 1.6 | 0.51 | 0.51 | 0.55 |
| 1.7 | 0.55 | 0.55 | 0.58 |
| 1.8 | 0.59 | 0.59 | 0.62 |
| 1.9 | 0.63 | 0.63 | 0.66 |
| 2.0 | 0.66 | 0.66 | 0.70 |
| 2.1 | 0.70 | 0.70 | 0.73 |
| 2.2 | 0.73 | 0.73 | 0.77 |
| 2.3 | 0.76 | 0.76 | 0.80 |
| 2.4 | 0.79 | 0.79 | 0.83 |
| 2.5 | 0.82 | 0.82 | 0.86 |

Table 4.A.11 Data for CSEMs—Grade Eight

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | 0.63 | 0.63 | 0.69 |
| -2.4 | 0.62 | 0.62 | 0.68 |
| -2.3 | 0.60 | 0.60 | 0.67 |
| -2.2 | 0.58 | 0.58 | 0.65 |
| -2.1 | 0.55 | 0.55 | 0.62 |
| -2.0 | 0.52 | 0.52 | 0.60 |
| -1.9 | 0.50 | 0.50 | 0.57 |
| -1.8 | 0.47 | 0.47 | 0.54 |
| -1.7 | 0.44 | 0.44 | 0.51 |
| -1.6 | 0.41 | 0.41 | 0.47 |
| -1.5 | 0.39 | 0.39 | 0.44 |
| -1.4 | 0.36 | 0.36 | 0.42 |
| -1.3 | 0.34 | 0.34 | 0.39 |
| -1.2 | 0.33 | 0.33 | 0.37 |
| -1.1 | 0.31 | 0.31 | 0.35 |
| -1.0 | 0.30 | 0.30 | 0.33 |
| -0.9 | 0.29 | 0.29 | 0.32 |
| -0.8 | 0.28 | 0.28 | 0.30 |
| -0.7 | 0.27 | 0.27 | 0.29 |
| -0.6 | 0.27 | 0.27 | 0.28 |
| -0.5 | 0.26 | 0.26 | 0.28 |
| -0.4 | 0.26 | 0.26 | 0.27 |
| -0.3 | 0.25 | 0.25 | 0.26 |
| -0.2 | 0.25 | 0.25 | 0.26 |
| -0.1 | 0.25 | 0.25 | 0.26 |
| 0.0 | 0.25 | 0.25 | 0.26 |
| 0.1 | 0.25 | 0.25 | 0.26 |
| 0.2 | 0.25 | 0.25 | 0.26 |
| 0.3 | 0.25 | 0.25 | 0.26 |
| 0.4 | 0.26 | 0.26 | 0.27 |
| 0.5 | 0.26 | 0.26 | 0.27 |
| 0.6 | 0.27 | 0.27 | 0.28 |
| 0.7 | 0.27 | 0.27 | 0.29 |
| 0.8 | 0.28 | 0.28 | 0.30 |
| 0.9 | 0.29 | 0.29 | 0.32 |
| 1.0 | 0.30 | 0.30 | 0.33 |
| 1.1 | 0.31 | 0.31 | 0.35 |
| 1.2 | 0.33 | 0.33 | 0.37 |
| 1.3 | 0.34 | 0.34 | 0.40 |

Table 4.A.11 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.4 | 0.36 | 0.36 | 0.42 |
| 1.5 | 0.39 | 0.39 | 0.45 |
| 1.6 | 0.41 | 0.41 | 0.48 |
| 1.7 | 0.44 | 0.44 | 0.52 |
| 1.8 | 0.47 | 0.47 | 0.55 |
| 1.9 | 0.50 | 0.50 | 0.59 |
| 2.0 | 0.53 | 0.53 | 0.62 |
| 2.1 | 0.56 | 0.56 | 0.65 |
| 2.2 | 0.59 | 0.59 | 0.68 |
| 2.3 | 0.61 | 0.61 | 0.70 |
| 2.4 | 0.63 | 0.63 | 0.72 |
| 2.5 | 0.65 | 0.65 | 0.74 |

Table 4.A.12 Data for CSEMs—High School

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | 0.64 | 0.64 | 0.73 |
| -2.4 | 0.63 | 0.63 | 0.73 |
| -2.3 | 0.61 | 0.61 | 0.71 |
| -2.2 | 0.59 | 0.59 | 0.70 |
| -2.1 | 0.57 | 0.57 | 0.67 |
| -2.0 | 0.54 | 0.54 | 0.65 |
| -1.9 | 0.51 | 0.51 | 0.62 |
| -1.8 | 0.48 | 0.48 | 0.59 |
| -1.7 | 0.45 | 0.45 | 0.55 |
| -1.6 | 0.42 | 0.42 | 0.52 |
| -1.5 | 0.39 | 0.39 | 0.48 |
| -1.4 | 0.37 | 0.37 | 0.45 |
| -1.3 | 0.35 | 0.35 | 0.42 |
| -1.2 | 0.33 | 0.33 | 0.39 |
| -1.1 | 0.32 | 0.32 | 0.37 |
| -1.0 | 0.30 | 0.30 | 0.35 |
| -0.9 | 0.29 | 0.29 | 0.33 |
| -0.8 | 0.28 | 0.28 | 0.31 |
| -0.7 | 0.28 | 0.28 | 0.30 |
| -0.6 | 0.27 | 0.27 | 0.29 |
| -0.5 | 0.27 | 0.27 | 0.28 |
| -0.4 | 0.26 | 0.26 | 0.27 |
| -0.3 | 0.26 | 0.26 | 0.27 |
| -0.2 | 0.26 | 0.26 | 0.27 |
| -0.1 | 0.26 | 0.26 | 0.26 |
| 0.0 | 0.26 | 0.26 | 0.26 |
| 0.1 | 0.26 | 0.26 | 0.26 |
| 0.2 | 0.26 | 0.26 | 0.26 |
| 0.3 | 0.26 | 0.26 | 0.27 |
| 0.4 | 0.27 | 0.27 | 0.27 |
| 0.5 | 0.27 | 0.27 | 0.27 |
| 0.6 | 0.28 | 0.27 | 0.28 |
| 0.7 | 0.28 | 0.28 | 0.28 |
| 0.8 | 0.29 | 0.28 | 0.29 |
| 0.9 | 0.29 | 0.29 | 0.30 |
| 1.0 | 0.30 | 0.30 | 0.31 |
| 1.1 | 0.30 | 0.30 | 0.32 |
| 1.2 | 0.31 | 0.31 | 0.33 |
| 1.3 | 0.32 | 0.32 | 0.34 |

Table 4.A.12 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.4 | 0.33 | 0.33 | 0.36 |
| 1.5 | 0.35 | 0.35 | 0.37 |
| 1.6 | 0.36 | 0.36 | 0.39 |
| 1.7 | 0.37 | 0.37 | 0.41 |
| 1.8 | 0.39 | 0.39 | 0.43 |
| 1.9 | 0.41 | 0.41 | 0.45 |
| 2.0 | 0.43 | 0.43 | 0.47 |
| 2.1 | 0.45 | 0.45 | 0.50 |
| 2.2 | 0.48 | 0.48 | 0.53 |
| 2.3 | 0.50 | 0.50 | 0.55 |
| 2.4 | 0.53 | 0.53 | 0.58 |
| 2.5 | 0.56 | 0.56 | 0.61 |

#### Data for Conditional Bias Graphs

Table 4.A.13 Data for Conditional Bias—Grade Five

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | -0.16 | -0.16 | -0.12 |
| -2.4 | -0.15 | -0.15 | -0.13 |
| -2.3 | -0.14 | -0.14 | -0.13 |
| -2.2 | -0.13 | -0.13 | -0.13 |
| -2.1 | -0.11 | -0.11 | -0.12 |
| -2.0 | -0.10 | -0.10 | -0.12 |
| -1.9 | -0.09 | -0.09 | -0.11 |
| -1.8 | -0.08 | -0.08 | -0.10 |
| -1.7 | -0.07 | -0.07 | -0.09 |
| -1.6 | -0.06 | -0.06 | -0.08 |
| -1.5 | -0.05 | -0.05 | -0.07 |
| -1.4 | -0.04 | -0.04 | -0.06 |
| -1.3 | -0.04 | -0.04 | -0.05 |
| -1.2 | -0.03 | -0.03 | -0.05 |
| -1.1 | -0.03 | -0.03 | -0.04 |
| -1.0 | -0.02 | -0.02 | -0.03 |
| -0.9 | -0.02 | -0.02 | -0.03 |
| -0.8 | -0.01 | -0.01 | -0.02 |
| -0.7 | -0.01 | -0.01 | -0.02 |
| -0.6 | -0.01 | -0.01 | -0.02 |
| -0.5 | -0.01 | -0.01 | -0.01 |
| -0.4 | -0.01 | -0.01 | -0.01 |
| -0.3 | -0.01 | -0.01 | -0.01 |
| -0.2 | 0.00 | 0.00 | 0.00 |
| -0.1 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.00 | 0.00 | 0.00 |
| 0.1 | 0.00 | 0.00 | 0.01 |
| 0.2 | 0.01 | 0.01 | 0.01 |
| 0.3 | 0.01 | 0.01 | 0.01 |
| 0.4 | 0.01 | 0.01 | 0.02 |
| 0.5 | 0.02 | 0.02 | 0.02 |
| 0.6 | 0.02 | 0.02 | 0.03 |
| 0.7 | 0.03 | 0.03 | 0.03 |
| 0.8 | 0.03 | 0.03 | 0.04 |
| 0.9 | 0.04 | 0.04 | 0.04 |
| 1.0 | 0.04 | 0.04 | 0.05 |
| 1.1 | 0.05 | 0.05 | 0.05 |

Table 4.A.13 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.2 | 0.06 | 0.06 | 0.06 |
| 1.3 | 0.07 | 0.07 | 0.07 |
| 1.4 | 0.07 | 0.07 | 0.08 |
| 1.5 | 0.09 | 0.09 | 0.09 |
| 1.6 | 0.10 | 0.10 | 0.10 |
| 1.7 | 0.11 | 0.11 | 0.12 |
| 1.8 | 0.12 | 0.12 | 0.13 |
| 1.9 | 0.14 | 0.14 | 0.15 |
| 2.0 | 0.16 | 0.16 | 0.16 |
| 2.1 | 0.17 | 0.17 | 0.18 |
| 2.2 | 0.19 | 0.19 | 0.20 |
| 2.3 | 0.20 | 0.20 | 0.21 |
| 2.4 | 0.22 | 0.22 | 0.23 |
| 2.5 | 0.23 | 0.23 | 0.24 |

Table 4.A.14 Data for Conditional Bias—Grade Eight

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | -0.16 | -0.16 | -0.18 |
| -2.4 | -0.16 | -0.16 | -0.18 |
| -2.3 | -0.15 | -0.15 | -0.17 |
| -2.2 | -0.13 | -0.13 | -0.16 |
| -2.1 | -0.12 | -0.12 | -0.15 |
| -2.0 | -0.11 | -0.11 | -0.14 |
| -1.9 | -0.10 | -0.10 | -0.13 |
| -1.8 | -0.09 | -0.09 | -0.11 |
| -1.7 | -0.07 | -0.07 | -0.10 |
| -1.6 | -0.06 | -0.06 | -0.09 |
| -1.5 | -0.06 | -0.06 | -0.08 |
| -1.4 | -0.05 | -0.05 | -0.07 |
| -1.3 | -0.04 | -0.04 | -0.06 |
| -1.2 | -0.03 | -0.03 | -0.05 |
| -1.1 | -0.03 | -0.03 | -0.04 |
| -1.0 | -0.02 | -0.02 | -0.04 |
| -0.9 | -0.02 | -0.02 | -0.03 |
| -0.8 | -0.02 | -0.02 | -0.03 |
| -0.7 | -0.01 | -0.01 | -0.02 |
| -0.6 | -0.01 | -0.01 | -0.02 |
| -0.5 | -0.01 | -0.01 | -0.02 |
| -0.4 | 0.00 | 0.00 | -0.01 |
| -0.3 | 0.00 | 0.00 | -0.01 |
| -0.2 | 0.00 | 0.00 | -0.01 |
| -0.1 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.00 | 0.00 | 0.00 |
| 0.1 | 0.00 | 0.00 | 0.00 |
| 0.2 | 0.00 | 0.00 | 0.01 |
| 0.3 | 0.00 | 0.00 | 0.01 |
| 0.4 | 0.00 | 0.00 | 0.01 |
| 0.5 | 0.01 | 0.01 | 0.02 |
| 0.6 | 0.01 | 0.01 | 0.02 |
| 0.7 | 0.01 | 0.01 | 0.02 |
| 0.8 | 0.02 | 0.02 | 0.03 |
| 0.9 | 0.02 | 0.02 | 0.03 |
| 1.0 | 0.03 | 0.03 | 0.04 |
| 1.1 | 0.03 | 0.03 | 0.05 |
| 1.2 | 0.04 | 0.04 | 0.05 |
| 1.3 | 0.04 | 0.04 | 0.06 |

Table 4.A.14 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.4 | 0.05 | 0.05 | 0.07 |
| 1.5 | 0.06 | 0.06 | 0.08 |
| 1.6 | 0.07 | 0.07 | 0.09 |
| 1.7 | 0.08 | 0.08 | 0.10 |
| 1.8 | 0.09 | 0.09 | 0.12 |
| 1.9 | 0.10 | 0.10 | 0.13 |
| 2.0 | 0.11 | 0.11 | 0.15 |
| 2.1 | 0.13 | 0.13 | 0.16 |
| 2.2 | 0.14 | 0.14 | 0.17 |
| 2.3 | 0.15 | 0.15 | 0.19 |
| 2.4 | 0.16 | 0.16 | 0.20 |
| 2.5 | 0.17 | 0.17 | 0.21 |

Table 4.A.15 Data for Conditional Bias—High School

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| -2.5 | -0.17 | -0.17 | -0.22 |
| -2.4 | -0.17 | -0.17 | -0.22 |
| -2.3 | -0.16 | -0.16 | -0.21 |
| -2.2 | -0.15 | -0.15 | -0.20 |
| -2.1 | -0.13 | -0.13 | -0.19 |
| -2.0 | -0.12 | -0.12 | -0.17 |
| -1.9 | -0.11 | -0.11 | -0.16 |
| -1.8 | -0.09 | -0.09 | -0.14 |
| -1.7 | -0.08 | -0.08 | -0.13 |
| -1.6 | -0.07 | -0.07 | -0.11 |
| -1.5 | -0.06 | -0.06 | -0.10 |
| -1.4 | -0.05 | -0.05 | -0.08 |
| -1.3 | -0.04 | -0.04 | -0.07 |
| -1.2 | -0.04 | -0.04 | -0.06 |
| -1.1 | -0.03 | -0.03 | -0.05 |
| -1.0 | -0.03 | -0.03 | -0.04 |
| -0.9 | -0.02 | -0.02 | -0.04 |
| -0.8 | -0.02 | -0.02 | -0.03 |
| -0.7 | -0.01 | -0.01 | -0.02 |
| -0.6 | -0.01 | -0.01 | -0.02 |
| -0.5 | -0.01 | -0.01 | -0.02 |
| -0.4 | 0.00 | 0.00 | -0.01 |
| -0.3 | 0.00 | 0.00 | -0.01 |
| -0.2 | 0.00 | 0.00 | -0.01 |
| -0.1 | 0.00 | 0.00 | 0.00 |
| 0.0 | 0.00 | 0.00 | 0.00 |
| 0.1 | 0.00 | 0.00 | 0.00 |
| 0.2 | 0.00 | 0.00 | 0.00 |
| 0.3 | 0.00 | 0.00 | 0.00 |
| 0.4 | 0.00 | 0.00 | 0.01 |
| 0.5 | 0.00 | 0.00 | 0.01 |
| 0.6 | 0.00 | 0.00 | 0.01 |
| 0.7 | 0.01 | 0.01 | 0.01 |
| 0.8 | 0.01 | 0.01 | 0.02 |
| 0.9 | 0.01 | 0.01 | 0.02 |
| 1.0 | 0.02 | 0.02 | 0.02 |
| 1.1 | 0.02 | 0.02 | 0.02 |
| 1.2 | 0.02 | 0.02 | 0.03 |
| 1.3 | 0.02 | 0.02 | 0.03 |

Table 4.A.15 *(continuation)*

|  |  |  |  |
| --- | --- | --- | --- |
| Proficiency | MST 1-2 | MST 1-3 | Linear |
| 1.4 | 0.03 | 0.03 | 0.04 |
| 1.5 | 0.03 | 0.03 | 0.04 |
| 1.6 | 0.04 | 0.04 | 0.05 |
| 1.7 | 0.04 | 0.04 | 0.05 |
| 1.8 | 0.05 | 0.05 | 0.06 |
| 1.9 | 0.05 | 0.05 | 0.07 |
| 2.0 | 0.06 | 0.06 | 0.08 |
| 2.1 | 0.07 | 0.07 | 0.09 |
| 2.2 | 0.08 | 0.08 | 0.10 |
| 2.3 | 0.09 | 0.09 | 0.11 |
| 2.4 | 0.10 | 0.10 | 0.12 |
| 2.5 | 0.11 | 0.11 | 0.13 |

#### Data for Relative Efficiency Graphs

Table 4.A.16 Data for Relative Efficiency—Grade Five

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1-2 | Lower Line | MST 1-3 | Higher Line |
| -2.5 | 0.78 | 0.7 | 0.78 | 1 |
| -2.4 | 0.77 | 0.7 | 0.77 | 1 |
| -2.3 | 0.77 | 0.7 | 0.77 | 1 |
| -2.2 | 0.76 | 0.7 | 0.76 | 1 |
| -2.1 | 0.76 | 0.7 | 0.76 | 1 |
| -2.0 | 0.76 | 0.7 | 0.76 | 1 |
| -1.9 | 0.76 | 0.7 | 0.76 | 1 |
| -1.8 | 0.76 | 0.7 | 0.76 | 1 |
| -1.7 | 0.77 | 0.7 | 0.77 | 1 |
| -1.6 | 0.77 | 0.7 | 0.77 | 1 |
| -1.5 | 0.78 | 0.7 | 0.78 | 1 |
| -1.4 | 0.79 | 0.7 | 0.79 | 1 |
| -1.3 | 0.81 | 0.7 | 0.81 | 1 |
| -1.2 | 0.82 | 0.7 | 0.82 | 1 |
| -1.1 | 0.84 | 0.7 | 0.84 | 1 |
| -1.0 | 0.86 | 0.7 | 0.86 | 1 |
| -0.9 | 0.88 | 0.7 | 0.88 | 1 |
| -0.8 | 0.90 | 0.7 | 0.90 | 1 |
| -0.7 | 0.92 | 0.7 | 0.92 | 1 |
| -0.6 | 0.94 | 0.7 | 0.94 | 1 |
| -0.5 | 0.95 | 0.7 | 0.95 | 1 |
| -0.4 | 0.96 | 0.7 | 0.96 | 1 |
| -0.3 | 0.97 | 0.7 | 0.97 | 1 |
| -0.2 | 0.97 | 0.7 | 0.97 | 1 |
| -0.1 | 0.96 | 0.7 | 0.96 | 1 |
| 0.0 | 0.96 | 0.7 | 0.96 | 1 |
| 0.1 | 0.95 | 0.7 | 0.95 | 1 |
| 0.2 | 0.95 | 0.7 | 0.95 | 1 |
| 0.3 | 0.94 | 0.7 | 0.94 | 1 |
| 0.4 | 0.94 | 0.7 | 0.94 | 1 |
| 0.5 | 0.94 | 0.7 | 0.94 | 1 |
| 0.6 | 0.93 | 0.7 | 0.93 | 1 |
| 0.7 | 0.93 | 0.7 | 0.93 | 1 |
| 0.8 | 0.93 | 0.7 | 0.93 | 1 |
| 0.9 | 0.93 | 0.7 | 0.93 | 1 |
| 1.0 | 0.93 | 0.7 | 0.93 | 1 |
| 1.1 | 0.93 | 0.7 | 0.93 | 1 |

Table 4.A.16 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1-2 | Lower Line | MST 1-3 | Higher Line |
| 1.2 | 0.93 | 0.7 | 0.93 | 1 |
| 1.3 | 0.93 | 0.7 | 0.93 | 1 |
| 1.4 | 0.93 | 0.7 | 0.93 | 1 |
| 1.5 | 0.93 | 0.7 | 0.93 | 1 |
| 1.6 | 0.93 | 0.7 | 0.93 | 1 |
| 1.7 | 0.93 | 0.7 | 0.93 | 1 |
| 1.8 | 0.93 | 0.7 | 0.93 | 1 |
| 1.9 | 0.93 | 0.7 | 0.93 | 1 |
| 2.0 | 0.93 | 0.7 | 0.93 | 1 |
| 2.1 | 0.93 | 0.7 | 0.93 | 1 |
| 2.2 | 0.94 | 0.7 | 0.94 | 1 |
| 2.3 | 0.94 | 0.7 | 0.94 | 1 |
| 2.4 | 0.94 | 0.7 | 0.94 | 1 |
| 2.5 | 0.94 | 0.7 | 0.94 | 1 |

Table 4.A.17 Data for Relative Efficiency—Grade Eight

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1–2 | Lower Line | MST 1–3 | Higher Line |
| -2.5 | 0.80 | 0.8 | 0.80 | 1 |
| -2.4 | 0.80 | 0.8 | 0.80 | 1 |
| -2.3 | 0.80 | 0.8 | 0.80 | 1 |
| -2.2 | 0.80 | 0.8 | 0.80 | 1 |
| -2.1 | 0.80 | 0.8 | 0.80 | 1 |
| -2.0 | 0.80 | 0.8 | 0.80 | 1 |
| -1.9 | 0.80 | 0.8 | 0.80 | 1 |
| -1.8 | 0.80 | 0.8 | 0.80 | 1 |
| -1.7 | 0.81 | 0.8 | 0.81 | 1 |
| -1.6 | 0.81 | 0.8 | 0.81 | 1 |
| -1.5 | 0.81 | 0.8 | 0.81 | 1 |
| -1.4 | 0.82 | 0.8 | 0.82 | 1 |
| -1.3 | 0.83 | 0.8 | 0.83 | 1 |
| -1.2 | 0.83 | 0.8 | 0.83 | 1 |
| -1.1 | 0.84 | 0.8 | 0.84 | 1 |
| -1.0 | 0.85 | 0.8 | 0.85 | 1 |
| -0.9 | 0.86 | 0.8 | 0.86 | 1 |
| -0.8 | 0.87 | 0.8 | 0.87 | 1 |
| -0.7 | 0.88 | 0.8 | 0.88 | 1 |
| -0.6 | 0.90 | 0.8 | 0.90 | 1 |
| -0.5 | 0.92 | 0.8 | 0.92 | 1 |
| -0.4 | 0.94 | 0.8 | 0.94 | 1 |
| -0.3 | 0.96 | 0.8 | 0.96 | 1 |
| -0.2 | 0.98 | 0.8 | 0.98 | 1 |
| -0.1 | 0.99 | 0.8 | 0.99 | 1 |
| 0.0 | 1.00 | 0.8 | 1.00 | 1 |
| 0.1 | 1.00 | 0.8 | 1.00 | 1 |
| 0.2 | 0.99 | 0.8 | 0.99 | 1 |
| 0.3 | 0.98 | 0.8 | 0.98 | 1 |
| 0.4 | 0.96 | 0.8 | 0.96 | 1 |
| 0.5 | 0.95 | 0.8 | 0.95 | 1 |
| 0.6 | 0.93 | 0.8 | 0.93 | 1 |
| 0.7 | 0.91 | 0.8 | 0.91 | 1 |
| 0.8 | 0.89 | 0.8 | 0.89 | 1 |
| 0.9 | 0.88 | 0.8 | 0.88 | 1 |
| 1.0 | 0.87 | 0.8 | 0.87 | 1 |
| 1.1 | 0.86 | 0.8 | 0.86 | 1 |
| 1.2 | 0.84 | 0.8 | 0.84 | 1 |
| 1.3 | 0.83 | 0.8 | 0.83 | 1 |

Table 4.A.17 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1–2 | Lower Line | MST 1–3 | Higher Line |
| 1.4 | 0.83 | 0.8 | 0.83 | 1 |
| 1.5 | 0.82 | 0.8 | 0.82 | 1 |
| 1.6 | 0.81 | 0.8 | 0.81 | 1 |
| 1.7 | 0.80 | 0.8 | 0.80 | 1 |
| 1.8 | 0.80 | 0.8 | 0.80 | 1 |
| 1.9 | 0.79 | 0.8 | 0.79 | 1 |
| 2.0 | 0.79 | 0.8 | 0.79 | 1 |
| 2.1 | 0.79 | 0.8 | 0.79 | 1 |
| 2.2 | 0.79 | 0.8 | 0.79 | 1 |
| 2.3 | 0.79 | 0.8 | 0.79 | 1 |
| 2.4 | 0.79 | 0.8 | 0.79 | 1 |
| 2.5 | 0.79 | 0.8 | 0.79 | 1 |

Table 4.A.18 Data for Relative Efficiency—High School

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1–2 | Lower Line | MST 1–3 | Higher Line |
| -2.5 | 0.71 | 0.7 | 0.71 | 1 |
| -2.4 | 0.71 | 0.7 | 0.71 | 1 |
| -2.3 | 0.71 | 0.7 | 0.71 | 1 |
| -2.2 | 0.72 | 0.7 | 0.72 | 1 |
| -2.1 | 0.72 | 0.7 | 0.72 | 1 |
| -2.0 | 0.73 | 0.7 | 0.73 | 1 |
| -1.9 | 0.73 | 0.7 | 0.73 | 1 |
| -1.8 | 0.74 | 0.7 | 0.74 | 1 |
| -1.7 | 0.75 | 0.7 | 0.75 | 1 |
| -1.6 | 0.76 | 0.7 | 0.76 | 1 |
| -1.5 | 0.77 | 0.7 | 0.77 | 1 |
| -1.4 | 0.78 | 0.7 | 0.78 | 1 |
| -1.3 | 0.80 | 0.7 | 0.80 | 1 |
| -1.2 | 0.81 | 0.7 | 0.81 | 1 |
| -1.1 | 0.82 | 0.7 | 0.82 | 1 |
| -1.0 | 0.84 | 0.7 | 0.84 | 1 |
| -0.9 | 0.85 | 0.7 | 0.85 | 1 |
| -0.8 | 0.87 | 0.7 | 0.87 | 1 |
| -0.7 | 0.88 | 0.7 | 0.88 | 1 |
| -0.6 | 0.90 | 0.7 | 0.90 | 1 |
| -0.5 | 0.92 | 0.7 | 0.92 | 1 |
| -0.4 | 0.94 | 0.7 | 0.94 | 1 |
| -0.3 | 0.96 | 0.7 | 0.96 | 1 |
| -0.2 | 0.98 | 0.7 | 0.98 | 1 |
| -0.1 | 1.01 | 0.7 | 0.99 | 1 |
| 0.0 | 1.02 | 0.7 | 1.01 | 1 |
| 0.1 | 1.03 | 0.7 | 1.02 | 1 |
| 0.2 | 1.04 | 0.7 | 1.02 | 1 |
| 0.3 | 1.04 | 0.7 | 1.02 | 1 |
| 0.4 | 1.03 | 0.7 | 1.01 | 1 |
| 0.5 | 1.01 | 0.7 | 1.00 | 1 |
| 0.6 | 1.00 | 0.7 | 0.99 | 1 |
| 0.7 | 0.98 | 0.7 | 0.98 | 1 |
| 0.8 | 0.97 | 0.7 | 0.97 | 1 |
| 0.9 | 0.95 | 0.7 | 0.96 | 1 |
| 1.0 | 0.94 | 0.7 | 0.94 | 1 |
| 1.1 | 0.93 | 0.7 | 0.93 | 1 |
| 1.2 | 0.92 | 0.7 | 0.92 | 1 |
| 1.3 | 0.91 | 0.7 | 0.91 | 1 |

Table 4.A.18 *(continuation)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proficiency | MST 1–2 | Lower Line | MST 1–3 | Higher Line |
| 1.4 | 0.90 | 0.7 | 0.90 | 1 |
| 1.5 | 0.89 | 0.7 | 0.90 | 1 |
| 1.6 | 0.89 | 0.7 | 0.89 | 1 |
| 1.7 | 0.88 | 0.7 | 0.88 | 1 |
| 1.8 | 0.88 | 0.7 | 0.88 | 1 |
| 1.9 | 0.87 | 0.7 | 0.87 | 1 |
| 2.0 | 0.87 | 0.7 | 0.87 | 1 |
| 2.1 | 0.87 | 0.7 | 0.87 | 1 |
| 2.2 | 0.86 | 0.7 | 0.86 | 1 |
| 2.3 | 0.86 | 0.7 | 0.86 | 1 |
| 2.4 | 0.86 | 0.7 | 0.86 | 1 |
| 2.5 | 0.85 | 0.7 | 0.85 | 1 |

## The Utility of a Content Screener

### Study Purpose

Students received two performance tasks (PTs) in Segment B of the California Science Test (CAST), where each PT has a primary domain—Life Sciences (LS), Earth and Space Sciences (ESS), or Physical Sciences (PS). For the field test administration, students were randomly assigned any two PTs from the pool of eight available PTs in each tested grade. However, in the 2019 operational administration, students must receive the two PTs in two different science domains.

There are a number of ways in which PTs can be assigned. For instance, students could be randomly assigned to the two PTs with the restriction that the PTs must be from two different domains, or all students could be assigned the same two PTs, each from a different domain. Both of these assignment rules may result in advantaging or disadvantaging certain students if students are found to perform better in contexts with which those students are interested and experienced.

Alternatively, performance in Segment A, which is comprised of 32 to 34 items roughly spread evenly across the three domains, could be used to “screen out” PTs in the domain in which the student demonstrates conspicuously poor performance, so as not to disadvantage any student in the assignment of PTs received in Segment B. Students who, for instance, performed poorly on LS items in Segment A would be assigned PTs in the other two domains (PS and ESS), whereas students who performed similarly across all three domains in Segment A would be randomly assigned two PTs from any two domains. In either case, the tasks presented would be randomly chosen from the collection of permissible tasks. Doing so would help ensure that each student takes a test that is sufficiently broadly based but that encourages each student’s best performance.

The purpose of this study is to evaluate the utility of the content screener for implementation in future administrations.

### Methods

The screen-out rule would be helpful to inform the selection of PTs only if (1) student performance tended to differ by science domain and (2) student performance in Segment A was predictive of performance in Segment B. That is, students who performed poorly on LS items in Segment A also performed poorly in PTs set in LS contexts in Segment B and vice versa.

The intent of the content screener study is to address the following three research questions:

1. Does the performance of some students vary significantly across the three science domains in Segment A? [[2]](#footnote-3)
2. If so, is a student’s performance in Segment B affected by the context of the tasks that student is assigned?
3. If so, is the targeted assignment of tasks likely to have a practical and statistically significant impact on scores?

The answer to the second question depends, in part, on how pronounced or pervasive performance differences across content areas are for students who received Segment A. Such differences are a necessary, but not a sufficient, condition to cause performance in Segment B to be influenced by the assigned PT. For example, if all students show similar performance profiles across content areas in Segment A, there is little reason to select Segment B PTs differentially. But even if performance differences across students are generally found, this does not in itself guarantee that measurement of the California Next Generation Science Standards (CA NGSS) in Segment B is influenced—positively or negatively—by the context of measurement.

For differential selection to have value it must be shown from the data both that (a) performance differences across content areas in Segment A are present *and* (b) that these differences significantly and differentially impact Segment B performance. The third question is answered by applying different decision thresholds retrospectively to the data and determining which is best able to minimize the impact of Segment B tasks on performance.

The following steps outline the procedure implemented for the content screener study.

#### Step 1: Data Preparation

The content screener study was conducted after the preliminary item analysis, the differential item functioning analysis, the dimensionality study, and the calibration had been completed.

This study used all items from the operational items in segments A and B. The sample includes all students who completed two Segment A blocks and two Segment B blocks. Unmotivated students were identified and removed. These were students who

1. completed the assigned items in less than the sum of the minimum testing times for each of the blocks, which was determined by the first percentile of average multiple-choice item time from the CAST operational year for each grade level, to provide conservative estimates of the minimum time a student who is not motivated would take to complete an item; or
2. scored a zero on all selected-response (SR) items administered to the student across segments A, B, and C.

#### Step 2: Comparable Scores

Performance across any given set of items was scaled to a common metric that allowed for direct comparison across student scores regardless of which blocks of items were received in Segment A or which PTs were received in segments B and C.

An estimation method based on the test characteristic curve (TCC) was used that is more robust than methods based on item response patterns, since it depends on the item response theory (IRT) model fit being acceptable only in the aggregate, across items, rather than for each individual item. This TCC method requires that the TCC for the set of items in question (e.g., the LS component of a student’s test) first be computed by summing across the corresponding item response functions.

Because the TCC is aggregated across items, it is much less sensitive to estimation error or model fit issues than are the parameters estimated for any individual item. For each student, the TCC mapped ability values to expected point totals, which ran from zero to the maximum points possible across items in the set. The student’s observed point total was then translated to the corresponding ability value by inverting the TCC. This step was performed for each of the following sets of items:

1. All Segment A items (to produce a total Segment A score, *TA*)
2. LS items in Segment A (to produce a LS Segment A score, *LA*)
3. PS items in Segment A (to produce a PS Segment A score, *PA*)
4. ESS items in Segment A (to produce an ESS Segment A score, *EA*)
5. All items in PTs administered to each student (to produce a total Segment B score, *TB*)
6. LS items in the PTs set in LS context (to produce a LS Segment B score, *LB*)
7. PS items in the PTs set in PS context (to produce a PS Segment B score, *PB*)
8. ESS items in PTs set in ESS context (to produce an ESS Segment B score, *EB*)

#### Step 3: Descriptive Analyses

To orient toward the data, a set of descriptive analyses were first run. These analyses included the following:

1. Internal consistency reliabilities of all of the segments A and B total and domain scores that will be used in subsequent variables and analyses

* Cronbach’s alpha is calculated as the reliability for each score *s* with the following formula (equation 5.1):

 (5.1)

Where *K* is the number of items contributing to score *s*, is the variance of score of item , and is the variance of the observed score *s.*

1. Correlations and disattenuated correlations between each pair of segments A and B total and domain scores to provide indicators of the extent Segment B scores are related to Segment A scores

* Correlations were disattenuated by dividing the Pearson correlation by the square root of the product of the reliabilities.
* This analysis also included correlating individual PTs with Segment A domain scores to determine if particular PTs better correlate with scores in the same domain.

#### Step 4: The Alignment Index

An alignment index was used to quantify the extent to which the composition of Segment B tasks by domain aligns with the student’s score profile of strengths and weaknesses across the Segment A domains. An example of poor alignment would be a student receiving PTs in the two domains in which the student performed the weakest in Segment A, while an example of good alignment would be a student receiving PTs in the two domains in which the student performed the best in Segment A.

The alignment index is a product of a difference measure and an indicator characterizing Segment B composition. The difference measure is the vector of standardized differences between each student’s Segment A domain score and the Segment A total score. That is, it is composed of the following three standardized differences for LS (equation 5.2), PS (equation 5.3), and ESS (equation 5.4), respectively, for student *i*:

 (5.2)

 (5.3)

 (5.4)

where the standard deviation in each denominator is the standard deviation of the corresponding numerator.

Positive differences indicate strengths—that the student performed better in that domain than overall in Segment A; and negative differences indicate weaknesses—that the student performed worse in that domain than overall in Segment A.

The other component of the alignment index is the Segment B composition indicator. This indicator is a vector of three elements for each student that records the proportion of Segment B items in each of the three domain contexts. Specifically, it contains, , , and , where each specifies the proportion of items in the PTs in the LS, PS, and ESS domains, respectively. For instance, if a student was randomly assigned a five-item PT only in LS and another five-item PT in PS, then , , and .

The alignment index is the product of the Segment A difference and Segment B composition vectors. Specifically, it is the sum of the difference values weighted by the proportion of task items set in each domain (equation 5.5):

 (5.5)

Positive values indicate a composition that aligns with a student’s strengths, and negative values indicate a composition that aligns with a student’s weaknesses.

Summary statistics of the alignment index were examined, including mean, standard deviation, median, minimum, maximum, twenty-fifth, and seventy-fifth percentiles for all students, and by the following student groups available at the time of this analysis:

* Gender
  + Male
  + Female
* Race/Ethnicity
  + Hispanic or Latino
  + American Indian or Alaska Native
  + Black or African American
  + White
  + Native Hawaiian or Other Pacific Islander
  + Filipino
  + Two or more races
* Special education
* English language fluency
* Economically disadvantaged

#### Step 5: Fitted Linear Models

To evaluate whether students’ performance in Segment B is affected by the context of the tasks the students are assigned, a series of linear models was fit to answer the following questions:

1. How much variation in Segment B total scores is explained by the Segment A total scores?
2. Do Segment A domain scores improve the prediction of Segment B total scores?
3. Is prediction of Segment B total scores improved by characterizing the composition of Segment B and the student’s profile of strengths and weaknesses (i.e., the alignment index)?

The Segment B total score is used as the outcome because the interest is in the impact of the assigned tasks on a student’s total performance in Segment B. In addition, the total score is more reliable than any of the individual Segment B domain scores, given the few numbers of items on which the scores are each based.

In the following models, let instances of  denote the unknown coefficients (i.e., intercept or slope) in different models, and let instances of  denote the random errors in different models, which are assumed to be normally distributed with mean 0.

**Model 1:** How much variation in the Segment B composite can be explained by the Segment A composite (equation 5.6)?

 (5.6)

**Model 2:** How much (more) variation in the Segment B composite can be explained by the Segment A domain scores (equation 5.7)?

 (5.7)

**Model 3:** How much (more) variation in the Segment B composite can be explained by the Segment A domain scores and an index that measures the extent to which Segment B composition aligns with a student’s profile of strengths and weaknesses (equation 5.8)?

 (5.8)

To answer each of the three questions associated with each model, the adjusted *R*2 values were compared across the models. Operational implementation of a screener mechanism is advisable if Segment A domain scores added substantially to the prediction of Segment B scores and if the alignment index also added substantially to the prediction.

**Step 6: Follow-up Analyses**

Depending on the results of step 5, additional analyses may be needed to probe further into the utility of a screener and to test the impact of implementing a screener with different threshold values for the Segment A domain scores.

### Study Data

The number of PTs and distribution of PTs by content domain varied by grade level for the 2018–19 operational administration of the CAST. PTs were characterized by their primary domain—the main domain in which the task was set. In some instances, PTs also had a secondary domain—a small number of items that aligned to another content domain than the primary domain—but in these cases, all items within the PT were counted as corresponding to those of the primary domain. For grade five, there were five PTs total: one for LS, one for ESS, and three for PS, resulting in seven possible pairings of PTs from two different contents for Segment B. For grade eight, there were six PTs total—two for each content domain, resulting in 12 possible pairings of PTs for Segment B. For high school, there were only three PTs—one for each content domain, resulting in three possible pairings of PTs for Segment B. Students were roughly evenly distributed across the possible Segment B PT pairings. There were a total of 454,657 students in the grade five dataset, 495,940 students in the grade eight dataset, and 551,757 students in the high school dataset.

### Results

The results are organized by descriptive statistics (steps [3](#_Step_3:_Descriptive) and [4](#_Step_4:_The)) and model results ([step 5](#_Step_5:_Fitted)).

#### Descriptive Statistics

##### Correlations

Descriptive statistics were obtained for the scores and the alignment index. Table 5.1 through Table 5.3 give the reliabilities for each score, with the raw or Pearson correlations above the diagonal and the disattenuated correlations (or estimated true-score correlations) below the diagonal for grade five, grade eight, and high school, respectively. In several cases, the disattenuated correlations were greater than one, which typically occurs when score reliabilities are low. In such cases, the reliabilities were displayed as 1.0\*. Note that it was not possible to estimate correlations for all pairings of the PTs as no students were assigned two PTs in the same domain. Thus, for instance, in grade five, there are no correlations among PTs B2, B3, and B4, which all have a primary domain of Physical Sciences.

Table 5.1 provides the score reliabilities and correlations for grade five.

Table 5.1 Score Reliabilities and Correlations Between Scores, Grade Five

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Segment | Total Segment A | Life Sciences A | Earth and Space Sciences A | Physical Sciences A | Total Segment B | B1 (Life Sciences) | B2 (Physical Sciences) | B3 (Physical Sciences) | B4 (Physical Sciences) | B5 (Earth and Space Sciences) |
| Total Segment A | 1.00 | 0.75 | 0.86 | 0.87 | 0.64 | 0.53 | 0.48 | 0.54 | 0.58 | 0.54 |
| Life Sciences A | 0.93 | 1.00 | 0.57 | 0.58 | 0.49 | 0.41 | 0.38 | 0.41 | 0.44 | 0.40 |
| Earth and Space Sciences A | 1.0\* | 0.77 | 1.00 | 0.64 | 0.55 | 0.45 | 0.41 | 0.47 | 0.50 | 0.47 |
| Physical Sciences A | 1.0\* | 0.79 | 0.88 | 1.00 | 0.55 | 0.46 | 0.42 | 0.46 | 0.49 | 0.46 |
| Total Segment B | 0.81 | 0.68 | 0.77 | 0.77 | 1.00 | 0.71 | 0.78 | 0.76 | 0.82 | 0.73 |
| B1 (Life Sciences) | 0.77 | 0.66 | 0.72 | 0.74 | 1.0\* | 1.00 | 0.29 | 0.32 | 0.35 | 0.33 |
| B2 (Physical Sciences) | 1.0\* | 1.0\* | 1.0\* | 1.0\* | 1.0\* | 1.0\* | 1.00 | NA | NA | 0.31 |
| B3 (Physical Sciences) | 0.66 | 0.55 | 0.63 | 0.62 | 1.0\* | 0.51 | NA | 1.00 | NA | 0.38 |
| B4 (Physical Sciences) | 0.89 | 0.74 | 0.85 | 0.84 | 1.0\* | 0.71 | NA | NA | 1.00 | 0.39 |
| B5 (Earth and Space Sciences) | 0.83 | 0.68 | 0.80 | 0.78 | 1.0\* | 0.67 | 1.0\* | 0.63 | 0.83 | 1.00 |
| Reliability | 0.89 | 0.73 | 0.74 | 0.73 | 0.70 | 0.53 | 0.12 | 0.76 | 0.47 | 0.47 |

Table 5.2 provides the score reliabilities and correlations for grade eight.

Table 5.2 Score Reliabilities and Correlations Between Scores, Grade Eight

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Segment | Total Segment A | Life Sciences A | Earth and Space Sciences A | Physical Sciences A | Total Segment B | B1 (Life Sciences) | B2 (Life Sciences) | B3 (Physical Sciences) | B4 (Physical Sciences) | B5 (Earth and Space Sciences) | B6 (Earth and Space Sciences) |
| Total Segment A | 1.00 | 0.82 | 0.82 | 0.89 | 0.65 | 0.50 | 0.53 | 0.59 | 0.52 | 0.51 | 0.56 |
| Life Sciences A | 1.0\* | 1.00 | 0.58 | 0.61 | 0.53 | 0.41 | 0.43 | 0.48 | 0.42 | 0.43 | 0.46 |
| Earth and Space Sciences A | 1.0\* | 0.92 | 1.00 | 0.61 | 0.52 | 0.40 | 0.43 | 0.48 | 0.41 | 0.41 | 0.46 |
| Physical Sciences A | 1.0\* | 0.89 | 0.90 | 1.00 | 0.58 | 0.46 | 0.46 | 0.54 | 0.47 | 0.45 | 0.50 |
| Total Segment B | 0.86 | 0.82 | 0.83 | 0.83 | 1.00 | 0.72 | 0.76 | 0.80 | 0.78 | 0.80 | 0.81 |

Table 5.2 (*continuation)*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Segment | Total Segment A | Life Sciences A | Earth and Space Sciences A | Physical Sciences A | Total Segment B | B1 (Life Sciences) | B2 (Life Sciences) | B3 (Physical Sciences) | B4 (Physical Sciences) | B5 (Earth and Space Sciences) | B6 (Earth and Space Sciences) |
| B1 (Life Sciences) | 1.0\* | 1.0\* | 1.0\* | 1.0\* | 1.0\* | 1.00 | NA | 0.36 | 0.31 | 0.32 | 0.33 |
| B2 (Life Sciences) | 0.89 | 0.83 | 0.86 | 0.84 | 1.0\* | NA | 1.00 | 0.38 | 0.33 | 0.32 | 0.36 |
| B3 (Physical Sciences) | 0.75 | 0.71 | 0.72 | 0.73 | 1.0\* | 1.0\* | 0.71 | 1.00 | NA | 0.37 | 0.41 |
| B4 (Physical Sciences) | 0.90 | 0.84 | 0.85 | 0.89 | 1.0\* | 1.0\* | 0.71 | NA | 1.00 | 0.30 | 0.37 |
| B5 (Earth and Space Sciences) | 0.98 | 0.96 | 0.94 | 0.93 | 1.0\* | 1.0\* | 0.84 | 0.79 | 0.88 | 1.00 | NA |
| B6 (Earth and Space Sciences) | 0.82 | 0.78 | 0.79 | 0.78 | 1.0\* | 1.0\* | 0.89 | 0.65 | 0.80 | NA | 1.00 |
| Reliability | 0.86 | 0.64 | 0.61 | 0.74 | 0.66 | 0.16 | 0.41 | 0.72 | 0.38 | 0.31 | 0.55 |

Table 5.3 provides the score reliabilities and correlations for high school.

Table 5.3 Score Reliabilities and Correlations Between Scores, High School

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Segment | Segment A | Life Sciences A | Earth and Space Sciences A | Physical Sciences A | Segment B | B1 (Life Sciences) | B2 (Physical Sciences) | B3 (Earth and Space Sciences) |
| Total Segment A | 1.00 | 0.86 | 0.81 | 0.82 | 0.62 | 0.46 | 0.55 | 0.56 |
| Life Sciences A | 1.0\* | 1.00 | 0.56 | 0.57 | 0.54 | 0.41 | 0.48 | 0.49 |
| Earth and Space Sciences A | 1.0\* | 0.89 | 1.00 | 0.55 | 0.50 | 0.36 | 0.44 | 0.46 |
| Physical Sciences A | 1.0\* | 0.85 | 0.88 | 1.00 | 0.50 | 0.37 | 0.45 | 0.44 |

Table 5.3 *(continuation)*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Segment | Segment A | Life Sciences A | Earth and Space Sciences A | Physical Sciences A | Segment B | B1 (Life Sciences) | B2 (Physical Sciences) | B3 (Earth and Space Sciences) |
| Total Segment B | 0.82 | 0.80 | 0.79 | 0.75 | 1.00 | 0.74 | 0.80 | 0.78 |
| B1 (Life Sciences) | 0.66 | 0.65 | 0.61 | 0.60 | 1.0\* | 1.00 | 0.32 | 0.31 |
| B2 (Physical Sciences) | 0.71 | 0.68 | 0.67 | 0.65 | 1.0\* | 0.50 | 1.00 | 0.43 |
| B3 (Earth and Space Sciences) | 0.74 | 0.72 | 0.73 | 0.66 | 1.0\* | 0.50 | 0.61 | 1.00 |
| Reliability | 0.84 | 0.68 | 0.59 | 0.65 | 0.67 | 0.58 | 0.72 | 0.67 |

First, correlation by domain within the same segment was examined to begin to address research question 1 regarding the extent performance varied by domain, particularly in Segment A. This descriptive analysis provided an overall sense of the relationship among domain scores—even if the correlations were near 1, there could still be outlying students with true differences in performance across domains.

The Segment A domain scores have raw correlations from .55 to .64 and disattenuated correlations, or estimates of the correlations after accounting for measurement error, from .77 to .92 across the three grade levels and three pairings of the domain scores. The Segment B domain scores, where each was based on about 6 items, have weak raw correlations of .29 to .43 but disattenuated correlations of .5 to 1, suggesting that in some cases once measurement error is accounted for, there is no variability in performance across the PTs by domain. In contrast, the Segment A domain correlations suggest that student performance varies to some degree by domain in Segment A. We mainly have moderate correlations between the domain scores due to measurement error but if the measurement error was accounted for, we would have a stronger association which would indicate that these domain scores are not measuring unique aspects of NGSS.

The dimensionality study (refer to Chapter 3: Test Dimensionality), which uses multidimensional item response theory (MIRT) models, also provided estimated true-score correlations by domain. Although it pooled across both segments A and B, using a correlated three-factor MIRT model among the three domains, the dimensionality study estimated true-score correlations of .88 to .97 among the domains.

Such high correlations provide evidence against much variation in performance by science content domain on the CAST, making the utility of a screener unlikely. However, further analysis is needed, as there may still be outlying students who would benefit from a screener. In general, it is expected that student performance across domains is comparable and no screener would be needed (i.e., students would be randomly assigned two PTs from two different domains), but the interest is in the outlying students and the extent that a screener would benefit them.

Next, the correlations between the PTs and the Segment A domain scores were examined. The raw correlations between each PT and the Segment A domain scores are generally weak to moderate, ranging from .36 to .54. However, once disattenuated for measurement error, they are moderate to strong, ranging from .55 to 1. Only in some cases is the Segment B PT correlated the most with the corresponding Segment A domain, suggesting that performance by domain in Segment A may not be strongly predictive of PT performance.

##### Alignment Index

The alignment index, which is central to this study, was also examined before fitting the linear models of interest. The index is a linear combination of each student’s standardized difference Segment A domain scores, weighted by the proportion of PT items in each domain. The standardized differences defined in [step 4](#_Step_4:_The) were the difference between each Segment A domain score and the total Segment A score divided by the standard deviation of this difference.

Table 5.4 through Table 5.6 provide the summary statistics of the alignment index for grade five, grade eight, and high school, respectively. If students were generally assigned PTs in domains that corresponded to the domains in which the students performed the best in Segment A, the alignment index would be positive, and vice versa if students were generally assigned PTs in domains in the students performed poorly in Segment A.

Mean and median values for all students and each student group are close to zero, as presented in Table 5.4 through Table 5.6. This indicates that, on average, either students were generally performing comparably across the Segment A domains the PTs came from; or any strong performance (i.e., positive standardized difference) for a domain in Segment A was counterweighted by a weaker performance (i.e., negative standardized difference) in the other domain for which the students received a PT.

Table 5.4 through Table 5.6 also show that the alignment index is distributed similarly across each student group. In the cases where the mean alignment index value tends to deviate from the overall near zero means, the median tends to be much smaller and closer to zero. Thus, no differential alignment by student group, which is to be expected given the random assignment of PTs to students. However, for grade five students, there was more uncertainty in the alignment index values for Asian and students who are designated as initially fluent English proficient. These student groups tended to outperform other student groups on the test, which might indicate that, for higher performing students, the assigned PTs might not align well with their domain-specific strengths. For grade eight and high school tests, the disparities in the alignment index standard deviations across student groups were less pronounced in comparison to the grade five test.

Approximately 66 percent of students were assigned a PT in the domain in which the students performed the weakest in Segment A. This percentage follows from what would be expected because of random assignment. For example, in high school, there are three possible pairs of PTs (ESS–LS, ESS–PS, and LS–PS), and each domain appears in exactly two of the three pairs, so the probability is two thirds, or .66.

Table 5.4 Summary Statistics for the Alignment Index Across Student Groups, Grade Five

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Student Group | N | Minimum | 25th Percentile | Median | Mean | 75th Percentile | Maximum | Standard Deviation |
| Overall | 454,657 | -6.06 | -0.19 | 0.02 | 0.07 | 0.24 | 5.20 | 0.62 |
| Female | 221,080 | -6.06 | -0.19 | 0.02 | 0.07 | 0.24 | 4.91 | 0.61 |
| Male | 233,577 | -5.78 | -0.19 | 0.02 | 0.07 | 0.24 | 5.20 | 0.64 |
| Hispanic or Latino | 252,478 | -6.06 | -0.21 | 0.00 | 0.00 | 0.20 | 5.06 | 0.56 |
| American Indian or Alaskan Native | 2,160 | -4.35 | -0.21 | 0.00 | 0.01 | 0.21 | 4.55 | 0.56 |
| Asian | 42,587 | -4.99 | -0.15 | 0.08 | 0.25 | 0.37 | 4.74 | 0.75 |
| Black or African American | 24,114 | -4.35 | -0.22 | -0.01 | -0.04 | 0.20 | 4.91 | 0.59 |
| White | 100,361 | -5.44 | -0.17 | 0.05 | 0.18 | 0.31 | 5.20 | 0.68 |
| Native Hawaiian or Other Pacific Islander | 2,069 | -4.32 | -0.22 | 0.01 | 0.00 | 0.22 | 3.14 | 0.61 |
| Filipino | 9,183 | -3.75 | -0.16 | 0.04 | 0.16 | 0.29 | 4.74 | 0.63 |
| Two or more races | 18,622 | -4.09 | -0.17 | 0.05 | 0.19 | 0.33 | 4.74 | 0.71 |
| English learners | 88,177 | -6.06 | -0.25 | -0.03 | -0.11 | 0.17 | 4.34 | 0.56 |
| English or American Sign Language Only | 262,411 | -5.78 | -0.18 | 0.03 | 0.11 | 0.26 | 5.20 | 0.64 |
| Initial fluent English proficient | 17,893 | -4.09 | -0.15 | 0.07 | 0.27 | 0.37 | 5.06 | 0.75 |
| Reclassified fluent English proficient | 85,500 | -4.32 | -0.17 | 0.03 | 0.10 | 0.24 | 4.91 | 0.55 |
| No special education services | 398,047 | -6.06 | -0.18 | 0.03 | 0.10 | 0.25 | 5.20 | 0.62 |
| Special education services | 56,610 | -5.78 | -0.26 | -0.03 | -0.10 | 0.18 | 4.55 | 0.64 |
| Not limited English proficient | 366,480 | -5.78 | -0.18 | 0.03 | 0.12 | 0.26 | 5.20 | 0.63 |
| Limited English proficient | 88,177 | -6.06 | -0.25 | -0.03 | -0.11 | 0.17 | 4.34 | 0.56 |
| No Section 504 Plan | 447,569 | -6.06 | -0.19 | 0.02 | 0.07 | 0.24 | 5.20 | 0.62 |
| With Section 504 Plan | 7,088 | -3.75 | -0.18 | 0.03 | 0.12 | 0.26 | 4.55 | 0.63 |
| Not Economically disadvantaged | 171,840 | -5.78 | -0.16 | 0.05 | 0.20 | 0.33 | 5.20 | 0.70 |
| Economically disadvantaged | 282,817 | -6.06 | -0.22 | 0.00 | 0.00 | 0.20 | 5.06 | 0.56 |
| Nonmigrant | 450,863 | -6.06 | -0.19 | 0.02 | 0.07 | 0.24 | 5.20 | 0.62 |
| Migrant | 3,794 | -4.35 | -0.23 | -0.02 | -0.05 | 0.18 | 4.91 | 0.53 |

Table 5.5 Summary Statistics for the Alignment Index Across Student Groups, Grade Eight

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Student Group | N | Minimum | 25th Percentile | Median | Mean | 75th Percentile | Maximum | Standard Deviation |
| Overall | 459,540 | -5.77 | -0.26 | 0.01 | 0.00 | 0.27 | 4.92 | 0.57 |
| Female | 225,199 | -5.71 | -0.25 | 0.01 | 0.01 | 0.27 | 4.85 | 0.54 |
| Male | 234,341 | -5.77 | -0.27 | 0.00 | -0.02 | 0.27 | 4.92 | 0.60 |
| Hispanic or Latino | 252,491 | -5.76 | -0.27 | -0.01 | -0.05 | 0.24 | 4.85 | 0.55 |
| American Indian or Alaskan Native | 2,359 | -3.35 | -0.29 | 0.00 | -0.07 | 0.24 | 2.98 | 0.58 |
| Asian | 43,514 | -4.70 | -0.23 | 0.05 | 0.14 | 0.38 | 4.92 | 0.65 |
| Black or African American | 23,931 | -5.31 | -0.30 | -0.03 | -0.09 | 0.23 | 3.57 | 0.60 |
| White | 104,964 | -4.87 | -0.24 | 0.03 | 0.05 | 0.31 | 4.85 | 0.57 |
| Native Hawaiian or Other Pacific Islander | 2,171 | -4.87 | -0.26 | -0.01 | -0.03 | 0.25 | 3.28 | 0.52 |
| Filipino | 10,785 | -4.52 | -0.24 | 0.03 | 0.06 | 0.31 | 3.56 | 0.54 |
| Two or more races | 16,809 | -4.52 | -0.24 | 0.03 | 0.06 | 0.31 | 4.50 | 0.60 |
| English learners | 56,678 | -5.77 | -0.35 | -0.06 | -0.17 | 0.20 | 3.04 | 0.64 |
| English or American Sign Language Only | 251,216 | -5.31 | -0.25 | 0.01 | 0.02 | 0.28 | 4.85 | 0.57 |
| Initial fluent English proficient | 19,874 | -4.87 | -0.23 | 0.05 | 0.12 | 0.36 | 4.85 | 0.62 |
| Reclassified fluent English proficient | 131,083 | -5.28 | -0.25 | 0.01 | 0.01 | 0.27 | 4.92 | 0.52 |
| No special education services | 407,881 | -5.77 | -0.25 | 0.02 | 0.02 | 0.28 | 4.92 | 0.56 |
| Special education services | 51,659 | -5.31 | -0.34 | -0.05 | -0.16 | 0.21 | 3.84 | 0.66 |
| Not limited English proficient | 402,862 | -5.31 | -0.25 | 0.02 | 0.02 | 0.28 | 4.92 | 0.56 |
| Limited English proficient | 56,678 | -5.77 | -0.35 | -0.06 | -0.17 | 0.20 | 3.04 | 0.64 |
| No Section 504 Plan | 449,640 | -5.77 | -0.26 | 0.01 | 0.00 | 0.27 | 4.92 | 0.57 |
| With Section 504 Plan | 9,900 | -4.18 | -0.26 | 0.01 | 0.00 | 0.28 | 4.10 | 0.57 |
| Not economically disadvantaged | 181,584 | -5.77 | -0.24 | 0.03 | 0.07 | 0.32 | 4.92 | 0.58 |
| Economically disadvantaged | 277,956 | -5.76 | -0.28 | -0.01 | -0.05 | 0.24 | 4.85 | 0.56 |
| Nonmigrant | 455,828 | -5.77 | -0.26 | 0.01 | 0.00 | 0.27 | 4.92 | 0.57 |
| Migrant | 3,712 | -4.87 | -0.26 | 0.00 | -0.06 | 0.24 | 3.41 | 0.55 |

Table 5.6 Summary Statistics for the Alignment Index Across Student Groups, High School

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Student Group | N | Minimum | 25th Percentile | Median | Mean | 75th Percentile | Maximum | Standard Deviation |
| Overall | 551,757 | -5.34 | -0.32 | -0.02 | -0.07 | 0.25 | 3.68 | 0.56 |
| Female | 272,317 | -5.34 | -0.31 | -0.02 | -0.07 | 0.24 | 3.32 | 0.53 |
| Male | 279,440 | -5.34 | -0.32 | -0.02 | -0.08 | 0.25 | 3.68 | 0.58 |
| Hispanic or Latino | 290,724 | -5.34 | -0.33 | -0.04 | -0.11 | 0.23 | 3.44 | 0.56 |
| American Indian or Alaskan Native | 2,875 | -4.08 | -0.32 | -0.03 | -0.08 | 0.25 | 3.11 | 0.56 |
| Asian | 54,309 | -4.81 | -0.27 | 0.01 | 0.02 | 0.30 | 3.68 | 0.54 |
| Black or African American | 28,993 | -5.31 | -0.39 | -0.05 | -0.16 | 0.22 | 2.70 | 0.62 |
| White | 134,684 | -5.34 | -0.29 | -0.01 | -0.03 | 0.26 | 3.50 | 0.54 |
| Native Hawaiian or Other Pacific Islander | 2,552 | -4.03 | -0.35 | -0.05 | -0.13 | 0.21 | 1.88 | 0.58 |
| Filipino | 16,639 | -4.08 | -0.28 | -0.02 | -0.03 | 0.25 | 3.02 | 0.50 |
| Two or more races | 17,396 | -4.81 | -0.29 | -0.01 | -0.04 | 0.26 | 2.86 | 0.55 |
| English learners | 45,634 | -4.99 | -0.45 | -0.09 | -0.22 | 0.19 | 2.75 | 0.67 |
| English or American Sign Language Only | 294,533 | -5.34 | -0.30 | -0.02 | -0.06 | 0.25 | 3.50 | 0.55 |
| Initial fluent English proficient | 29,090 | -4.67 | -0.27 | 0.00 | 0.00 | 0.27 | 3.25 | 0.53 |
| Reclassified fluent English proficient | 181,741 | -5.34 | -0.31 | -0.02 | -0.07 | 0.24 | 3.68 | 0.53 |
| No special education services | 501,387 | -5.34 | -0.31 | -0.02 | -0.06 | 0.25 | 3.68 | 0.54 |
| Special education services | 50,370 | -5.14 | -0.44 | -0.08 | -0.21 | 0.19 | 3.13 | 0.68 |
| Not limited English proficient | 506,123 | -5.34 | -0.31 | -0.02 | -0.06 | 0.25 | 3.68 | 0.54 |
| Limited English proficient | 45,634 | -4.99 | -0.45 | -0.09 | -0.22 | 0.19 | 2.75 | 0.67 |
| No Section 504 Plan | 537,083 | -5.34 | -0.32 | -0.02 | -0.07 | 0.25 | 3.68 | 0.56 |
| With Section 504 Plan | 14,674 | -4.81 | -0.30 | -0.02 | -0.06 | 0.25 | 3.00 | 0.56 |
| Not economically disadvantaged | 239,936 | -5.14 | -0.29 | -0.01 | -0.03 | 0.26 | 3.68 | 0.54 |
| Economically disadvantaged | 311,821 | -5.34 | -0.33 | -0.04 | -0.11 | 0.23 | 3.19 | 0.57 |
| Nonmigrant | 547,941 | -5.34 | -0.32 | -0.02 | -0.07 | 0.25 | 3.68 | 0.56 |
| Migrant | 3,816 | -4.67 | -0.34 | -0.03 | -0.11 | 0.23 | 2.05 | 0.56 |

#### Model Results

The linear models assessing the extent that Segment A performance predicts Segment B performance were evaluated using R2 and the adjusted-R2. Analysis shows that the R2 and the adjusted-R2 values are all very similar, meaning that adding the Segment A domain scores, the alignment index, or both the Segment A domain scores and the alignment index does not explain much more variance in the Segment B scores.

As the models are nested within each other, likelihood-ratio tests can be used to test whether a model fits significantly better than the other models. The second model, which contains the total Segment A scores and domain Segment A subscores, fits significantly better than the first model, which contains just the total Segment A scores. However, given the very small increase in R2 values, this statistical significance does not seem to have high practical significance.

The third model that adds to the alignment index also fits significantly better than model two. However, the R2 and the adjusted-R2 values are almost identical up to the fourth decimal place for these two models. Upon further inspection, it is found that the alignment index is correlated about .27 with the total Segment B score in grade five, .21 in grade eight, and .18 in high school, indicating that there is not a strong relationship between the alignment index and the total Segment B score. Moreover, it is only for students with alignment index values in the bottom 5 percent that tend to have lower Segment B scores and those with alignment index values in the top 5 percent that tend to have higher Segment B scores. For all other students, there is no relationship between the Segment B score and the alignment.

Table 5.7 provides the results for the three models.

Table 5.7 R2 and the Adjusted-R2 Values for the Fitted Linear Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Grade | Model Number | Model Description | R2 | Adjusted R2 |
| Five | One | Total Segment A score only | 0.4097 | 0.4097 |
| Five | Two | + Segment A domain subscores | 0.4100 | 0.4100 |
| Five | Three | + Alignment index | 0.4100 | 0.4100 |
| Eight | One | Total Segment A score only | 0.4199 | 0.4199 |
| Eight | Two | + Segment A domain subscores | 0.4204 | 0.4204 |
| Eight | Three | + Alignment index | 0.4204 | 0.4204 |
| High School | One | Total Segment A score only | 0.3835 | 0.3835 |
| High School | Two | + Segment A domain subscores | 0.3844 | 0.3844 |
| High School | Three | + Alignment index | 0.3845 | 0.3845 |

#### Follow-Up Analyses

If the fitted models had indicated that a screener may be useful, additional analyses testing threshold values would have been implemented to determine the actual impact of a screener on student scores. In this case, because the linear models do not provide strong evidence in favor of a screener, a small experiment was conducted to probe further into the possibility that the screener may be useful. For students who were assigned a PT in the domain that corresponds to the students’ weakest Segment A domain score, evidence for the utility of a screener would be that such students have significantly lower total Segment B scores than students who are not assigned a PT in the weakest domain. Given that it is outlying students who are of most interest, for this experiment the sample was restricted to those students whose standardized difference in the weakest Segment A domain was in the bottom 25 percent (i.e., less than or equal to the 25th percentile). This cutoff is one possible threshold for identifying students most likely to benefit from the screener.

First, students were separated into groups by which Segment A domain the students were ranked the lowest in. These groups were further subset to students who demonstrated conspicuous weakness in the domain of interest; that is, if the students’ standardized difference fell in the bottom 25 percent. Then, within each group, students’ total Segment B scores were compared for each pair of PTs.

Table 5.8 provides the mean and standard deviation (SD) of total Segment B scores for each pair of PTs for each group of students and indicates whether the PT pair included the weakest domain, as indicated by Segment A. The means are given for each pair of PTs instead of the simple dichotomy of “assigned weakest” versus “not assigned weakest,” as these categories generally combine more than one pair of PTs and may mask effects of particular PTs on a student’s total Segment B score.

Table 5.8 Average Total Segment B Scores for Students by the Students’ Weakest Domain

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | Weakest Domain | PTs | Assigned Weakest | N | Average B Score | SD B Score |
| Five | Life Sciences | B1 (LS)—B2 (PS) | Yes | 5212 | 0.36 | 2.05 |
| Five | Life Sciences | B1 (LS)—B3 (PS) | Yes | 5452 | 0.36 | 1.86 |
| Five | Life Sciences | B1 (LS)—B4 (PS) | Yes | 5420 | 0.38 | 1.97 |
| Five | Life Sciences | B1 (LS)—B5 (ESS) | Yes | 4330 | 0.19 | 2.17 |
| Five | Life Sciences | B2 (PS)—B5 (ESS) | No | 5373 | 0.27 | 2.04 |
| Five | Life Sciences | B3 (PS)—B5 (ESS) | No | 5418 | 0.37 | 1.81 |
| Five | Life Sciences | B4 (PS)—B5 (ESS) | No | 5196 | 0.27 | 1.96 |
| Five | Earth and Space Sciences | B1 (LS)—B5 (ESS) | Yes | 5101 | -0.78 | 2.27 |
| Five | Earth and Space Sciences | B2 (PS)—B5 (ESS) | Yes | 6197 | -0.70 | 2.07 |
| Five | Earth and Space Sciences | B3 (PS)—B5 (ESS) | Yes | 6214 | -0.48 | 1.72 |
| Five | Earth and Space Sciences | B4 (PS)—B5 (ESS) | Yes | 6243 | -0.58 | 1.90 |
| Five | Earth and Space Sciences | B1 (LS)—B2 (PS) | No | 6363 | -0.54 | 2.01 |
| Five | Earth and Space Sciences | B1 (LS)—B3 (PS) | No | 6304 | -0.39 | 1.62 |
| Five | Earth and Space Sciences | B1 (LS)—B4 (PS) | No | 6417 | -0.54 | 1.89 |
| Five | Physical Sciences | B1 (LS)—B2 (PS) | Yes | 5414 | -0.31 | 2.06 |
| Five | Physical Sciences | B1 (LS)—B3 (PS) | Yes | 5325 | -0.20 | 1.79 |
| Five | Physical Sciences | B1 (LS)—B4 (PS) | Yes | 5485 | -0.26 | 1.96 |
| Five | Physical Sciences | B2 (PS)—B5 (ESS) | Yes | 5428 | -0.38 | 2.09 |
| Five | Physical Sciences | B3 (PS)—B5 (ESS) | Yes | 5278 | -0.25 | 1.77 |
| Five | Physical Sciences | B4 (PS)—B5 (ESS) | Yes | 5267 | -0.39 | 2.01 |
| Five | Physical Sciences | B1 (LS)—B5 (ESS) | No | 4239 | -0.54 | 2.34 |

Table 5.8 *(continuation one)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | Weakest Domain | PTs | Assigned Weakest | N | Average B Score | SD B Score |
| Eight | Life Sciences | B1 (LS)—B3 (PS) | Yes | 3362 | -0.35 | 1.96 |
| Eight | Life Sciences | B1 (LS)—B4 (PS) | Yes | 3290 | -0.42 | 2.12 |
| Eight | Life Sciences | B1 (LS)—B5 (ESS) | Yes | 3185 | -0.59 | 2.31 |
| Eight | Life Sciences | B1 (LS)—B6 (ESS) | Yes | 2647 | -0.21 | 1.93 |
| Eight | Life Sciences | B2 (LS)—B3 (PS) | Yes | 3227 | -0.28 | 1.87 |
| Eight | Life Sciences | B2 (LS)—B4 (PS) | Yes | 3383 | -0.36 | 1.90 |
| Eight | Life Sciences | B2 (LS)—B5 (ESS) | Yes | 3192 | -0.51 | 2.09 |
| Eight | Life Sciences | B2 (LS)—B6 (ESS) | Yes | 2761 | -0.18 | 1.84 |
| Eight | Life Sciences | B3 (PS)—B5 (ESS) | No | 3262 | -0.36 | 1.93 |
| Eight | Life Sciences | B3 (PS)—B6 (ESS) | No | 2852 | -0.12 | 1.75 |
| Eight | Life Sciences | B4 (PS)—B5 (ESS) | No | 3293 | -0.44 | 2.05 |
| Eight | Life Sciences | B4 (PS)—B6 (ESS) | No | 2717 | -0.14 | 1.88 |
| Eight | Earth and Space Sciences | B1 (LS)—B5 (ESS) | Yes | 3861 | -0.55 | 2.20 |
| Eight | Earth and Space Sciences | B1 (LS)—B6 (ESS) | Yes | 3315 | -0.26 | 1.86 |
| Eight | Earth and Space Sciences | B2 (LS)—B5 (ESS) | Yes | 3898 | -0.49 | 1.95 |
| Eight | Earth and Space Sciences | B2 (LS)—B6 (ESS) | Yes | 3194 | -0.25 | 1.71 |
| Eight | Earth and Space Sciences | B3 (PS)—B5 (ESS) | Yes | 3923 | -0.38 | 1.78 |
| Eight | Earth and Space Sciences | B3 (PS)—B6 (ESS) | Yes | 3196 | -0.12 | 1.67 |
| Eight | Earth and Space Sciences | B4 (PS)—B5 (ESS) | Yes | 3938 | -0.42 | 2.01 |
| Eight | Earth and Space Sciences | B4 (PS)—B6 (ESS) | Yes | 3241 | -0.15 | 1.89 |
| Eight | Earth and Space Sciences | B1 (LS)—B3 (PS) | No | 3970 | -0.31 | 1.82 |
| Eight | Earth and Space Sciences | B1 (LS)—B4 (PS) | No | 3866 | -0.47 | 2.12 |
| Eight | Earth and Space Sciences | B2 (LS)—B3 (PS) | No | 3871 | -0.29 | 1.72 |
| Eight | Earth and Space Sciences | B2 (LS)—B4 (PS) | No | 3921 | -0.37 | 1.88 |

Table 5.8 *(continuation two)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | Weakest Domain | PTs | Assigned Weakest | N | Average B Score | SD B Score |
| Eight | Physical Sciences | B1 (LS)—B3 (PS) | Yes | 3134 | -0.16 | 1.41 |
| Eight | Physical Sciences | B1 (LS)—B4 (PS) | Yes | 3080 | -0.29 | 1.81 |
| Eight | Physical Sciences | B2 (LS)—B3 (PS) | Yes | 3022 | -0.05 | 1.47 |
| Eight | Physical Sciences | B2 (LS)—B4 (PS) | Yes | 3114 | -0.17 | 1.64 |
| Eight | Physical Sciences | B3 (PS)—B5 (ESS) | Yes | 3013 | -0.11 | 1.48 |
| Eight | Physical Sciences | B3 (PS)—B6 (ESS) | Yes | 2714 | 0.01 | 1.38 |
| Eight | Physical Sciences | B4 (PS)—B5 (ESS) | Yes | 3003 | -0.17 | 1.63 |
| Eight | Physical Sciences | B4 (PS)—B6 (ESS) | Yes | 2698 | -0.08 | 1.48 |
| Eight | Physical Sciences | B1 (LS)—B5 (ESS) | No | 2987 | -0.35 | 1.80 |
| Eight | Physical Sciences | B1 (LS)—B6 (ESS) | No | 2680 | -0.12 | 1.52 |
| Eight | Physical Sciences | B2 (LS)—B5 (ESS) | No | 3024 | -0.19 | 1.68 |
| Eight | Physical Sciences | B2 (LS)—B6 (ESS) | No | 2706 | -0.09 | 1.49 |
| High School | Life Sciences | B1 (LS)—B2 (PS) | Yes | 15520 | -0.75 | 2.01 |
| High School | Life Sciences | B1 (LS)—B3 (ESS) | Yes | 15107 | -0.88 | 1.85 |
| High School | Life Sciences | B2 (PS)—B3 (ESS) | No | 15279 | -0.82 | 1.86 |
| High School | Earth and Space Sciences | B1 (LS)—B3 (ESS) | Yes | 15370 | -0.56 | 1.98 |
| High School | Earth and Space Sciences | B2 (PS)—B3 (ESS) | Yes | 15271 | -0.47 | 1.87 |
| High School | Earth and Space Sciences | B1 (LS)—B2 (PS) | No | 15340 | -0.36 | 2.18 |
| High School | Physical Sciences | B1 (LS)—B2 (PS) | Yes | 16056 | -0.90 | 1.93 |
| High School | Physical Sciences | B2 (PS)—B3 (ESS) | Yes | 16057 | -0.95 | 1.77 |
| High School | Physical Sciences | B1 (LS)—B3 (ESS) | No | 16031 | -0.98 | 1.90 |

The results were mixed. There were several cases for which students performed as well as or better on average when assigned a PT in the students’ conspicuously weak Segment A domain than when not assigned a PT in the weakest domain. In most cases, the rank ordering of performance on the Segment B PT pairs generally corresponded to that of the overall population and was more indicative of the difficulty of the particular PTs than on whether the students were assigned a PT in the students’ weak Segment A domain. For instance, in grade five, students who are conspicuously weak in Physical Sciences and are not assigned any of the three Physical Sciences PTs—B2, B3, or B4—(i.e., who received B1 and B5) perform worse than (mean performance = -0.54) students who receive any of the Physical Science PTs (mean performance ranges from -0.39 to -0.20). In the overall grade five population, B1 (LS) and B5 (ESS) are the PTs that students perform the worst on in grade five, and, in this case, this relative difficulty holds even for students who perform better in LS and ESS in Segment A PS.

### Limitations

The CAST 2018–19 operational administration offered much richer data to use for this analysis than the 2017–18 field test. However, for grade five and high school, two to three of the content domains only had one PT, limiting the extent that results are generalizable beyond these particular PTs.

### Implications for Test Blueprint Change

The CAST 2018–19 operational administration screener study provided weak evidence that a screener from Segment A to Segment B would be useful in limiting the extent that students are advantaged or disadvantaged by the domain-specific PTs the students are assigned. It is recommended that a screener not be implemented for the current test blueprint. Note that the content screener is only applicable when students are not assigned PTs from all science domains; as a result, different students might be receiving PTs from different domains. The content screener was therefore proposed to minimize the chance of some students being disadvantaged. If the updates to the test are adopted, where each student will be receiving three PTs, one from each domain, the need for the content screener is eliminated.

## Summary of Recommendations

The recommendations in this report are informed by psychometric studies conducted using data from the 2018–19 operational administration of the California Science Test (CAST).

Table 6.1 summarizes the California State Board of Education (SBE) approved design or blueprint, the recommended improvement, and the rationale for each improvement based on the results from these studies.

Table 6.1 Summary of CAST Blueprint Improvements

|  |  |  |
| --- | --- | --- |
| **SBE-Approved Design** | **Recommended Improvement** | **Rationale** |
| This design has a two‑stage adaptive session for Segment A that presents items that are substantively equivalent in content but may differ in difficulty to match each student’s level of performance. | Administer Segment A as a linear, nonadaptive test and reevaluate when the bank is more robust. | Based on the MST study, the number of items in the pool at higher difficulty ranges is insufficient to support building second stage blocks. |
| Use a screener to select two performance tasks (PTs) from three science domains based on student’s performance in Segment A. | Eliminate the use of a screener. | Results from the screener study show no clear evidence that a screener would result in improved performance on the CAST by eliminating content where student performance was conspicuously poor in relation to a student being randomly assigned content from any science domain. (A summary of findings can be found in this memorandum in the section titled “[Summary of Psychometric Analyses](#_Summary_of_Psychometric).”) |

Table 6.1 *(*c*ontinuation one)*

| **SBE-Approved Design** | **Recommended Improvement** | **Rationale** |
| --- | --- | --- |
| The existing test blueprint shows the same number of discrete operational items in Segment A across the three grade bands:   * 32–‍34 operational discrete items worth 42–44 points in Segment A   Based on response-time analyses, 42 percent of grade five students, 58 percent of grade eight students, and 92 percent of high school students completed the test in two hours. | Revise the test blueprint to establish a fixed number of discrete items in Segment A that will vary across grades based on the number of performance expectations (PEs) to be covered in a three-year cycle. The recommended numbers of items for Segment A are as follows:   * Grade five—26 items worth 28–32 points * Grade eight—28 items worth 30–34 points * High school—32 items worth 34–38 points | Based on response-time analyses, a high percentage of students in grades five and eight did not complete the test within two hours. The expected improvement under the recommended design would yield a 12 percent increase in students completing the test within two hours in grade five, a 14 percent increase in grade eight, and a slight increase in high school.  A breakdown of response‑time data for each grade can be found in this memorandum in the section titled “[Summary of Psychometric Analyses](#_Summary_of_Psychometric).” The high school design was modified to have a fixed number of discrete items to maintain consistency with the other grades. |
| The existing test blueprint has two operational PTs in Segment B. | Revise the test blueprint to have three PTs, one PT for each science domain (Physical Sciences, Life Sciences, and Earth and Space Sciences), in Segment B. | Three PTs will allow all three science domains to be represented in both operational segments A and B. |

Table 6.1 *(continuation two)*

| **SBE-Approved Design** | **Recommended Improvement** | **Rationale** |
| --- | --- | --- |
| The existing test blueprint for Segment C has students receiving one of the following:   * 12 to 14 discrete items (operational and field test) * One operational PT (4 to 6 items) * One field test PT (4 to 7 items) | Revise the test blueprint for Segment C so students receive one of the following:   * 6 discrete field test items * one field test PT (6 items) | To accommodate the additional PT in Segment B and remain within the two‑hour testing window, the maximum number of items will need to be reduced.  The number of items recommended for Segment C is supported by the response-time analyses located in the section titled “[Summary of Psychometric Analyses](#_Summary_of_Psychometric).” |
| Using the operational items in segments A, B, and C provide additional, unique information at the group level. This is in addition to the aggregation of the student individual scores. | Investigate new types of information that can be reported at the group level utilizing items from segments A, B, and C. For example, reports may provide information about students’ relative stregths and areas of weakness. | It is important to provide more information to improve instruction. This has been supported by both national experts and California educators. |
| The test is administered as a single test session and can be paused as needed based on class period or student needs. The test is estimated to take approximately two hours. | Based on timing data across grades and advice from the technical advisory group, investigate current test administration practices, such as class period time constraints. | Feedback from the field suggests there may be a need to recognize class period constraints in the test administration, which may differ by grade. |

1. At the time of this writing, the 2019–20 test forms are nearing finalization. It is possible that the anticipated number of PEs covered could change. [↑](#footnote-ref-2)
2. Items written to assess PEs associated with Engineering, Technology, and the Applications of Science was assigned to one of the three science domains, depending upon the context of their stimulus, and was not considered in the assignment of Segment B. [↑](#footnote-ref-3)